Bibliothèques numériques et crowdsourcing

Mathieu Andro

▶ **To cite this version:**

Mathieu Andro. Bibliothèques numériques et crowdsourcing : Expérimentations autour de Numalire, projet de numérisation à la demande par crowdfunding. Bibliothèque électronique [cs.DL]. Université Paris 8 Vincennes Saint-Denis, 2016. Français. NNT : . tel-01384692

HAL Id: tel-01384692
https://hal.science/tel-01384692

Submitted on 25 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

THÈSE DOCTORALE (2012-2016)

Bibliothèques numériques et *crowdsourcing* : expérimentations autour de Numalire, projet de numérisation à la demande par *crowdfunding*

Mathieu Andro

Laboratoire Paragraphe [EA 349]

Doctorat en Sciences de l'Information et de la Communication

Université Paris 8 Vincennes – Saint Denis

Soutenue le 10 octobre 2016

Jury :

Directeur de la thèse	M. Imad SALEH	Professeur à l'Université Paris 8 – Saint-Denis
Co-Directeur de la thèse	M. Samuel SZONIECKY	Maître de Conférences à l'Université Paris 8 – Saint-Denis
Rapporteur	Mme Ghislaine CHARTRON	Professeur au Conservatoire National des Arts et Métiers
Rapporteur	M. Stéphane CHAUDIRON	Professeur à L'Université Charles-de-Gaulle, Lille 3
Examinateur	Mme Céline PAGANELLI	Maître de conférences - HDR à l'Université Paul Valéry, Montpellier 3
Examinateur	M. Alain GARNIER	Directeur Général de la société Jamespot

Vous pouvez diffuser, partager, adapter et même utiliser commercialement cette œuvre en contrepartie de sa citation

Table des matières

Résumé ... 8
Mots clés libres de l'auteur ... 9
Indexation RAMEAU ... 9
Indexation LCSH ... 9
Remerciements ... 11
Avertissement ... 16
Introduction générale ... 17
0- Méthodologie utilisée dans le cadre de la thèse ... 22
0.1- Analyse de la littérature ... 22
0.1.1- Constitution d'une bibliographie ... 23
0.1.2- Analyses bibliométriques ... 23
0.1.2.1- Analyses à partir d'un corpus sans a priori ... 23
0.1.2.2- Analyses à partir de la bibliographie de la thèse ... 26
0.2- Dispositif de veille ... 34
0.3- Expérimentations conduites dans le cadre de la thèse ... 35
0.3.1- Observation participante ... 35
0.3.2- Enquêtes ... 36
0.3.3- Recherche-action ... 36
0.4- Rédaction en Google Doc : une thèse sur le *crowdsourcing* qui bénéficie du *crowdsourcing ?* ... 38
Chapitre 1- Introduction conceptuelle sur la notion de *crowdsourcing* en bibliothèque : un nouveau paradigme ? ... 41
1.1- Un modèle économique en plein essor ... 41
1.1.1- Ce qui a rendu possible ce nouveau modèle économique ... 41
1.1.2- et son application aux bibliothèques numériques ... 45
1.1.3- et suscite l'intérêt croissant des politiques, des internautes et des universitaires ... 48
1.2- Une origine, une traduction, une définition, et un périmètre du *crowdsourcing* ... 53
1.3- Chronologie historique du *crowdsourcing* ... 63
1.4- Controverses philosophiques et politiques ... 68
1.5- Conséquences économiques, sociologiques et juridiques ... 83
1.5.1- Économie du *crowdsourcing* ... 83
1.5.1.1- La disparition du travail nécessaire ? ... 84

1.5.1.2- *crowdsourcing*, revenu de base et théorie des biens communs..... 86
1.5.1.3- L'amateur, nouveau moteur de l'économie et du développement ?88
1.5.2- Les usagers du *crowdsourcing*..... 91
1.6- Conséquences managériales, bibliothéconomiques et technologiques..... 93
1.6.1- L'exception française..... 93
1.6.2- Les bibliothèques françaises, exception dans l'exception ?..... 94
1.6.3- Le règne de l'amateur : vers une médiocratie ?..... 98
1.6.4- Le *crowdsourcing*, stade suprême de l'externalisation ?..... 100
Chapitre 2- Panorama de quelques projets de *crowdsourcing* appliqués à la numérisation des bibliothèques..... 102
2.1- Mise en ligne et curation participatives : l'Oxford's great war archive et Europeana 1914-1918..... 102
2.2- la numérisation à la demande sous forme de *crowdfunding* appliquée aux bibliothèques numériques : le réseau européen ebooks en demand (EOD)..... 104
2.3- L'impression à la demande (*Print on Demand*, POD) : l'Espresso Book Machine..... 121
2.4- La correction participative de l'OCR et la transcription participative de manuscrits..... 130
2.4.1- Le *crowdsourcing* explicite : la correction / transcription volontaire..... 133
2.4.1.1- La correction participative et volontaire de l'OCR : l'Australian Newspapers Digitisation Program (TROVE)..... 133
2.4.1.2- La transcription participative et volontaire de manuscrits : Transcribe Bentham..... 140
2.4.2- La *gamification*, la correction de l'OCR en jouant : Digitalkoot (Bibliothèque Nationale de Finlande)..... 149
2.4.3- Le *crowdsourcing* implicite : la correction involontaire de l'OCR via reCAPTCHA au service de Google Books..... 154
2.4.4- Le *crowdsourcing* rémunéré : l'Amazon mechanical turk marketplace (AMT)..... 162
2.5- Folksonomie, catalogage et indexation participatives..... 184
2.5.1- Le *crowdsourcing* explicite par tagging volontaire : Flickr: The Commons..... 184
2.5.2- Le recours à la *gamification* : Art Collector..... 186
2.6- La traduction participative..... 191
Chapitre 3- Analyses, du point de vue des sciences de l'information et de la communication, sur le *crowdsourcing* pour les bibliothèques numériques 193
3.1- Typologies et taxonomies des projets..... 193

3.1.1- Le *crowdsourcing* explicite 208
3.1.1.1- Le *crowdsourcing* explicite gratuit 208
3.1.1.2- Le *crowdsourcing* explicite rémunéré 208
3.1.2- Le *crowdsourcing* implicite 209
3.1.3- La *gamification* 209
3.2- Communication et marketing pour recruter les bénévoles 218
3.3- La question des motivations 222
3.3.1- Les motivations intrinsèques 223
3.3.2- Les motivations extrinsèques 226
3.3.3- L'opposition entre les motivations intrinsèques et extrinsèques 227
3.3.4- Les motivations spécifiques des projets de *gamification* 228
3.3.5- *Crowdsourcing* et récompenses 230
3.3.6- Les autres théories sur les motivations 232
3.3.7- Les motivations des institutions culturelles et les pré-requis pour lancer un projet de *crowdsourcing* 235
3.4- Sociologie des contributeurs et *community management* 238
Dans l'introduction conceptuelle, nous avons déjà évoqué la sociologie des contributeurs des projets de crowdsourcing de manière générale. Nous traiterons ici plus précisément des usagers dans le domaine des bibliothèques numériques en particulier et à la lumière des projets que nous avons analysés. 238
3.4.1- Sociologie des contributeurs 238
3.4.2- *Crowdsourcing* ou *communitysourcing* ? 241
3.4.3- Le travail des professionnels autour de ces projets et le *community management* 242
3.5- La question de la qualité des contributions 247
3.5.1- Les systèmes d'évaluation et de modération des contributions 247
3.5.2- Comparaison entre la qualité des données produites par les amateurs et celles produites par les professionnels 252
3.5.3- La réintégration des données produites 255
3.5.4- Le statut juridique des contributions : *crowdsourcing* et web sémantique 257
3.6- L'évaluation des projets de *crowdsourcing* 258
3.6.1- les facteurs de réussite et d'échecs 261
3.6.2- Évaluation quantitative des projets de *crowdsourcing* et de leurs coûts 263
3.7- La conduite du changement 268

3.8- Les grandes étapes d'un projet de *crowdsourcing* 271

Chapitre 4- Expérimentations autour de la correction participative de l'OCR et autour du *crowdfunding* 273

4.1- Première expérimentation autour de Wikisource à l'Ecole Nationale Vétérinaire de Toulouse en 2008 274

4.2- Le projet de plateforme mutualisée et participative du PRES Sorbonne Paris-Cité 275

4.3- Participation au lancement du projet de *crowdfunding* Numalire (Yabé) ... 280

4.3.1- Présentation du projet 280

4.3.2- Référencement web et profil des visiteurs du site numalire.com 284

4.3.3- Propositions d'améliorations 289

4.3.3.1- Numériser sans devis et favoriser l'achat impulsif 289

4.3.3.2- Diminuer les coûts de numérisation par son "ubérisation" 294

4.3.3.3- Élargir à d'autres types de documents que le livre imprimé 297

4.3.3.4- Communiquer d'avantage sur les réseaux sociaux et s'appuyer sur les investisseurs, les mécènes, les libraires 298

4.3.3.5- Améliorer le référencement du site en multipliant les liens vers ses notices et en créant une bibliothèque numérique 301

4.3.3.6- Élargir à l'international et devenir partenaire du réseau européen Ebooks on Demand (EOD) 303

4.3.3.7- Numériser le contenus des bibliothèques sans convention préalable et changer de statut 304

4.3.4- Enquêtes auprès des bibliothèques et auprès des clients 305

4.3.4.1- Auprès de bibliothèques 306

4.3.4.2- Auprès des clients 308

4.3.5- Résultats et conclusions de l'expérimentation 308

Conclusion de la thèse 310

Annexe 1 : Autres projets non évoqués dans le chapitre 2 (Panorama des projets de *crowdsourcing* appliqués à la numérisation des bibliothèques) 317

1- Mise en ligne et curation participatives : Internet Archive 317

2- la numérisation à la demande sous forme de *crowdfunding* appliquée aux bibliothèques numériques 321

2.1- Le livre à la carte, Phénix Éditions 321

2.2- Juan Pirlot de Corbion : de Chapitre.com à YouScribe 323

2.3- Adopter un livre sur Gallica 323

2.4- Le projet *crowdfunding* Numalire 327

2.5- revealdigital.com et Lyrasis 329

2.6- Le projet FeniXX 331
3- L'impression à la demande (*Print on Demand*, POD) 332
3.1- Electronic Library (eLib) et Higher Education Resources ON Demand (HERON) 332
3.2- Amazon BookSurge (CreateSpace) 333
3.3- Gallica et *Print on Demand* 333
4- La correction participative de l'OCR et la transcription participative de manuscrits 335
4.1- La correction participative de l'OCR 335
4.1.1- Distributed Proofreaders (DP ou PGDP) 335
4.1.2- Wikisource 338
4.1.3- California Digital Newspaper Collection (CDNC) 342
4.1.4- La Bibliothèque nationale de France et la plateforme Correct (projet FUI12 Ozalid) 343
4.1.5- Franscriptor 347
4.2- La transcription participative de manuscrits 347
4.2.1- What's on the menu ? (WOTM) 347
4.2.2- Ancient Lives 348
4.2.3- ArcHIVE 350
4.2.4- What's the score (WTS) 350
4.2.5- Transkribus 351
4.3- Les logiciels de correction / transcription participatives 354
4.3.1- T-Pen 354
4.3.2- FromThePage 354
4.3.3- Refine! 355
4.3.4- Scripto 356
4.4- La *gamification*, la correction de l'OCR en jouant 357
4.4.1- COoperative eNgine for Correction of ExtRacted Text (CONCERT) 357
4.4.2- TypeAttack 360
4.4.3- Word Soup Game 362
4.4.4- Un jeu pour corriger l'OCR en arabe 363
4.4.5- Biodiversity Heritage Library (BHL) : Smorball et Beanstalk 363
5- Folksonomie, catalogage et indexation participatives 367
5.1- Le *crowdsourcing* explicite : le tagging volontaire 367
5.1.1- le steve.museum 367

5.1.2- GLAM Wikimedia 368
5.1.3- Les herbonautes 370
5.1.4- Les projets néerlandais Glashelder! et VeleHanden 371
5.2- Le recours à la *gamification* 373
5.2.1- Google Image Labeler 373
5.2.2- ESP Game puis GWAP 374
5.2.3- Peekaboom 376
5.2.4- KissKissBan (KKB) 380
5.2.5- PexAce 381
5.2.6- museumgam.es 382
5.2.7- Metadata Games 383
5.2.8- SaveMyHeritage 385
5.2.9- Picaguess 386
Annexe 2- Une analyse du panorama des projets de crowdsourcing en bibliothèques 390
Histoire des projets de *crowdsourcing* en bibliothèques 390
Géographie des projets de crowdsourcing en bibliothèques 393
Les tâches externalisées par *crowdsourcing* en bibliothèques 397
Taxonomie du *crowdsourcing* en bibliothèques 400
Annexe 3- Résultats de l'enquête auprès des usagers de Numalire 404
Annexe 4- Equations de recherche utilisée pour constituer le corpus 416
Annexe 5- Illustration de la manière dont communiquent les projets 419
Annexe 6- Articles publiés dans le cadre de la thèse 423
Figures et des tableaux contenus dans la thèse 429
Figures 429
Tableaux 434
Bibliographie 436

Résumé

Au lieu d'externaliser certaines tâches auprès de prestataires ayant recours à des pays dont la main d'œuvre est bon marché, les bibliothèques dans le monde font de plus en plus appel aux foules d'internautes, rendant plus collaborative leur relation avec les usagers. Après un chapitre conceptuel sur les conséquences de ce nouveau modèle économique sur la société et sur les bibliothèques, un panorama des projets est présenté dans les domaines de la numérisation à la demande, de la correction participative de l'OCR notamment sous la forme de jeux (*gamification*) et de la folksonomie. Ce panorama débouche sur un état de l'art du *crowdsourcing* appliqué à la numérisation et aux bibliothèques numériques et sur des analyses dans le domaine des sciences de l'information et de la communication. Enfin, sont présentées des apports conceptuels et des expérimentations originales, principalement autour du projet Numalire de numérisation à la demande par *crowdfunding*.

Mots clés libres de l'auteur

crowdsourcing, communitysourcing, nichesourcing, financement participatif, crowdfunding, numérisation à la demande, digitization on demand, ebooks on demand, impression à la demande, print on demand, POD, correction participative de l'OCR, transcription participative de manuscrits, folksonomie, gamification, ludification, game with a purpose, GWAP, human computation, user generated content, bibliothèques, libraries, bibliothèques numériques, digital libraries.

Indexation RAMEAU

Bibliothèques virtuelles

Documentation de bibliothèque -- Numérisation

Crowdsourcing

Financement participatif

Gamification

Reconnaissance optique des caractères

Indexation LCSH

Digital libraries

Library materials – Digitalization

Crowdsourcing

Crowd funding

Human computation

User-generated content

On-demand printing

Optical character recognition

Figure 1. Nuage de mots portant sur le contenu de la thèse et réalisé avec tagxedo.com

Remerciements

- Imad Saleh, Professeur au laboratoire Paragraphe de l'Université Paris 8, pour avoir accepté d'encadrer ce projet de thèse, pour sa gentillesse et pour ses conseils tout au long du projet.
- Ghislaine Chartron, Professeur au Conservatoire National des Arts et Métiers pour avoir accepté d'être rapporteur de cette thèse.
- Stéphane Chaudiron, Professeur à L'Université Charles-de-Gaulle, Lille 3 pour avoir accepté d'être rapporteur de cette thèse.
- Céline Paganelli, Maître de conférences - HDR à l'Université Paul Valéry, Montpellier 3 pour avoir accepté d'être examinateur de cette thèse
- Alain Garnier, Directeur Général de Jamespot et référent *crowdsourcing* auprès du Groupement Français des Industries de l'Information (GFII) pour avoir accepté d'être examinateur de cette thèse
- Samuel Szoniecky, Maître de Conférences à l'Université Paris 8 – Saint Denis pour avoir accepté d'être examinateur de cette thèse et pour m'avoir invité à intervenir auprès de ses étudiants.
- François Houllier, Président de l'Institut National de la Recherche Agronomique pour m'avoir permis de participer, à ses côtés, à un groupe de travail sur les sciences citoyennes afin de remettre un rapport sur le sujet à la demande de nos Ministres de tutelles.
- Odile Hologne, de la Direction de la Valorisation, Information Scientifique et Technique de l'Institut National de la Recherche Agronomique, pour avoir encouragé les expérimentations autour du projet Numalire à l'Inra dans le cadre de mon travail.
- Filippo Gropallo et Denis Maingreaud, de la société Orange et de la société Yabé, pour leur projet Numalire auquel ils m'ont permis de participer et leur collaboration tout au long de ce travail de recherche.
- Marc Maisonneuve et Emmanuelle Asselin, de la société de consulting TOSCA pour leur collaboration dans le livre que nous avons publié ensemble sur les logiciels et les plateformes pour développer des bibliothèques numériques.

- Gaëtan Tröger de l'Ecole Nationale des Ponts pour sa collaboration dans l'étude que nous avons menée sur la visibilité et les statistiques de consultation des bibliothèques numériques.
- Pauline Rivière, de la Bibliothèque Sainte-Geneviève et Anaïs Dupuy-Olivier, de l'Académie de Médecine pour leur collaboration dans le retour de l'expérience Numalire que nous avons rédigé ensemble.
- Robert Miller, de Internet Archive, pour la collaboration que nous avons eue à la Bibliothèque Sainte-Geneviève qui est devenue la première bibliothèque en France à participer à Internet Archive.
- Stéphane Ipert du Centre de Conservation du Livre pour les collaborations et les intéressantes discussions que nous avons eues.
- Pierre Beaudoin et Rémi Mathis, précédent et actuel président de Wikimédia France, association avec laquelle des collaborations avec Wikisource ont été concrétisées (École Nationale Vétérinaire de Toulouse en 2008) ou seulement envisagées (Bibliothèque Sainte-Geneviève).
- Valérie Chansigaud, historienne des sciences et contributrice Wikipédia avec qui un premier contact avait été établi au Muséum puis une expérimentation pilote de numérisation et de correction participative de l'OCR conduite dès 2008 à l'Ecole Nationale Vétérinaire de Toulouse.
- Gilonne d'Origny de la société ondemandbook.com avec laquelle une collaboration pour une première implantation en France d'une Espresso Book Machine a bien failli se concrétiser.
- Daniel Teeter, de la société Amazon pour l'intéressante opportunité de partenariat que nous avons failli construire.
- Juan Pirlot de Corbion, fondateur de Chapitre.com et de YouScribe pour les passionnants échanges que nous avons eus au cours de nos rencontres.
- Daniel Benoilid, fondateur de la société de *crowdsourcing* rémunéré Foule Factory pour les discussions que nous avons eues.
- Jean-Pierre Gerault, Directeur Général de la société I2S, leader dans le domaine de la fabrication de scanner pour la numérisation patrimoniale, Président du

Comité Richelieu et Directeur Général de Publishroom, pour les intéressantes discussions que nous avons eues

- Arnaud Beaufort de la Bibliothèque nationale de France, rencontré à l'occasion de journées Wikimédia à l'Assemblée Nationale et avec lequel j'ai eu un intéressant entretien par la suite.
- Silvia Gstrein et Veronika Gründhammer, de l'Université de Innsbruck pour m'avoir invité à intervenir à la conférence Ebooks on Demand 2014.
- Yves Desrichard et Armelle de Boisse, de l'Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques pour m'avoir permis d'intervenir aux journées "Quoi de neuf en bibliothèques ?" ces 5 dernières années.
- Thierry Claerr du Ministère de la Culture qui m'a permis d'intervenir régulièrement à l'ENSSIB, qui m'a sollicité pour la rédaction d'un ouvrage collectif et avec lequel j'ai eu des discussions très enrichissantes.
- Jean-Marie Feurtet de l'Agence Bibliographique de l'Enseignement Supérieur pour la collaboration que nous avons eue autour d'un projet de mutualisation d'une bibliothèque numérique et pour m'avoir invité à intervenir aux journées ABES 2011.
- Nicolas Turenne, de l'Institut National de la Recherche Agronomique pour m'avoir invité à exposer les premiers résultats de ses travaux au séminaire de l'axe "traces digitales" (groupe Cortext, Institut Francilien Recherche Innovation Société).
- Pierre-Benoît Joly, directeur de l'Institut Francilien Recherche, Innovation, Société (IFRIS) pour m'avoir invité à donner un cours au master Etudes numériques et innovation (NUMI)
- La SNCF pour le confort des voyages en train pendant lesquels la thèse a été rédigée.
- Google pour le service Google Drive qui a été utilisé pour rédiger la thèse tout en donnant accès en temps réel à la rédaction du document à mon directeur de thèse, à mes collaborateurs et à mes contacts qui avaient ainsi la possibilité d'y ajouter des commentaires.
- Mon épouse Véronique et mes 3 enfants Terence, Orégane et Eloïse.

Je tiens également à remercier les personnes suivantes pour les commentaires constructifs qu'elles ont mis sur le texte de la thèse diffusée dans sa version première en Google Drive :

Christine Young (1 relecture d'article en anglais), Wilfrid Niobet (1 idée, 8 pistes, 6 corrections), Célya Gruson-Daniel (3 pistes, 4 corrections), Olivia Dejean : (9 corrections), Michaël Jeulin (7 corrections), Catherine Thiolon (10 pistes), Caroline Dandurand (5 pistes), Diane Le Hénaff (3 pistes), Sophie Aubin (2 pistes), Nicolas Ricci (1 piste), Pauline Rivière (1 piste), Frédérique Bordignon (1 piste), Sylvie Cocaud (1 piste), Marjolaine Hamelin (1 piste), Silvère Hanguehard (1 piste), Christine Sireyjol (1 piste), Odile Viseux (1 piste), Véronique Decognet (1 piste), Dominique Fournier (2 corrections), et tous les "soldats inconnus" qui sont restés anonymes dans leurs commentaires (82 corrections)

« Que la force me soit donnée de supporter ce qui ne peut être changé et le courage de changer ce qui peut l'être mais aussi la sagesse de distinguer l'un de l'autre »

(Marc Aurèle)

« *La complaisance fait les amis, la vérité engendre la haine* »

(Terence)

« *L'amitié, le patriotisme, l'amour, tous les sentiments nobles, sont aussi une espèce de foi. C'est parce qu'ils ont cru que les Codrus, les Pylade, les Régulus, les Arrie, ont fait des prodiges. Et voilà pourquoi ces cœurs qui ne croient rien, qui traitent d'illusions les attachements de l'âme et de folie les belles actions, qui regardent en pitié l'imagination et la tendresse du génie, voilà pourquoi ces cœurs n'achèveront jamais rien de grand, de généreux : ils n'ont de foi que dans la matière et dans la mort, et ils sont déjà insensibles comme l'une et glacés comme l'autre*. »

(Chateaubriand)

« Ils ne savaient pas que c'était impossible, alors ils l'ont fait. »

(Mark Twain)

Avertissement

Le sujet du *crowdsourcing* est loin d'être consensuel. Il génère des divergences, des tensions et des polémiques. Dans la première partie conceptuelle de la thèse, nous avons été amenés à exposer les différents points de vues sur le sujet de la manière la plus exhaustive, équilibrée et équitable possible, en évitant de prendre part à toute polémique. Les points de vues rapportés ici n'engagent donc ni leur auteur ni le jury de la thèse. A l'issue de l'examen de ces points de vues, l'auteur de ces travaux reste d'ailleurs personnellement très partagé à leur sujet.

Certains modes de fonctionnement et états de faits comme l'externalisation dans des pays en voie de développement du travail de correction de l'OCR, le recours au travail involontaire et inconscient ou encore au travail faiblement rémunéré sur des plateformes comme l'Amazon Mechanical Turk Marketplace ont également été rapportés dans cette première partie. Ils sont également susceptibles de choquer les consciences en fonction des convictions ou des croyances. Nous avons néanmoins décidé d'évoquer leur existence sans éluder ce qui est susceptible de heurter. Notre propos n'a pas été de juger tel ou tel modèle économique sur le plan de la morale mais de rapporter leur mode de fonctionnement et de les analyser de la manière la plus rationnelle et scientifique possible. Leur évaluation n'a été analysée qu'au seul point de vue des avantages et des inconvénients, à la mesure de ce qu'ils rapportent et de ce qu'ils coûtent sans prendre part à un quelconque jugement moral.

Introduction générale

Les bibliothèques ont déjà eu recours à l'externalisation de certaines tâches de saisies de notices bibliographiques, de catalogage, d'indexation ou encore de correction de l'OCR auprès de prestataires dans des pays où la main d'œuvre est dite à bas coût. Cette externalisation est demeurée dans un cadre contractuel et limité et n'a pas bouleversé en profondeur le mode de fonctionnement sur lequel repose les bibliothèques. Mais, avec le développement du *crowdsourcing*, il pourrait être envisagé d'externaliser (« *outsourcing* ») certaines de ces tâches, non plus auprès de prestataires, mais auprès de foules (« *crowd* ») d'internautes et donc de faire faire une partie du travail des professionnels par des amateurs. Le « *crowdsourcing* » modifie ainsi le paradigme sur lequel repose des bibliothèques largement centrées sur la constitution et la conservation de collections. Il modifie également le rapport entre les producteurs d'un service que sont les bibliothécaires et ses consommateurs que sont les usagers, ces derniers devenant également des producteurs actifs du service. Le *crowdsourcing* pourrait aussi interroger les politiques documentaires des bibliothèques qui anticipent les besoins dans une logique d'offre qui n'est pas directement et immédiatement déterminée par la demande. C'est particulièrement le cas avec la numérisation à la demande par *crowdfunding*, une forme de *crowdsourcing* faisant appel, non pas au travail des foules mais à leurs ressources financières ou avec l'impression à la demande qui lui est indissociable. Avec ces modèles économiques à la demande, la politique documentaire est finalement partagée avec les usagers qui décident de ce qui sera numérisé et/ou imprimé. Les collections deviennent ainsi l'œuvre des usagers.

Cette thèse a pour objet d'apporter des éléments de réponse à la question du recours au *crowdsourcing* à destination des professionnels des bibliothèques. Au delà des questions coûts / bénéfices et avantages / inconvénients, la question d'une évolution du métier de bibliothécaire recentré sur ses compétences singulières sera abordée. Cette thèse a également pour objectif scientifique d'apporter une contribution à la connaissance du *crowdsourcing* sur le plan théorique et conceptuel autour des modèles économiques. Enfin, elle livre des

résultats et des analyses originales dans le cadre d'une expérimentation autour d'un projet de numérisation à la demande par *crowdfunding*.

La partie introductive de la thèse permet d'expliciter son contexte et la méthodologie qui a été utilisée.

Le premier chapitre conceptuel de la thèse aborde les représentations philosophiques, politiques, économiques du *crowdsourcing* et ses conséquences sur le mode de fonctionnement des bibliothèques. Ce chapitre conceptuel contient, en particulier :

- une discussion critique à propos de la définition de *crowdsourcing* ;
- une chronologie originale de ses origines historiques ;
- une analyse au sujet de ses origines conceptuelles auprès de courants philosophiques parfois diamétralement opposés et, en particulier, un apport conceptuel autour de la loi de la valeur ;
- une réflexion sur le concept de sagesse des foules ;
- une analyse des diverses critiques du *crowdsourcing* appliqué aux bibliothèques numériques que certains pourraient qualifier, aujourd'hui, de « ubérisation » des bibliothèques numériques.

Le second chapitre contient une sélection de projets par types de tâches avec :

- la mise en ligne et la curation participatives ;
- la numérisation et l'impression à la demande sous forme de *crowdfunding* ;
- la correction participative de l'OCR et la transcription participative de manuscrits ;
- et enfin, la folksonomie.

Ce chapitre contient des données et des informations récoltées dans la littérature pour chaque projet. Un panorama plus exhaustif des projets est fourni en annexes. Des analyses originales pour chaque grand type de projets sont données en conclusion de ce second chapitre.

En troisième chapitre des analyses du point de vue des sciences de l'information et de la communication sont proposées avec, en particulier :

- une taxonomie originale du *crowdsourcing* en bibliothèque numérique distinguant *crowdsourcing* explicite (ou conscient) bénévole et rémunéré, *crowdsourcing* implicite (ou inconscient), *gamification* et *crowdfunding* ;
- une analyse des motivations des bibliothèques et des conditions nécessaires au développement de projets de *crowdsourcing* ;
- une taxonomie des motivations des internautes qui contribuent à leurs projets ;
- des analyses sur les récompenses et rémunérations éventuelles ;
- un éclairage à propos de la communication nécessaire au recrutement ;
- des développements sur le *community management* spécifique de ce type de projets ;
- des analyses sur la question de la qualité et de la réintégration des données produites ;
- et enfin, une réflexion sur l'évaluation des projets de *crowdsourcing*. .

Le dernier chapitre, débouche sur des expérimentations conduites autour du *crowdsourcing* avant le doctorat (à l'Ecole Nationale Vétérinaire de Toulouse avec Wikisource en 2008, à la Bibliothèque Sainte-Geneviève avec le projet de plateforme de *crowdsourcing* de Sorbonne Paris-Cité entre 2009 et 2012) et surtout, dans le cadre du doctorat depuis 2013.

C'est, en particulier, autour de Numalire, un projet de numérisation à la demande par crowdfunding qu'ont été faites la plupart de nos expérimentations. Ce chapitre contient la description et les résultats du projet et les propositions d'améliorations du projet dans lequel nous sommes intervenu à la fois en tant que partenaire d'une institution participante, porte parole des bibliothèques participantes, veilleur, et consultant spécialiste du *crowdsourcing*. Le choix de mener l'essentiel de nos expérimentations autour du *crowdfunding*, c'est à dire d'une forme très particulière de *crowdsourcing*, est lié au fait qu'il existe très peu d'études sur ce sujet dans la littérature. Nous souhaitions également profiter de cette opportunité pour participer à la création d'une entreprise autour d'un modèle économique innovant. Au delà de cette seule expérimentation, cette thèse est

également le résultat d'une observation participante et d'une pratique de plusieurs années au sein des bibliothèques scientifiques sur lesquelles elle vise à avoir une action. Elle s'inscrit donc pleinement dans le cadre d'une recherche-action.

Enfin, on trouvera en annexe un complément à notre panorama de projets. En effet, afin de ne pas déséquilibrer cette thèse, nous avons choisi de ne présenter, en partie centrale qu'un seul projet représentatif de chaque type et de présenter les autres en annexe. On trouvera également des analyses de ce panorama, les résultats de l'enquête conduite dans le cadre de l'expérimentation Numalire, les équations de recherche utilisées pour constituer le corpus de publications analysées dans la thèse, des illustrations de manière de communiquer des projets et, pour finir, la liste des articles publiés dans le cadre de ce travail de recherche.

Les principaux apports de cette thèse résident dans les taxonomies originales qui sont données du *crowdsourcing* en bibliothèques avec une matrice des types de projets restant à inventer mais aussi une taxonomie des motivations des contributeurs. Cette dernière taxonomie est principalement le résultat d'une analyse de la littérature. Une contribution à la définition du *crowdsourcing* et une discussion au sujet de la notion de *crowdsourcing* implicite sont également proposées. Le calcul des coûts rapportés aux bénéfices d'un certain nombre de projets nous semble également être un sujet encore peu abordé et pour lequel des résultats sont présentés dans la thèse. Dans le cadre d'une observation participante, nous proposons des analyses originales des réticences dans les bibliothèques de France. Nous avons également identifié deux conceptions opposées concernant la démocratisation et la communication autour d'un sujet à la mode d'une part et la diminution des coûts corrélée à l'augmentation des résultats d'autre part. Enfin, des expérimentations originales ont été relatées, en particulier dans le domaine du *crowdfunding*, une forme de *crowdsourcing* très peu étudiée en bibliothèques comme nous l'avons déjà dit et pour laquelle nous proposons une analyse originale. L'expérimentation de ce modèle économique appliqué à la numérisation des documents conservés dans les bibliothèques et la proposition de

solutions concrètes pour le rendre rentable et viable pour une mise en œuvre future nous semble être l'apport central de cette thèse.

0- Méthodologie utilisée dans le cadre de la thèse

Comme nous l'avons précédemment explicité, l'objectif premier de ce travail de recherche a été d'analyser les avantages et les bénéfices comparés aux inconvénients et aux coûts du *crowdsourcing* pour les projets de numérisation des bibliothèques. Pour cela, la méthodologie suivante a été utilisée :

0.1- Analyse de la littérature

Une bibliographie la plus exhaustive possible a été réalisée sur le sujet du *crowdsourcing* et du *crowdfunding* appliqué à la numérisation des bibliothèques à partir d'équations de recherches dans des bases bibliographiques. Cette bibliographie a fait l'objet d'un plan de lectures de chaque publication au cours de la première année de doctorat qui a essentiellement été consacrée à de l'autoformation. Une analyse bibliométrique de ce corpus a également permis d'obtenir une première représentation générale et introductive du sujet. Elle nous a également amenés à mettre en place un dispositif de veille éditoriale qui a permis d'être systématiquement alerté de toute nouvelle publication sur le sujet. La bibliographie a ainsi été régulièrement mise à jour à partir de la veille et aussi à partir de l'exploitation des bibliographies contenues dans les publications lues.

Parmi toutes les publications qui ont été lues, seule une mineure partie a été utilisée et a été référencée dans la bibliographie finale de cette thèse. Ainsi, sur 821 documents, la bibliographie finale ne compte que 248 références bibliographiques soit moins du tiers. En fin de thèse, une nouvelle analyse bibliométrique a été produite sur ces 248 références bibliographiques sélectionnées et utilisées. Cette analyse donne une représentation originale et plus fine du domaine.

0.1.1- Constitution d'une bibliographie

Les sources suivantes ont été interrogées en fonction des requêtes suivantes :

	A- Web of Science	B- ScienceDirect	D- Google Scholar	Total
1- *crowdsourcing* et bibliothèques	A1 = 37 notices validées	B1 = 4 notices validées	C1 = 116 notices validées	157 notices validées
2- *Print on demand* et bibliothèques	A2 = 4 notices validées	B2 = 0 notices	C2 = 34 notices validées	38 notices validées

Tableau 1. Sources utilisées pour constituer le corpus analysé dans la thèse

0.1.2- Analyses bibliométriques

Les notices bibliographiques ont ensuite été importées sous EndNote, dédoublonnées, nettoyées manuellement, retravaillées avec Open Refine et analysées statistiquement avec Excel afin de produire les analyses bibliométriques suivantes.

Le nombre d'articles académiques et scientifiques sur le *crowdsourcing* est relativement faible par rapport à la littérature grise sur le sujet (rapports, actualités, newsletters, manifestations...) comme le confirme (Zhao, 2012). Le *crowdsourcing* est, en effet, encore un domaine émergeant.

0.1.2.1- Analyses à partir d'un corpus sans a priori

Le corpus de final de 156 notices est assez peu volumineux. La recherche dans le domaine ne semble donc pas encore très développée. Néanmoins, ce corpus a été utilisé afin d'identifier les pays, les institutions et les auteurs qui travaillent le plus sur ce sujet.

On observe une forte croissance de la recherche mondiale sur le sujet ces dernières années.

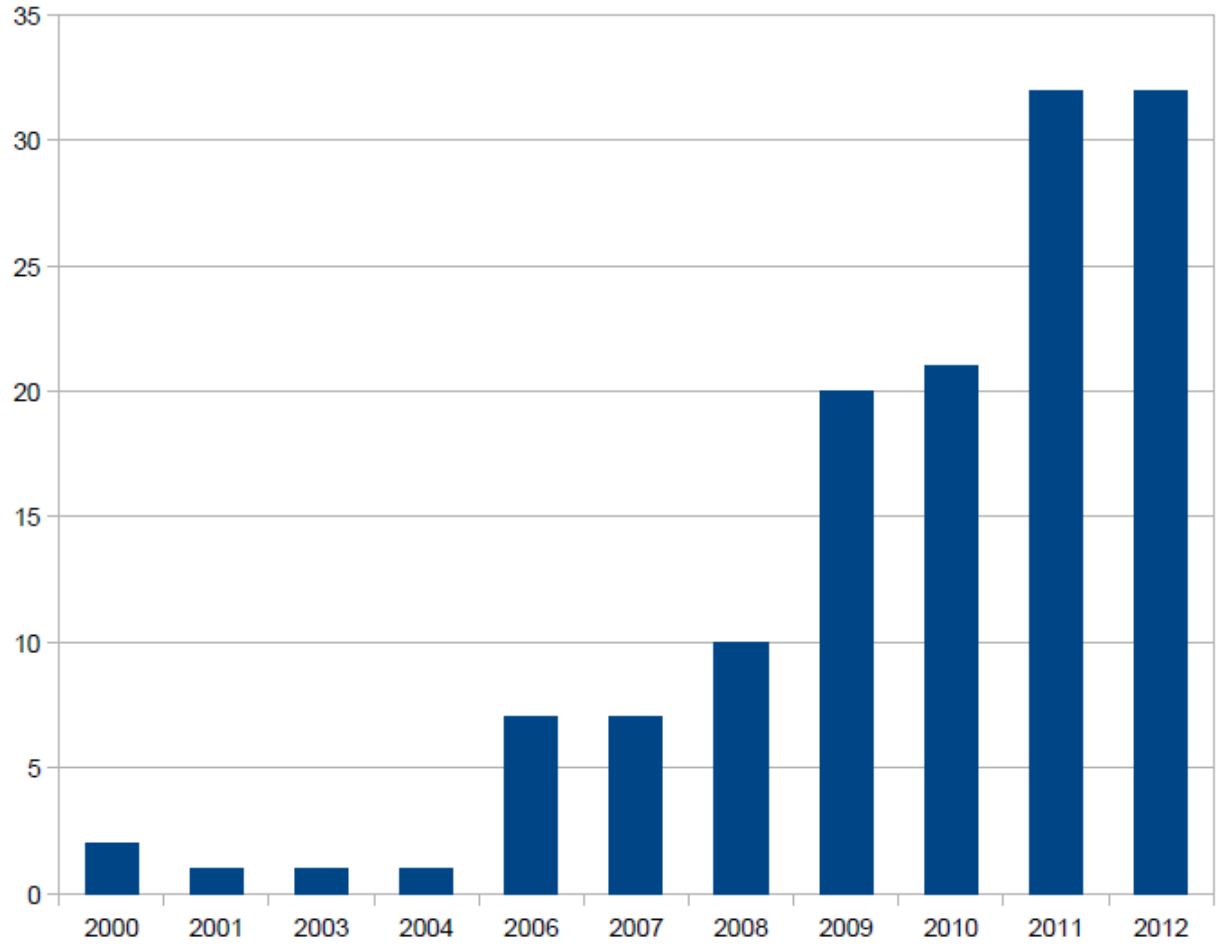


Figure 2. Analyse temporelle de l'évolution du nombre de publications dans le corpus bibliométrique

Néanmoins, il est possible qu'on ait fait du *crowdsourcing* sans le savoir et surtout bien avant que le mot ne soit inventé par Jeff Howe en 2006. Par conséquent, certains articles n'apparaissent pas dans ce corpus préliminaire, mais ont été découverts au fur et à mesure des lectures, notamment via les bibliographies des publications.

Les principales revues ayant publié des articles sur le sujet sont les suivantes :

- D-Lib Magazine 4
- Literary and Linguistic Computing 4

- Program-Electronic Library and Information Systems 3
- Research and Advanced Technology for Digital Libraries, Proceedings 3
- ZooKeys 3
- Archival Science 2
- Electronic Library 2
- Journal of Documentation 2
- Journal of Library Administration 2
- Liber Quarterly 2
- Library Hi Tech 2
- Library Philosophy and Practice (e-journal) 2
- Libri 2
- SLIS Student Research Journal 2
- The Journal of Academic Librarianship 2

Les principaux auteurs spécialistes du sujet sont les suivants :

- Holley, Rose 7
- Gstrein, Silvia 6
- Trant, Jennifer 4
- Wallace, Valerie 4
- Causer, Tim 3
- Hall, Catherine 3
- Mühlberger, Günter 3
- Nichols, David M 3
- Svoljšak, Sonja 3
- Terras, Melissa 3
- Tonra, Justin 3
- Wyman, Bruce 3
- Zarro, Michael 3

Le corpus étant principalement constitué de notices en provenance de Google Scholar, il n’a pas été possible d’identifier des institutions significatives qui sont

renseignées dans seulement 37 notices en provenance du Web of Science. Concernant les pays, les suivants semblent dominer :

- USA 5
- Finlande 2
- Australie 2

0.1.2.2- Analyses à partir de la bibliographie de la thèse

Les analyses bibliométriques sont totalement dépendantes des corpus de départ. Or, ces corpus dépendent eux-mêmes fortement des équations de recherches qui ont présidé à leurs constitutions et qui présentent tantôt des lacunes (“silence documentaire”) tantôt des documents non pertinents (“silence documentaire”). Ces corpus sont, de surcroît, souvent insuffisamment validés et les documents qui les composent ne sont que très rarement lus ni mêmes survolés. A la différence des analyses bibliométriques proposées dans la partie précédente, nous nous proposons donc, à présent, de réaliser une analyse bibliométrique sur un corpus arrêté le 17 juin 2015 et portant sur 219 références bibliographiques toutes lues et sélectionnées une à une dans le cadre de recherches bibliographiques, mais aussi à partir d’utilisation des bibliographies des articles lus et enfin, via un dispositif de veille sous Digimind.

De manière générale, la répartition des publications en fonction des années, se fait de la manière suivante :

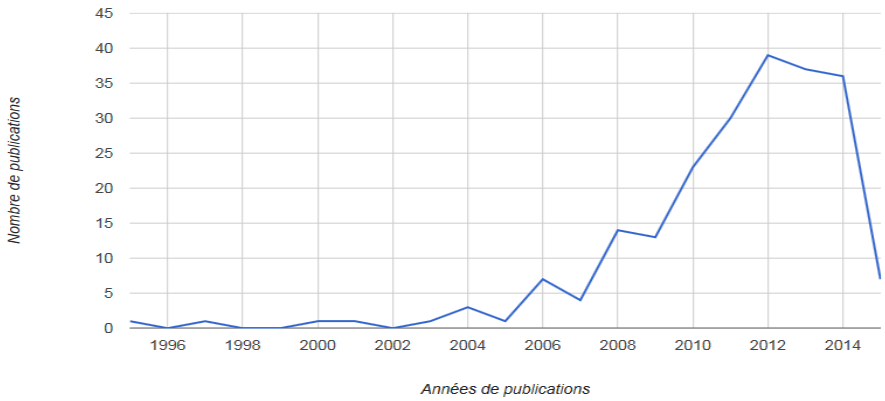


Figure 3. Nombre de publications par an dans la bibliographie de la thèse

On observe une croissance des publications particulièrement marquée entre 2005 et 2012 suivie d'une stagnation depuis 2013.

Le corpus contient 361 auteurs au total, 321 d'entre eux ne présentent qu'une seule occurrence. Les autres auteurs, les plus prolifiques, sont les suivants :

Holley R. 9	Adda, G 2
	Bozzon, A 2
McKinley, D. 7	Brabham, D. C. 2
Von Ahn, L 7	Causer, T 2
	Dabbish, L. 2
Aroyo, L 4	Dixon, D 2
Moirez, P 4	Dunn, S 2
Renault S. 4	Eveleigh, A. 2
Smith-Yoshimura, K 4	Fort, K 2
	Harris, C. G. 2
Alam, S. L 3	Hedges, M. 2
Blum, M. 3	Houben, G.J 2
Campbell, J. 3	Josse, I. 2
Deterding, S 3	Khaled, R 2
Dijkshoorn, C 3	Lakhani, K. 2
Gstrein, S 3	Nottamkandath, A 2
Ipeirotis, P. G 3	Onnée, S 2
Mühlberger, G 3	Oomen, J 2
Tonra, J 3	Oosterman, J 2
Wallace, V. 3	Paraschakis, D 2
	Revitt, M 2
	Ridge, M. 2
	Tzadok, A. 2

Concernant la France, les auteurs suivants ont, en particulier, été identifiés : Isabelle Josse, Pauline Moirez, Sophie Renault, Stéphane Onnée, Karen Fort, Gilles Adda, Alain Pierrot, Marthe Lagarrigue, Florence Rossant, Joël Gardes, Christophe Maldivi, Eric Petit, Edouard Bouyé, Benoit Sagot, Bernard Lang, Eric Schenk, Marc Pignal, Henri Le More, laude Guittard et Nathalie Casemajor Loustau.

Les 219 documents du corpus comprennent majoritairement des articles de revues et des actes de conférences :

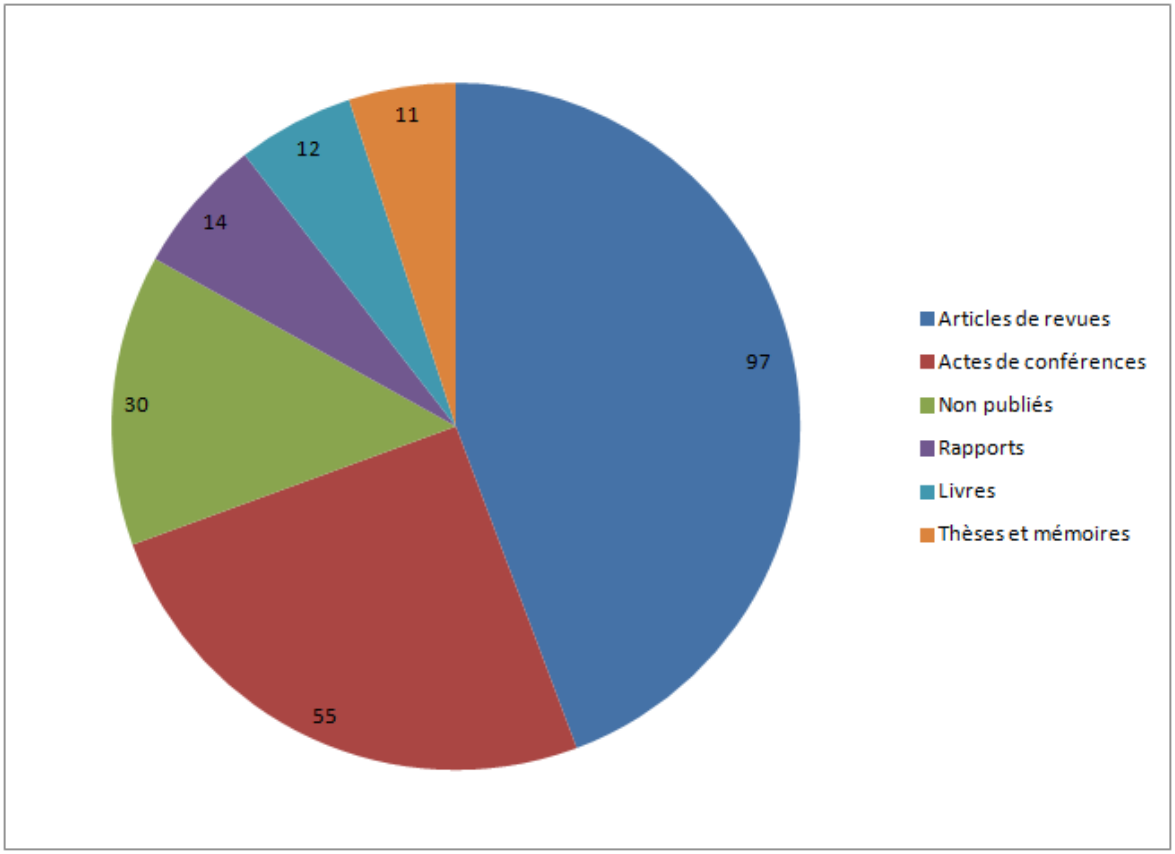


Figure 4. Répartition du corpus selon les types de documents

Parmi les revues, voici les plus occurrentes dans le corpus :

- Bulletin des Bibliothèques de France 6 (mais cette revue ne serait pas du tout apparue si nous avions considéré seul le corpus en langue anglaise)
- D-Lib Magazine 5
- First Monday 4
- Library Hi Tech 3
- Archival Science 2
- Communications of the ACM 2
- Documentaliste - Sciences de l'information 2 (cette revue ne serait pas apparue si nous avions considéré seul le corpus en langues anglaise)
- Information, Communication & Society 2
- Interlending & Document Supply 2
- International Academic MindTrek Conference 2
- International Journal on Digital Libraries 2
- Journal of Information Science 2
- Liber Quarterly 2
- Literary and Linguistic Computing 2
- Science 2
- Serials 2
- The Museum Journal 2

Parmi les conférences :

- Conference on Human Factors in Computing Systems, CHI (ACM) 8
- Museums and the Web 2006 5
- Australasian Conference on Information Systems (ACIS) 3
- Hawaii International Conference on System Sciences 3
- IFLA World Library and Information Congress 3
- International Conference on Information Systems 2
- SIGCHI Conference on Human Factors in Computing Systems 2

Concernant les institutions des auteurs, nous avons saisis les institutions de chaque auteur dans un tableur et avons attribué 1 point par publication dans laquelle une institution apparaissait. Afin de ne pas biaiser cette analyse bibliométrique, nous avons fait abstraction des articles en langue française que nous avons naturellement consulté de manière privilégiée dans la mesure où il s'agit de notre langue maternelle.

Nous obtenons le classement suivant pour les articles internationaux parmi les 149 institutions du corpus :

- Australie - National Library of Australia - 11
- Nouvelle Zélande - Victoria University of Wellington - 9
- USA - Carnegie Mellon University - 8
- UK - University College of London - 6
- Danemark - University of Copenhagen - 5
- Autriche - Innsbruck University - 4
- Pays Bas - University Amsterdam - 4
- USA - OCLC - 4
- USA - University of Illinois - 4
- USA - University of Iowa - 4
- Allemagne - Hamburg University - 3
- Australie - University of Canberra - 3
- UK - Science Museum - 3
- UK - University of the West of England - 3
- USA - Harvard University - 3
- USA - New York University - 3
- USA - University of Hawaï - 3
- USA - University of Virginia - 3
- Canada - University of Ontario - 2
- Canada - University of Toronto - 2
- Finlande - University of Tampere - 2
- France (langue anglaise) - INIST - 2

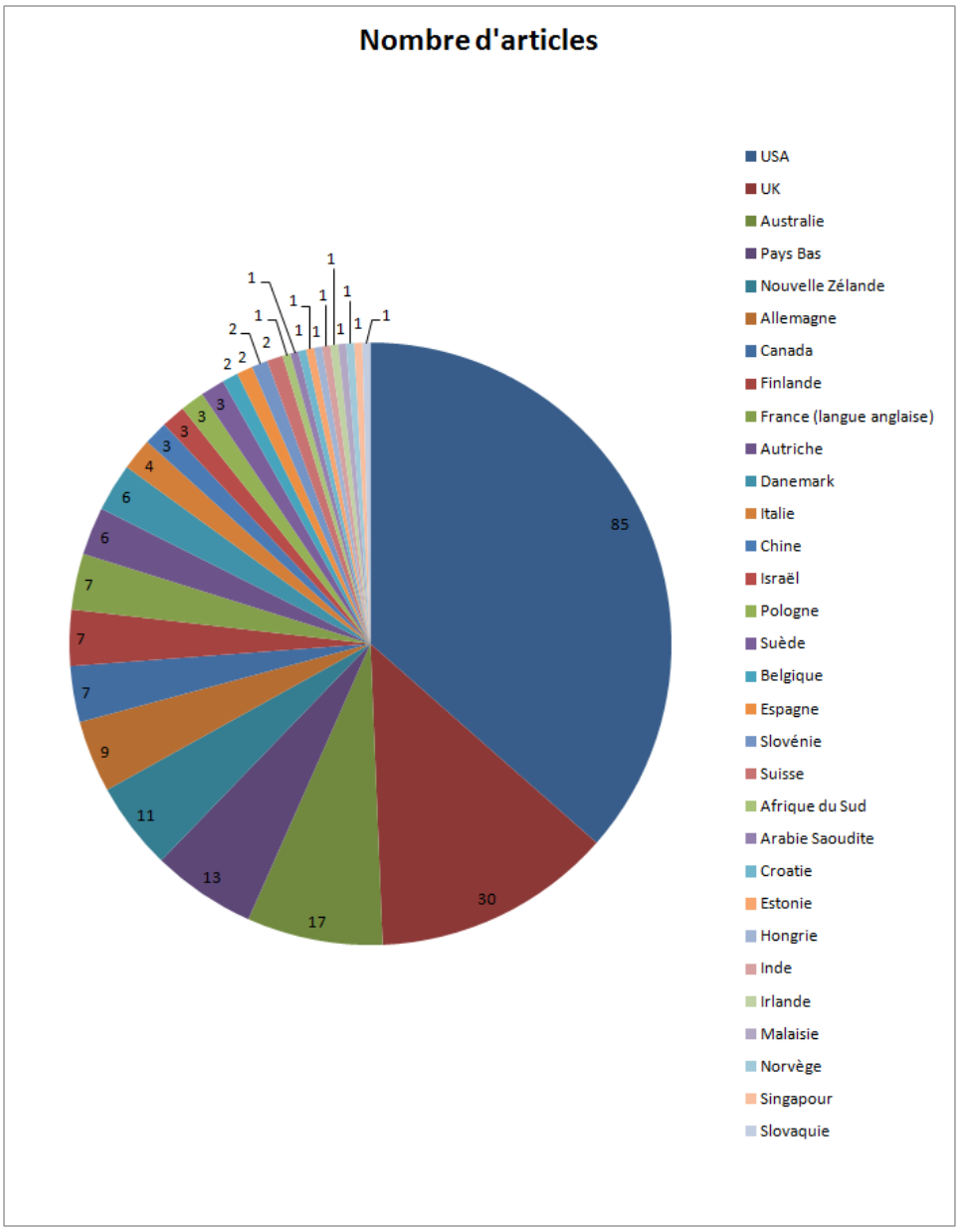
- France (langue anglaise) - Limsi - 2
- Israël - IBM Israël - 2
- Nouvelle Zélande - University of Waikato - 2
- Pologne - Nicolaus Copernicus University - 2
- Slovénie - National and University Library - 2
- Suède - Malmö University - 2
- UK - Harvard Business School - 2
- UK - King's College London - 2
- UK - University of Sheffield - 2
- USA - Cornell University - 2
- USA - Library of Congress - 2
- USA - School of Journalism & Mass Communication , UNC-Chapel Hill - 2
- USA - Stanford University - 2
- USA - University of Maine - 2
- USA - University of Maryland - 2
- USA - University of Michigan - 2
- USA - University of Utah - 2

Concernant plus spécifiquement la France, en considérant, à présent, à la fois les articles en langues anglaise et en langue française, le classement est le suivant :

- France (langue française) - Bibliothèque nationale de France - 5
- France (langue française) - Université d'Orléans - 4
- France (langue anglaise) - INIST - 2
- France (langue anglaise) - Limsi - 2
- France (langue anglaise) - I2S - 1
- France (langue anglaise) - ISEP - 1
- France (langue anglaise) - Orange - 1
- France (langue française) - Archives Départementales du Cantal - 1
- France (langue française) - Bibliothèque Municipale de Troyes - 1
- France (langue française) - collège Léon Jozeau Marigné d'Isigny-le-Buat - 1

- France (langue française) - Ever - 1
- France (langue française) - INRIA - 1
- France (langue française) - INSA de Strasbourg - 1
- France (langue française) - Muséum national d'Histoire naturelle - 1
- France (langue française) - Phénix éditions - 1
- France (langue française) - Rennes 2 - 1 -
- France (langue française) - Université de Strasbourg - 1
- France (langue française) - Université Lille 3 - 1

Si on compile les pays de ces institutions, on obtient la répartition suivante qui illustre la nette domination des Etats Unis et du Royaume Uni sur le reste des pays et l'absence de la plupart des pays des sud. Cette observation pourrait apparaître biaisée compte tenu de la domination générale de ces pays dans de domaine de la recherche scientifique et de la prise en compte, dans notre analyse, des seules publications en langue anglaise. Néanmoins, la même observation a été relevée à la fois par le rapport rédigé par Gille Bœuf, alors Président du Muséum national d'Histoire naturelle au sujet des sciences citoyennes à la demande du Ministère de l'Environnement (Bœuf, 2012) et par le rapport de François Houllier, alors Président de l'Institut National de la Recherche Agronomique au sujet également des sciences participatives rédigé à la demande du Ministère de l'Enseignement Supérieur et auquel nous avons eu le privilège de participer (Houllier, 2016). Par ailleurs, et en ce qui concerne plus particulièrement les bibliothèques, cette observation est également confirmée par le panorama des projets que nous avons identifiés.


Nombre d'articles
85
30
17
13
11
9
7
7
7
6
6
4
3
3
3
3
2
2
2
1
1
1
1
1
1
1
1
1
USA
UK
Australie
Pays Bas
Nouvelle Zélande
Allemagne
Canada
Finlande
France (langue anglaise)
Autriche
Danemark
Italie
Chine
Israël
Pologne
Suède
Belgique
Espagne
Slovénie
Suisse
Afrique du Sud
Arabie Saoudite
Croatie
Estonie
Hongrie
Inde
Irlande
Malaisie
Norvège
Singapour
Slovaquie

Figures 5 et 5bis. Poids des pays des auteurs des articles dans la bibliographie de la thèse

0.2- Dispositif de veille

Un dispositif de veille éditoriale a été mis en place via des alertes sur les bases bibliographiques et via Google Reader dans un premier temps, puis, à la disparition de cet outil gratuit, via Digimind, un outil plus élaboré dédié à la veille et auquel l'Institut National de la Recherche Agronomique est abonné.

Le dispositif de veille nous a permis de surveiller les bases bibliographiques filtrées par les équations de recherche déjà mentionnées précédemment, mais aussi les revues et les auteurs identifiés via l'analyse bibliométrique introductive et enfin, des sites institutionnels, des blogs, et la presse. Ce dispositif a permis de surveiller l'environnement du projet de thèse et d'être rapidement informé de toute actualité ou article en rapport avec le sujet de la thèse, d'identifier des opportunités (nouvelles publications, événements, collaborations possibles avec des équipes de recherche travaillant sur des sujets proches, innovations, bourses) et, éventuellement, de détecter aussi des menaces (nouveaux projets de thèses proches du notre en France, évolution défavorable du contexte, évolution juridique).

Ainsi, au lieu d'essayer de reformuler périodiquement les mêmes requêtes dans les sources d'informations et probablement de relire plusieurs fois les mêmes

informations, les requêtes et les sources ont été capitalisées, enregistrées, et complétées de manière itérative. Au lieu d'aller périodiquement chercher l'information, c'est l'information qui est systématiquement et automatiquement remontée jusqu'à nous.

0.3- Expérimentations conduites dans le cadre de la thèse

Dans le cadre de cette thèse, et parfois aussi quelques années auparavant, des expérimentations ont été conduites autour de la correction participative de l'OCR sous Wikisource à l'Ecole Nationale Vétérinaire de Toulouse, d'un cahier des charges de bibliothèque numérique mutualisée faisant appel au *crowdsourcing* pour le PRES Sorbonne Paris-Cité et la Bibliothèque Sainte-Geneviève et, enfin, du développement d'un projet de service de numérisation à la demande par *crowdfunding*. C'est autour de ce projet, nommé Numalire, qu'a porté l'essentiel de nos expérimentations qui ont permis d'alimenter notre propre réflexion conceptuelle.

0.3.1- Observation participante

Au delà de l'analyse de la littérature, nous avons ainsi mobilisé des méthodes pragmatiques d'observation participante et d'expérimentation empirique. Les résultats obtenus ont fait l'objet d'analyse sous la forme d'enquêtes mais aussi d'entretiens téléphoniques.

L'observation participante ou plutôt la participation observante (Soulé, 2007) est principalement liée à notre position de bibliothécaire documentaliste salarié et de notre expérience en bibliothèques. Ainsi, dès 2008, avant de commencer une thèse, nous avons conduit des expérimentations pionnières de correction participative de l'OCR de thèses vétérinaires sous Wikisource grâce à un mécénat de la fondation Wikimédia lorsque nous dirigions la bibliothèque de l'Ecole Nationale Vétérinaire de Toulouse. Quelques années plus tard, nous avons rédigé un cahier des charges de bibliothèque numérique mutualisée faisant appel au *crowdsourcing* et au *crowdfunding* pour le PRES Sorbonne Paris-Cité et la Bibliothèque Sainte-Geneviève. Ces expériences ont été accompagnées de veille, d'études, et de nombreux entretiens informels avec des prestataires et de riches

interactions avec des acteurs du domaine. Ces expériences ont donné lieu à une continuelle observation, à une prise de recul, une analyse critique de la pratique, une théorisation et à un effort de publication bien en amont de ce travail de recherche. Le *crowdsourcing* lui-même a d'ailleurs d'abord été une pratique professionnelle bien avant de devenir un concept et un sujet d'étude. La cohérence des diverses expériences rapportées dans cette thèse tient beaucoup à celle d'une carrière et à la permanence d'intérêt pour le sujet qui nous préoccupe.

0.3.2- Enquêtes

Dans le cadre de l'expérimentation, des enquêtes ont été menées auprès des responsables des bibliothèques partenaires sous la forme d'entretiens téléphoniques et auprès des internautes usagers du projet Numalire sous la forme d'un questionnaire. Ce dernier a été envoyé à 380 personnes possédant un compte sur le site web, mais seules 118 d'entre elles l'ont renseigné.

Les analyses de ces enquêtes sont proposées dans le chapitre expérimentation, les données collectées et les diagrammes produits sont accessibles en annexes.

0.3.3- Recherche-action

(Saint-Luc, 2014) rapporte une définition de la recherche-action de qui a été publiée dans le glossaire de l'Université Coopérative Internationale par Batide et Desroche et qui correspond bien à notre démarche : « *La recherche-action est un processus de recherche en sciences sociales donnant une large place à la prise en compte de l'expérience des acteurs dans l'analyse de pratiques concrètes (praxéologie) ; à l'implication des acteurs au processus d'objectivation et de formalisation (recherche impliquée) et enfin à la production d'un savoir utile dans l'action (recherche appliquée). C'est aussi une recherche d'explication ou recherche sur l'action ; une recherche d'application ou recherche pour l'action ; une recherche d'implication ou recherche par l'action* »

Cette thèse s'inscrit pleinement dans une démarche de type recherche-action. Elle cherche à dépasser la séparation entre théorie et pratique au travers d'une conceptualisation de l'expérience professionnelle d'un sujet y participant

pleinement en tant que professionnel de l'information scientifique et technique « converti » aux sciences de l'information et de la communication, en tant que « praticien-chercheur » (Morvan, 2013). La pratique professionnelle et l'expérimentation ont été utilisées comme sources de connaissances dans une démarche de co-formation et avec une confrontation des hypothèses théoriques avec la pratique concrète. Nous avons ainsi pu vérifier la pertinence d'un recours au *crowdfunding* pour financer la numérisation de livres anciens. Il a été nécessaire de trouver un équilibre entre l'abstraction théorique, parfois objective mais souvent sans lien avec le réel et une pratique concrète trop subjective pour être source de connaissances scientifiques. Cette thèse est le résultat de cette tension.

Il ne s'agissait pas simplement d'interpréter l'existant mais de chercher aussi à le transformer. La finalité et l'ambition de cette thèse est aussi que les connaissances produites dans le cadre de cette action aient, en retour, un effet sur cette pratique professionnelle, participe concrètement au changement dans les bibliothèques à la transformation de la réalité tout en produisant des connaissances sur cette transformation. Cette recherche est donc autant une recherche impliquée qu'une recherche appliquée (Saint-Luc, 2014). C'est l'une des caractéristiques de la recherche-action.

En soi, la recherche-action est également assez proche du sujet des sciences citoyennes et du *crowdsourcing* dans la mesure où il s'agit d'*empowerment,* mais aussi de dépasser le clivage entre pratiques profanes d'amateurs et théories scientifiques de spécialistes. C'est donc une forme de recherche particulièrement adaptée à ce sujet.

Au delà de cette observation participante qui est le résultat d'un parcours professionnel, une expérimentation empirique a été conduite autour du projet Numalire de numérisation à la demande par *crowdfunding*. Elle occupe une place centrale dans cette partie consacrée aux apports expérimentaux de la thèse. Si le pragmatisme a principalement guidé cette expérimentation, nous avons cherché à objectiver les résultats obtenus grâce à des enquêtes auprès des usagers et des entretiens auprès des bibliothèques partenaires. Une enquête, qu'on retrouvera en

annexes, a été conduite auprès de 380 usagers inscrits sur le site numalire.com. Elle a donné lieu à 118 réponses (31 %) dont les analyses sont livrées dans la partie expérimentale de la thèse. Des entretiens téléphoniques semi directifs ont été conduits auprès des responsables de bibliothèques partenaires. Les principales informations obtenues dans le cadre de ces échanges téléphoniques ont donné lieu à une synthèse qui a également été rapportée dans la partie expérimentale.

L'expérimentation a porté principalement sur le *crowdfunding*, une forme bien spécifique de *crowdsourcing*, une forme qui lorsqu'elle a été appliquée aux bibliothèques numériques a encore très peu été étudiée dans la littérature. C'est pour cette raison, mais aussi afin de participer à la création d'une entreprise que nous avons choisi de porter l'essentiel de nos efforts expérimentaux sur le *crowdfunding*.

0.4- Rédaction en Google Doc : une thèse sur le *crowdsourcing* qui bénéficie du *crowdsourcing* ?

L'exercice doctoral de rédaction d'une thèse est, par nature, individuel. Nous avons toutefois trouvé intéressant, pour une thèse portant sur le *crowdsourcing*, et avec une démarche assez proche de hackyourphd.org, de mettre en œuvre une démarche participative et d'expérimenter la rédaction d'une thèse exposée en temps réel sur le web et sur laquelle le directeur de thèse, les partenaires de l'expérimentation Numalire et des internautes pouvaient ajouter à tout moment leurs commentaires, leurs suggestions et propositions de corrections. Au total, nous avons ainsi bénéficié, de la part de personnes de connaissance et de collègues de travail, d'une idée, de 35 pistes de lectures, et de 101 corrections de fautes d'orthographe, de frappe ou de sens. Ces contributeurs ont été remerciés dans la thèse mais aussi sur un site web, www.bibliotheque-numerique.fr, qui a été développé afin d'exposer, aux commentaires, sur le web les parties de la thèse rédigées sur Google Doc lorsqu'elles étaient suffisamment achevées et afin d'afficher le classement des contributeurs.

Afin d'attirer, au-delà, des contributeurs de connaissance, d'éventuels spécialistes du sujet et d'aboutir à d'éventuelles collaborations de publication, la bibliographie de la thèse a été mise en ligne sur le site afin d'être plus facilement repérée par les auteurs grâce au référencement, par les moteurs de recherche, du contenu du site. Le site est bien référencé et apparaît en première page d'une requête Google avec les mots clés bibliothèque et *crowdsourcing*, par exemple. Néanmoins, la majeure partie des auteurs n'est pas de langue française et n'ont probablement pas pu prendre connaissance du contenu de la thèse. Les statistiques de consultation, extraites de Google Analytics, indiquent que le site, bien qu'en langue française, a majoritairement été consulté aux USA.

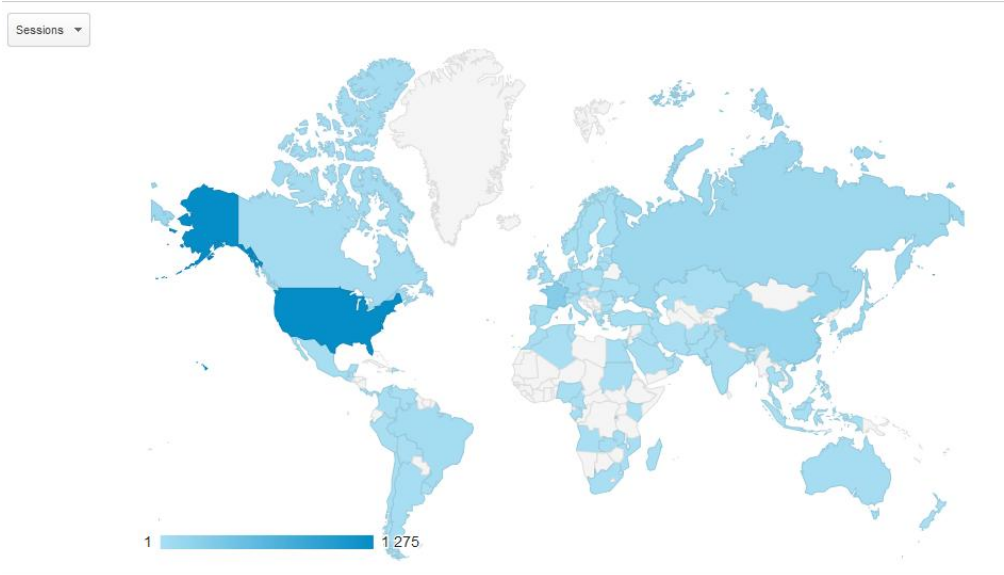


Figure 6. Origine géographique des visites sur le site bibliotheque-numerique.fr
d'après Google Analytics

Entre le 29 avril et le 30 novembre 2015, le site a généré 3 288 visites de 3009 visiteurs dont : headquarters usaischeadquarters usaisc (27 sessions), ecole normale superieure (24 sessions), ford motor company (19 sessions), massachusetts institute of technology (7 sessions), amazon technologies inc. (3 sessions)

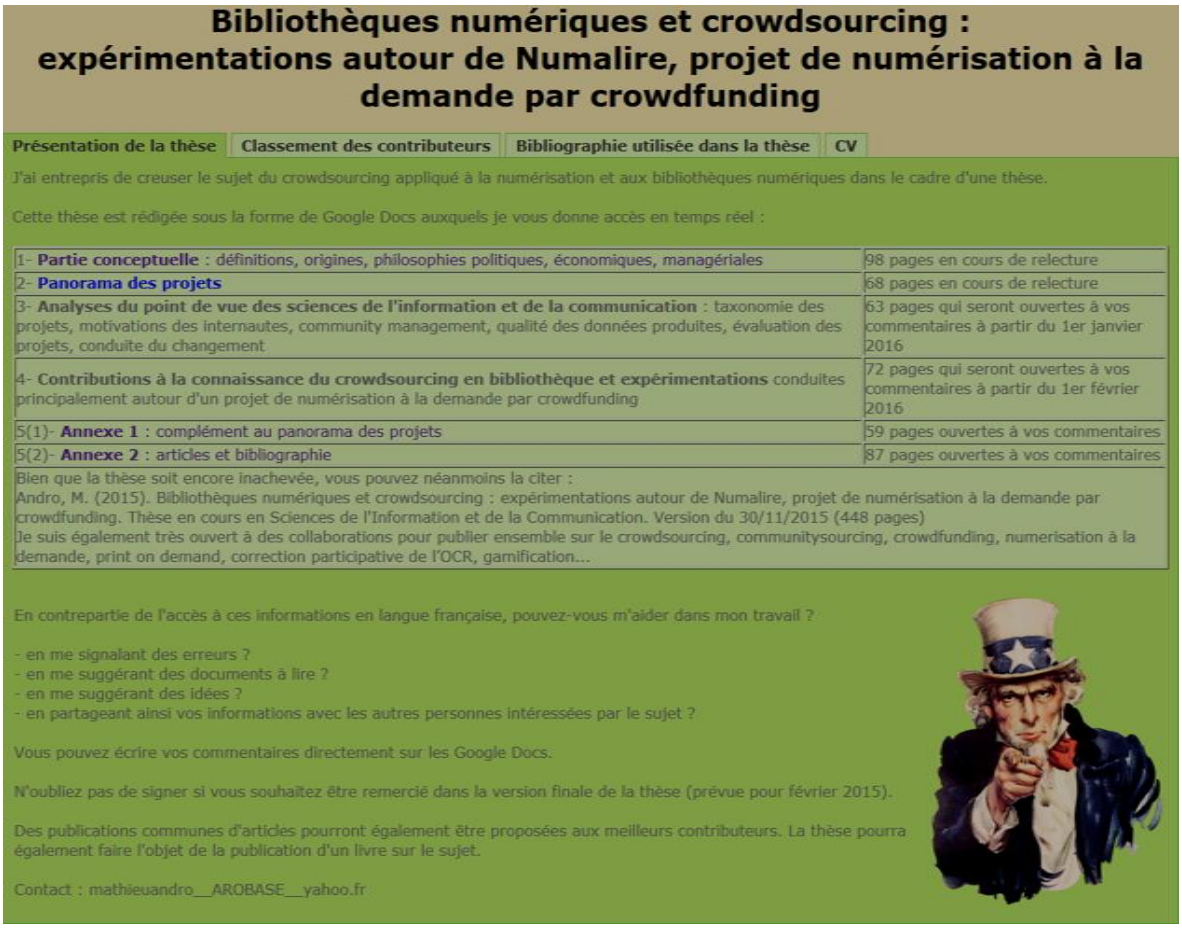


Figure 7. Capture d'écran du site www.bibliotheque-numerique.fr

Le risque que des éléments originaux de la thèse puissent être publiés par une personne à la moralité limitée a été écarté en ne diffusant ses parties qu'après qu'elles aient fait l'objet de publications d'articles. L'antériorité et la paternité de ces éléments aurait pu toutefois être prouvée assez facilement dans la mesure où l'historique des Google Docs peut être aisément retrouvé et établi. Nous avons aussi préféré parier sur l'esprit de confiance et de collaboration. Cette manière d'opérer n'est pas sans rappeler celle des brevets. Il faut diffuser l'invention pour en protéger la paternité.

Chapitre 1- Introduction conceptuelle sur la notion de *crowdsourcing* en bibliothèque : un nouveau paradigme ?

1.1- Un modèle économique en plein essor

1.1.1- Ce qui a rendu possible ce nouveau modèle économique

Les internautes sont de plus en plus nombreux et le temps qu'ils passent à surfer sur Internet est croissant. En France, il y avait déjà, en 2012, 40 millions d'internautes (contre 16 millions en 2002) et 19 millions d'entre eux disposaient d'un smartphone leur permettant d'accéder au web partout et à tout moment, d'après une étude de l'agence Médiamétrie du 14 mars 2012. En janvier 2014, ce sont plus de 54 millions de français (soit 83 % de la population) qui disposeraient d'un accès internet. Chacun de ces internautes consacrerait 4 heures et 7 minutes de son temps chaque jour à surfer sur le web, sans compter les connexions mobiles via smartphone qui occuperaient les français 58 minutes par jour en moyenne, d'après une étude de l'agence We are social. Ces chiffres, en progression permanente, classent la France parmi les pays les plus connectés au monde. Si on considère que chaque internaute consacre environ 4 heures de son temps sur Internet. Chaque jour, en France, le temps de connexion sur Internet avoisine donc les 160 000 000 d'heures, soit l'équivalent du temps de travail d'une équipe de 10 000 salariés travaillant pendant 10 ans. L'encyclopédie Wikipédia aurait, quant à elle, nécessité 100 millions d'heures de temps cumulé pour être construite, c'est-à-dire beaucoup moins que le temps cumulé par les seuls internautes de France sur une seule journée. Par ailleurs, comme l'affirmait Clay Shirky en 2008 à la Wiki-Conference NYC du 28 août 2008, si les américains, qui regardent la télévision, tous les ans, 200 billions d'heures, consacraient plutôt ce temps à des activités créatives, ils pourraient créer 2000 projets comme Wikipédia chaque année au lieu de regarder la télévision. Luis Von Ahn, au cours d'une conférence Ted de 2011[1], affirmait, quant à lui, qu'avec 100 000 hommes l'humanité était parvenue à construire des pyramides et à creuser le canal de

[1]https://www.ted.com/talks/luis_von_ahn_massive_scale_online_collaboration (consulté le 23 juin 2016)

Panama et que grâce à Internet et aux réseaux sociaux, il devenait possible d'en réunir 750 millions comme avec un projet de correction de l'OCR comme reCAPTCHA. Un fabuleux "réservoir de bonnes volontés" est donc potentiellement disponible pour les institutions culturelles si elles savent en tirer profit.

Les modèles participatifs sont nés avec le développement du web 2.0, et le terme aurait été inventé par DiNucci en 1999 (Nguyen, 2012) ou par Dale Dougherty en 2004 (Sarrouy, 2014) et popularisé par Tim O'Reilly en 2005 (Trainor, 2008). Le *crowdsourcing* permet désormais aux internautes de ne plus se contenter de consommer passivement le contenu du web sur un modèle de diffusion hiérarchique, unilatéral et statique (web 1.0) mais de participer activement à son développement. La diffusion d'information est devenue réciproque, interactive et dynamique. L'internaute cesse donc d'être un consommateur, un lecteur et un récepteur passif qui se contente de surfer pour devenir un producteur, un auteur, un émetteur actif d'information, un contributeur qui peut participer à la rédaction et à la modification de contenus sur le web (commentaires, tags, wikis, réseaux sociaux…) et à la production de données et de métadonnées. L'autorité des données a ainsi été déplacée du serveur vers le client (Bainbridge, 2012). Et, comme le souligne l'expert en télécommunication Benjamin Bayart, si l'imprimerie a appris au peuple à lire, Internet lui apprend aujourd'hui à écrire[2].

Bien avant le web 2.0, l'invention du "libre service" qui permettait au consommateur l'accès direct aux marchandises sans passer par l'intermédiaire du vendeur et qui s'est appliqué en bibliothèque sous la forme des collections en libre accès a été une première forme d'intégration du consommateur dans le processus de production. Ce modèle économique a été inventé par Aristide Boucicaut avec le magasin "le Bon Marché" dont le slogan était "libre accès, libre toucher" laissant aux clientes, décrites dans "Le bonheur des dames" de Zola, la possibilité d'accéder, activement et librement, sans l'intermédiaire d'un marchand, aux marchandises et, in fine, de prendre en charge, une part de l'activité du marchand et du magasinier. De manière générale, la production semble avoir ainsi perdu

[2]http://www.gameblog.fr/blogs/poufy/p_58428_l-imprimerie-aura-permis-au-peuple-de-lire-internet-lui-a-pe (consulté le 23 juin 2016)

progressivement la place centrale qu'elle occupait au profit de la consommation et de la société de consommation qui s'est développée après la seconde guerre mondiale.

Plus tard, le modèle du "just in time", développé chez Toyota, a consisté à produire "à la demande" du client afin d'éviter les stocks d'invendus en produisant en flux tendus une offre de manière synchronisée avec la demande et tirée par la demande. Ce modèle de "fabrication sans gaspi", de "fabrication maigre" ou de "fabrication sans gras" consistant à produire ce dont on a strictement besoin, avec les moyens justes nécessaires, au moment où on en a besoin et à moindre coût ont permis au producteur d'externaliser la décision de lancer la production auprès du consommateur. Ce modèle est né de la difficulté des échoppes japonaises à stocker faute d'espace suffisant et de la nécessité de n'approvisionner que lorsque le stock faisait défaut. Il a également largement été inspiré par le fonctionnement des supermarchés. De la même manière, la chaîne de commerces vestimentaires Zara ne conserve ainsi qu'un seul mois d'inventaire et adapte ainsi mieux sa production aux tendances du marché, produisant des modèles en fonction des ventes. (Surowiecki, 2008). La publicité elle-même participe à l'intégration du consommateur dans le procès de production. En effet, quand on visualise une émission de télévision ou un site web, on produit des statistiques et des données, ou quand on visualise des publicités, on produit aussi de la valeur. On peut ainsi parler d'une économie de l'attention (Citton, 2014). Le choix de visiter tel site ou tel site pourrait donc s'apparenter à un vote, un vote qui participe à la production et aux revenus des producteurs. Ce modèle a trouvé son application, en bibliothèque, dans la numérisation à la demande par financements participatifs (*crowdfunding*) et dans l'impression à la demande qui seront largement abordés dans la thèse.

Le *crowdsourcing* prolonge aujourd'hui ce mouvement, relativement ancien, d'intégration du consommateur dans le processus de production. Il a été rendu possible par le développement des technologies du web 2.0. Né d'une évolution culturelle vers des approches plus participatives et plus collaboratives, le *crowdsourcing* a été techniquement rendu possible, par le web 2.0, c'est-à-dire la possibilité de faire travailler à distance un grand nombre de personnes, ayant du

temps disponible sur le web, sur des projets communs. Il s'est particulièrement inspiré du mode de fonctionnement des communautés de développeurs du logiciel libre. En faisant appel à la foule des internautes, on peut réaliser en peu de temps des tâches qui auraient été impossibles à réaliser et même à imaginer ou qui auraient demandé énormément de temps auparavant. En somme, le *crowdsourcing* "est un moyen de trouver une aiguille dans une botte de foin" selon l'expression de (Lebraty, 2015). (Sagot, 2011) parle de "myriadisation du travail parcellisé" et de microworking. On pourrait également parler de « tâchification » du travail. Le *crowdsourcing* n'est pas sans rappeler la construction des cathédrales qui ont nécessité la capacité de "penser grand", de déléguer et d'organiser toutes les tâches et surtout, de mobiliser un grand nombre de personnes autour d'une vision et d'un objectif communs, comme le rappelle (Levi, 2014). C'est aussi, pour prendre un exemple moins ancien, ce qu'Alfred Sloan qualifiait de « management de groupe » au sein de General Motors, c'est à dire la sollicitation de nombreux collaborateurs pour prendre les décisions les plus importantes.

Nous illsutrons cette idée avec les œuvres d'art contemporain ci-après.

Figure 8. L'œuvre d'art Ten Thousands Cents[3]

[3]Cette œuvre d'art contemporain créée par Aaron Koblin a été réalisée par 1000 personnes ayant travaillé séparément, sous Amazon Mechanical Turk Marketplace (AMT), à la réalisation d'un millième du billet de 100 dollars sans avoir conscience du but final.

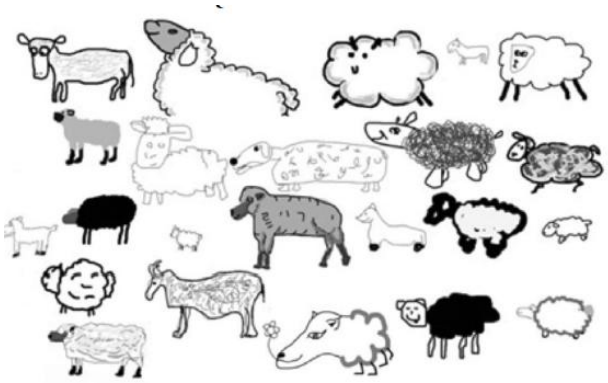


Figure 9. Œuvre d'art de juxtaposition de moutons[4]

Outre l'art, le *crowdsourcing* trouve déjà des applications dans de nombreux domaines. Par exemple, dans le domaine de la vidéo, YouTube ou DailyMotion ne fonctionnent que grâce aux contenus miss en ligne par des internautes. Le *crowdsourcing* a également trouvé des applications dans la musique, la politique, la mode, la banque, le tourisme, l'innovation, la cartographie, la recherche d'avions disparus, la médecine, la recherche scientifique, l'édition, la traduction ou encore le journalisme. Le recours aux foules est également d'actualité dans le domaine des GLAM (Galleries, Libraries, Archives and Museums) et des bibliothèques numériques en particulier qui font l'objet de cette thèse.

1.1.2- et son application aux bibliothèques numériques

Concernant les bibliothèques, avec la numérisation et la diffusion sur le web de leurs collections, elles se retrouvent, sur le web, dans un même espace que leurs usagers. Cette situation rend possible de multiples synergies et collaborations. Du coté des institutions culturelles, la masse des contenus qu'elles ont rendus disponibles sur le web a connu une croissance exponentielle et le "travail de fourmi" ne manque pas pour indexer, qualifier et corriger ces contenus. Mais leurs budgets et leurs effectifs ont connu un mouvement inverse et leur font souvent cruellement défaut. Cet état de fait rend l'achèvement de nombreux objectifs irréalisables et la réalisation d'autres projets inimaginables sans une aide

[4]Ces moutons ont été dessinés par des internautes rémunérés sur la même plateforme AMT et ont été rassemblés par l'artiste Aaron Koblin (http://www.thesheepmarket.com)

extérieure. Et, du coté des publics réels ou virtuels de ces institutions, ils se contentent de moins en moins du rôle de passifs consommateurs de données culturelles et souhaitent de plus en plus s'engager au service du patrimoine et de la culture. Dans les institutions culturelles, l'idée d'une ouverture à l'interaction avec un public participant et des bénévoles a largement précédé l'émergence du web 2.0. Mais le web relationnel a largement favorisé l'émergence d'une culture participative dont se nourrit le modèle du *crowdsourcing* en bibliothèque.

Dans les bibliothèques numériques, le *crowdsourcing* permet ainsi d'achever des tâches qu'il serait impossible de prendre en charge sans l'aide des internautes bénévoles, faute de moyens financiers et humains, d'améliorer, par exemple, la qualité des métadonnées ou de les enrichir (commentaires, tags, analyses...), de bénéficier des connaissances et des compétences des érudits, de développer des communautés autour des projets, d'augmenter la fréquentation des ressources produites, de mieux sensibiliser le grand public à la conservation du patrimoine commun, de susciter plus d'interactions, des idées innovantes et des collaborations. Par exemple, dans le public en ligne, il y a peut être quelqu'un qui saurait identifier cette église photographiée, un érudit pourrait apporter des renseignements sur sa construction, son histoire, un villageois âgé identifier une personne sur la photo… Les connaissances dont disposent les équipes de bibliothécaires sont bien trop limitées pour pouvoir répondre à toutes ces questions. Celles qui sont présentes dans la foule des internautes sont sans limites.

Le British Museum l'a bien compris. Le 3 août 2015, il publie sur britishlibrary.typepad.co.uk, un appel aux internautes sous le titre "aidez nous à déchiffrer cette inscription". Entre le 3 et le 18 août 2015, le billet a été partagé près de 32 000 fois, et a généré plus de 11 000 partages sur Facebook et 9000 tweets, mais aussi 115 commentaires directement sur le blog entre le 3 et le 10 août.

Figure 10. Épée du 13è siècle dont la photographie a été publiée par la British Library[5]

Afin de mobiliser les internautes, les institutions culturelles disposent, en effet, de solides atouts. Elles disposent déjà souvent d'une solide expérience dans la mobilisation de bénévoles et dans l'organisation de concours, de réunions de lecteurs, d'événements, et même dans l'"adoption" de livres dont l'achat a été financé par des lecteurs ou des mécènes. Par ailleurs, ces institutions jouissent d'une bonne image auprès des populations et apparaissent comme dignes de confiance, au service de l'intérêt général et n'ayant pas de buts lucratifs mais des finalités culturelles. Ces finalités sont donc susceptibles d'attirer des bénévoles et de susciter des contributions.

Le *crowdsourcing* au service des bibliothèques numériques est également le moyen de passer d'un travail parfois ingrat demandé à un salarié particulier à une activité valorisante proposée à un groupe indéfini d'internautes bénévoles et de "petites mains" souhaitant activement contribuer au développement du web culturel. Les documents numérisés et mis en ligne font ainsi l'objet d'une

[5] http://britishlibrary.typepad.co.uk/digitisedmanuscripts/2015/08/help-us-decipher-this-inscription.html (consulté le 23 juin 2016)

redocumentarisation participative, d'une remédiation permettant de traiter de nouveau et collectivement des collections de documents, en faisant appel tantôt au témoignage et à la mémoire, tantôt à l'expertise et à l'érudition des internautes. Les collections sont ainsi revisitées, réinventées, ré-imaginées.

1.1.3- et suscite l'intérêt croissant des politiques, des internautes et des universitaires

Le succès des projets de *crowdsourcing* et l'intérêt porté pour ces projets par les internautes, par les politiques, et par la recherche académique, sont croissants. Comme le rapporte (Sarrouy, 2014), une étude de 2011 de massolutions.com estimait le marché du *crowdsourcing* à plus de 300 000 000 de dollars avec un taux de croissance supérieur à 75 % entre 2010 et 2011. Une autre étude de McKinsey évaluait en 2012 à 25 % les gains en productivité issus des médias sociaux et des plateformes de *crowdsourcing* dans les biens de consommation, les services financiers, la production avancée et les services professionnels. Enfin, le cabinet Gartner prévoyait, quant à lui, à la fin 2013, qu'à l'horizon 2017, plus de la moitié des producteurs de biens de consommation baseront plus de 75 % de leur Recherche & Développements sur le *crowdsourcing*. Dans le domaine des sciences citoyennes pour la biodiversité seule, des chercheurs de l'Université de Washington estiment que les contributions en nature des 1,3 à 2,3 millions de bénévoles auraient une valeur économique supérieure à 2,5 milliards de dollars.

Le *crowdfunding*, en particulier, aurait permis de financer un million de projet en 2012 et de lever 2 milliards d'euros (Onnée, 2013). Bien que le financement de projets par des particuliers n'ait en soi aucun caractère nouveau, Internet permet de faciliter et de donner une envergure nouvelle au financement participatif qui représentait déjà un marché de 3 milliards de dollars dans le monde en 2012 et dont la croissance est exponentielle. Entre 2007 et 2015, plus d'un million de français, soit 7 % d'entre eux, aurait ainsi contribué, plaçant la France en championne d'Europe du financement participatif, d'après le journal Le Monde du 1er juin 2015. L'argent collecté, par exemple, par la fondation Wikimédia France

par les internautes français était, à titre d'exemple, de 2 987 935 € pour 95 554 dons entre janvier 2009 et décembre 2012. En 2011, dans le monde, 15,3 millions d'euros avaient déjà été versés à Wikimédia par un million de donateurs afin de financer les 95 salariés de Wikipédia dans le monde, mais aussi la maintenance, l'achat de serveurs et le développement de nouvelles spécifications.

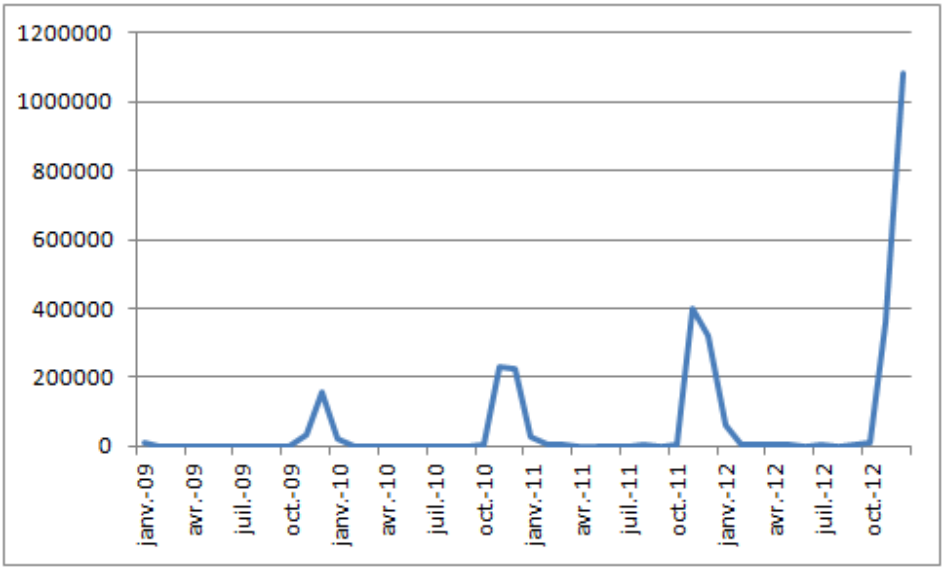


Figure 11. Evolution des dons de particulier à Wikimédia France entre janvier 2009 et octobre 2012

En utilisant le service Google Trends, c'est à dire les traces laissées involontairement[6] par les internautes qui interrogent Google, on constate aussi que, au-delà des politiques, de plus en plus d'internautes ont saisi le mot *crowdsourcing*, qui n'a guère de traductions courantes dans les autres langues que l'anglais, dans le moteur de recherche Google à partir de 2006, lorsque le terme a été popularisé par Jeff Howe. Sur une base 100, les pays dont les internautes ont fait le plus de recherches contenant le mot *crowdsourcing* sont, dans l'ordre, les Pays-Bas (100), le Portugal (60), l'Allemagne (60), l'Espagne (56), Singapour (55), l'Autriche (54),

[6] On pourrait parler dans ce cas de « crowdsourcing implicite » c'est à dire de contribution involontaire comme nous le verrons ultérieurement.

la Suisse (54), les États-Unis (48), le Brésil (43) et le Danemark (38), le Royaume Uni (31)...

La France (23), quant à elle arrive loin derrière et bien après les internautes américains puis allemands qui, chronologiquement, se sont rapidement intéressés au phénomène :

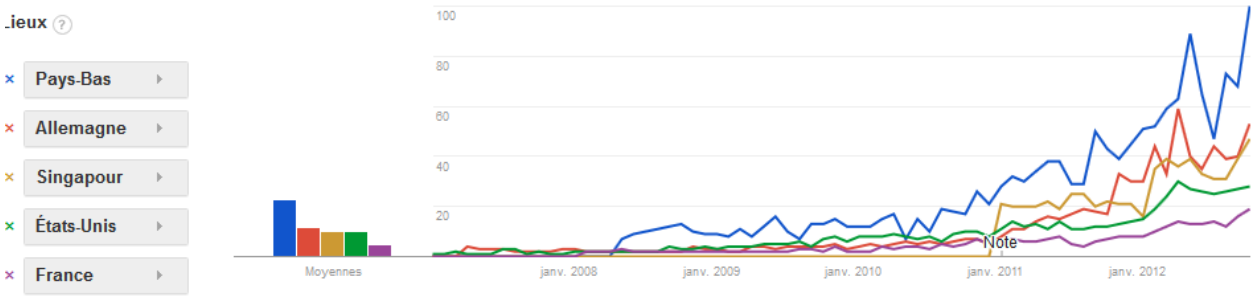


Figure 12. Evolution du nombre de recherche du mot "*crowdsourcing*" dans Google selon les pays d'après Google Trends

Le relativement faible intérêt de la France pour le *crowdsourcing* s'observe tout particulièrement dans le domaine de la numérisation et des bibliothèques numériques. L'enquête sur les projets de *crowdsourcing* appliqué aux bibliothèques la plus exhaustive qui ait été rencontrée dans la littérature a été menée par l'OCLC (Smith-Yoshimura, 2011), elle montre que, parmi les projets recensés dont les responsables ont été sollicités pour l'enquête, 60 % sont américains, 19 % australiens, 10 % anglais, 5 % néo-zélandais et 7 % seulement dans d'autres pays du monde.

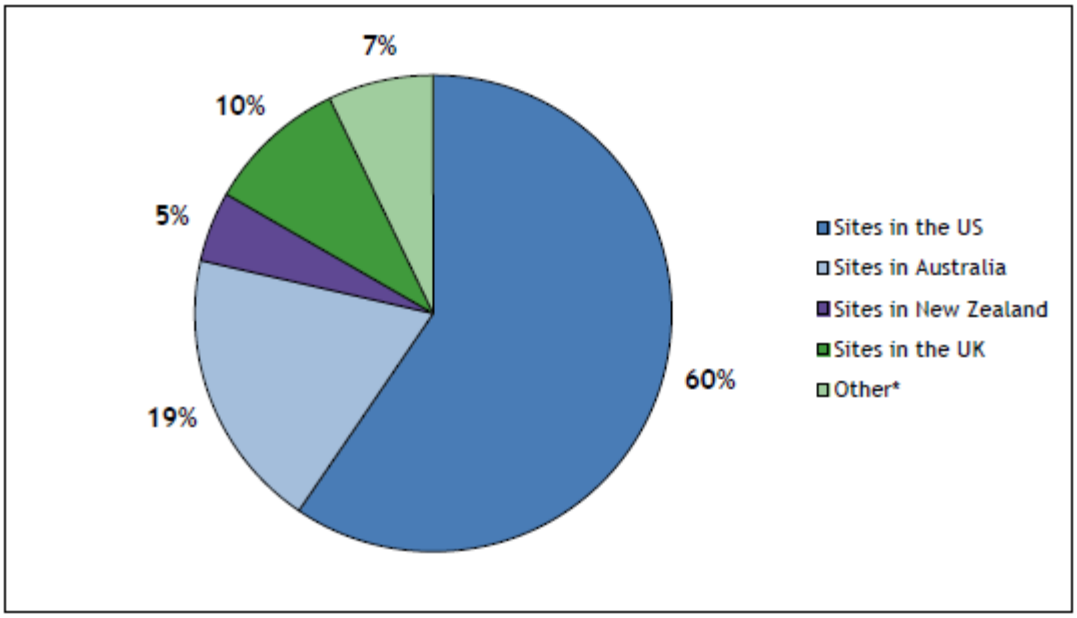


Figure 13. Pays représentés dans l'enquête conduite par l'OCLC à propos des métadonnées sociales (d'après Smith-Yoshimura, 2011)

Le *crowdsourcing* appliqué aux projets de numérisation pourrait donc n'être considéré que comme un phénomène purement anglo-saxon. La France et ses projets ne sont d'ailleurs mentionnés nulle part dans les 4 volumes et près de 350 pages de l'enquête.

Pourtant, l'intérêt de la recherche scientifique mondiale pour le phénomène du *crowdsourcing* est également croissant notamment pour ce qui concerne ses applications à la numérisation du patrimoine conservé dans les bibliothèques. Ce constat peut être révélé en observant, par exemple, le nombre d'articles indexés dans Google Scholar concernant ce sujet spécifique :

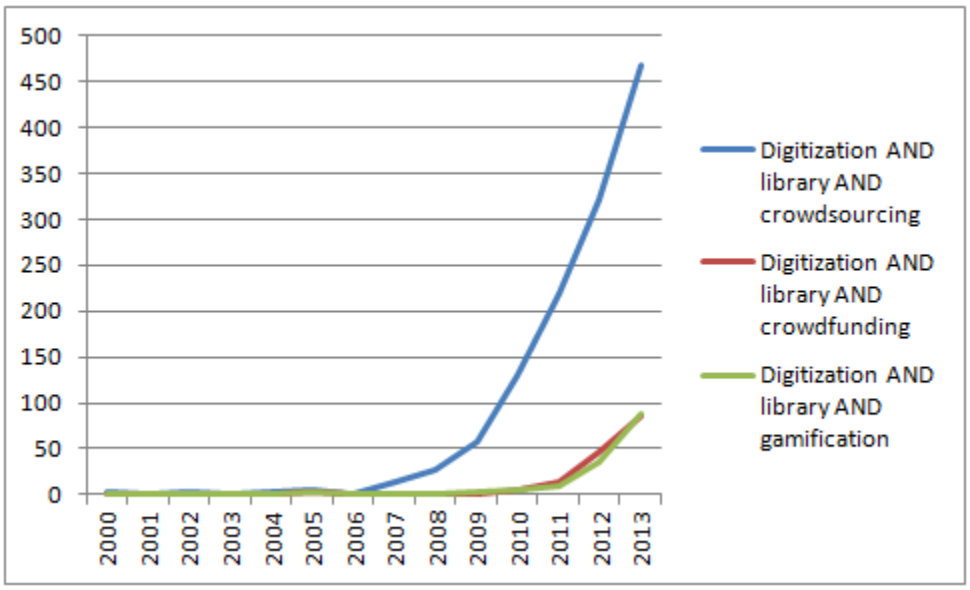


Figure 14. Evolution du nombre de publications indexés par Google Scholar sur le *crowdsourcing* appliqué à la numérisation des bibliothèques

Mais on observe toutefois, qu'à l'instar du relativement faible nombre d'internautes et de bibliothèques françaises s'intéressant au sujet comme nous l'avons précédemment mis en évidence, un nombre encore très restreint de publications professionnelles ou universitaires en France se sont penchées sur le sujet du *crowdsourcing* appliqué à la numérisation. La bibliographie la plus exhaustive possible qui a été produite dans le cadre de cette thèse ne laisse apparaître qu'un nombre très faible d'articles rédigés par des auteurs français et il aura fallu attendre février 2013 pour qu'une première étude française sur le sujet soit publiée, bien après le commencement de cette thèse, par la BnF, dans le cadre du projet Ozalid (Moirez, 2013).

Nous espérons donc que cette thèse permettra de combler des lacunes et, outre les apports conceptuels et expérimentaux, donnera accès, aux bibliothécaires et aux universitaires en sciences de l'information et de la communication, à un état de l'art et à une synthèse en langue française de la littérature internationale produite sur le sujet.

1.2- Une origine, une traduction, une définition, et un périmètre du *crowdsourcing*

Le crowdsourcing a longtemps été une pratique professionnelle pragmatique bien avant d'être conceptualisé et de devenir un sujet de recherche académique. Dans ces conditions, son origine, sa traduction, sa définition et son périmètre peuvent être malaisés à établir. Avant de devenir un « *buzzword* », le terme de *crowdsourcing* a été utilisé pour la première fois par Jeff Howe dans le titre d'un article publié dans Wired Magazine en juin 2006 et qui s'intitulait "the rise of *crowdsourcing*". D'après (Schenk, 2010), le terme aurait cependant été utilisé précédemment par un internaute anonyme sur un forum. D'autres auteurs préfèrent parler de "Open work" ou de "fair trade work" qui pourraient être traduits par "travail ouvert" ou par "travail libre".

Lorsqu'il est traduit en français, le terme de *crowdsourcing* l'est le plus souvent, littéralement, sous l'expression de "approvisionnement par la foule" ou plus rarement de "externalisation distribuée à grande échelle". France Terme (http://www.culture.fr/franceterme consulté le 23 juin 2016) suggère l'emploi de "production participative". Ce terme a été officialisé par le Journal Officiel de la République Française le 15 août 2014 et définit comme un "*mode de réalisation d'un projet ou d'un produit faisant appel aux contributions d'un grand nombre de personnes, généralement des internautes*" (rapporté par Néroulidis, 2015). L'Office québécois de la langue française préconise, quant à lui, le terme de "externalisation ouverte" qui semble plus judicieux mais qui ne semble malheureusement pas connaître, pour le moment, plus de succès dans la littérature francophone. Dans la littérature allemande, le terme synonyme de création interactive de valeur et qu'on retrouve sous la forme de "interaktive Wertschöpfung" (Kleemann, 2008) ne semble pas avoir eu non plus un très grand succès.

Dans le cas de projets de bibliothèques numériques dont les réels contributeurs ne sont qu'une minorité active de bénévoles et ne sauraient, en tous cas, être assimilés à une foule, certains auteurs préfèrent employer le terme de

“*nichesourcing*” ou de “*community sourcing*” préférant le terme de communauté déterminée à celui de foule plus indéterminée. Il s’agit, en effet, moins d’utiliser le public que de recruter des bénévoles motivés dans un esprit de collaboration, de co-création et de co-construction. Cette idée renvoie à celle énoncée par Jakob Nielsen[7], selon laquelle 80 % des internautes sont de passifs consommateurs et 20 % d’entre eux, d’actifs contributeurs et producteurs de contenus sur le web. D’après Holly Goodier[8], ces proportions auraient évolué depuis et seraient désormais plutôt de 25 % d’inactifs, 45 % qui commentent et enrichissent et 30 % d’internautes qui produisent des contenus. Concernant les bibliothèques numériques, le terme de « *community sourcing* » nous semble le plus judicieux. Nous emploierons néanmoins le terme de *crowdsourcing*, qui est plus courant, et permettra de rendre plus intelligible notre propos et nous permettra d’éviter le recours à un jargon complexe qui masque trop souvent la faiblesse des contenus.

Des auteurs (Estellés-Arolas, 2012), dont le travail fait autorité, ont cherché à travailler spécifiquement sur la question de la définition du *crowdsourcing* en collectant, dans la littérature, la diversité des définitions qui y ont été trouvées. Pas moins de 40 citations au sein de 32 articles publiés entre 2006 et 2011 ont ainsi été collectées dans cette étude qui a catégorisé les différents éléments nécessaire à la construction d’une définition de synthèse :

[7]https://www.nngroup.com/articles/community-is-dead-long-live-mega-collaboration (consulté le 23 juin 2016)

[8]http://www.bbc.co.uk/blogs/bbcinternet/2012/05/bbc_online_briefing_spring_201_1.html (consulté le 23 juin 2016)

Qui forme la foule ?	Des amateurs
Que fait la foule ?	Elle accomplit volontairement et consciemment des tâches et des microtâches pour résoudre des problèmes
Qu'obtient la foule en retour ?	la distraction, le plaisir, le développement de compétences, d'expériences, de connaissances, le partage de connaissances, l'amour d'une communauté, des récompenses économiques, une reconnaissance sociale, une meilleure estime de soi
Qui est l'initiateur ?	des sociétés publiques ou privées
De quel type de processus s'agit-il ?	Un mode de production, un modèle économique, l'externalisation participative d'une tâche après un appel ouvert à tous
Quel médium est utilisé ?	Internet

A partir de ces éléments, voici la définition en anglais à laquelle ces auteurs aboutissent :

> *« is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken.»*
>
> (d'après Estellés-Arolas, 2012)

Cette définition pourrait être traduite en français, par nos soins, de la manière suivante :

> « Le *crowdsourcing* est un type d'activité participative en ligne pour laquelle un individu, une institution, une organisation à but non lucratif ou une société propose à un groupe hétérogène d'individus de compétences variables, à travers un appel ouvert, la volontaire sous-traitance d'une tâche. L'externalisation de cette tâche, de complexité et de modularité variable, et pour laquelle, une foule d'internautes pourraient participer, apportant leur travail, leur argent, leurs connaissances et / ou expériences bénéfice toujours mutuellement à tous les associés. Les usagers recevront la satisfaction d'un type donné de besoins, qu'ils soient économiques, basés sur la reconnaissance sociale, l'estime de soi, ou le développement de compétences individuelles, dans la mesure où les commanditaires obtiendront et utiliseront à leur avantage ce que les participants ont apporté, dont la forme dépendra du type d'activité externalisée. »

La question du caractère volontaire ou involontaire de la participation des internautes pourrait toutefois être discutée. En effet, si on considère que cette contribution est nécessairement volontaire comme l'affirme cette définition, on exclut du champ du *crowdsourcing*, des sites comme YouTube, la correction de l'OCR grâce à ReCAPTCHA et une grande partie des projets récoltant les contributions des internautes sous la forme de jeux (*gamification*). Si on considère que cette contribution n'est pas nécessairement volontaire, le périmètre s'en trouve effectivement considérablement élargi. Dans tous les cas, exclure du champ du *crowdsourcing* les formes de participation non pleinement conscientes mériterait au moins d'être justifié, ce qui semble difficile. Peut être est-il donc préférable, de notre point de vue, de parler plutôt de *crowdsourcing* explicite lorsque la contribution des internautes est volontaire et de *crowdsourcing* implicite (ou *crowdsourcing* involontaire ou encore *crowdsourcing* passif) quand elle ne l'est pas. (Harris, 2013). (Renault, 2014bis) estime ainsi également que cette définition est

quelque peu naïve car nombreux sont les contributeurs du *crowdsourcing* qui n'ont pas conscience de contribuer. Néanmoins, on pourrait considérer le *crowdsourcing* implicite comme une sorte de trahison du *crowdsourcing* initialement pensé comme un moyen de ré-humaniser le web et le considérer comme une revanche du web marchand sur le pouvoir des internautes. En effet, avec le *crowdsourcing* implicite, le risque est grand d'instrumentaliser les citoyens au profit de lobbys, de considérer les internautes et les traces qu'ils laissent sur le web notamment avec leurs appareils connectés comme de simples moyens sans les associer aux projets (Le Crosnier, 2013).

(Schenk, 2012) a également fait le choix de faire rentrer cette forme de *crowdsourcing* dans sa typologie en la qualifiant de « non volontaire » et en la rapprochant de la notion d'externalité. Le *crowdsourcing implicite* pourrait, en effet, être considéré à la lumière de la notion d'externalité positive (ou économie externe). Ainsi, par les traces qu'ils laissent ou par leur travail inconscient, les internautes, en tant qu'agents économiques rendent un service économique valorisable par d'autres agents sans en être rémunérés. Ainsi Google bénéficie du travail des internautes qui corrigent ses textes océrisés sans le savoir en ressaisissant des reCAPTCHA afin de prouver qu'ils ne sont pas de robots et créer des comptes sur des sites web. De la même manière, un apiculteur bénéficie implicitement du travail d'un arboriculteur grâce aux fleurs des arbres que ce dernier cultive et qui pourront être butinées par les abeilles de ce premier sans compensation financière. En contrepartie, les abeilles vont également favoriser la fécondation des arbres. (Meade, 1952). Dans le cas de Google Books, la firme pourrait effectivement remercier ses contributeurs involontaires ou être taxée pour ce travail dissimulé. Néanmoins, on pourrait également considérer que l'amélioration par les internautes de la qualité de textes accessibles gratuitement aux internautes leur bénéficie directement en retour.

Toutes ces considérations étant prises en compte, le *crowdsourcing* pourrait donc plutôt être défini, à l'issue de la lecture d'un ensemble représentatif de publications, et selon la définition que nous proposons, comme :

Le *crowdsourcing* est une forme d'externalisation qui permet l'apport de travail, d'argent (« *crowdfunding* »), de compétences, de connaissances, d'intelligence, de créativité ou d'expérience, par engagement volontaire (« *crowdsourcing explicite* ») ou involontaire (« *crowdsourcing implicite* ») d'internautes. Cette externalisation fait suite à l'appel d'un individu, d'une institution ou d'une organisation. Les internautes bénéficieront, en échange de leur apport, d'une reconnaissance sociale, d'une expérience, de l'acquisition de compétences, de récompenses ou d'une rémunération (« *crowdsourcing rémunéré* »). Ils peuvent aussi agir pour améliorer l'estime de soi, par distraction, par plaisir, par amour pour une communauté ou par altruisme désintéressé.

Cette définition étant proposée, afin de bien appréhender ce qu'est le *crowdsourcing*, il apparaît nécessaire d'en dessiner le périmètre en énonçant ce que le *crowdsourcing* n'est pas. En effet, la notion de *crowdsourcing* est assez voisine, par exemple, de celle de *human computation* qui évoque la possibilité de faire faire aux humains et à leur intelligence collective des tâches que les programmes informatiques sont encore incapables d'effectuer de manière automatisée. Néanmoins, le *crowdsourcing* s'en distingue par des outils et des tâches plus simples et moins sophistiqués et par des règles de contribution construites de manière plus collaborative.

Avec le *crowdsourcing*, la force de la foule résiderait d'avantage dans l'agrégat d'idées indépendantes que dans leur collaboration (Szoniecky, 2012). Il se distingue donc aussi de l'intelligence collective.

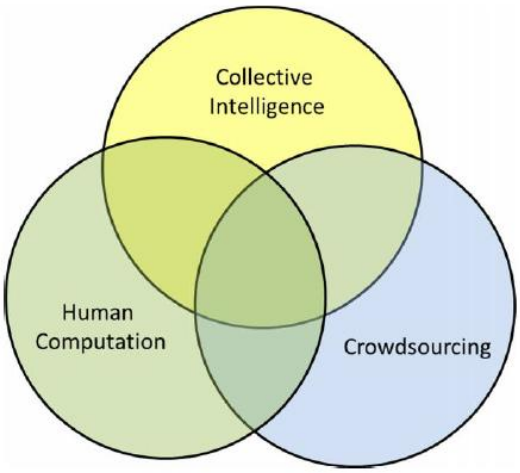


Figure 15. Relations entre *human computation*, *collective intelligence* et *crowdsourcing*, d'après (Harris, 2013)

La « *User innovation* » est une forme d'ouverture de la recherche aux internautes. Elle est beaucoup plus ouverte que le *crowdsourcing* qui est finalement très cadré par une procédure de contribution. Elle consiste à récolter les idées de recherches et innovations des internautes la plupart du temps sous la forme de concours et d'appels à contributions donnant généralement lieu à des récompenses. L'histoire des sciences est, en effet, remplie d'innovations venant d'amateurs extérieurs au métier, de bricoleurs et qui, ne cherchant pas à reproduire les modèles établis avec lesquels les professionnels ont été formés, sont parfois susceptibles de provoquer des ruptures innovantes. Le chercheur au MIT, Von Hippel, qui parle d'innovations par l'usage ou d'innovations ascendantes estime que 46 % des sociétés états-uniennes dans des secteurs innovants ont pour origine un utilisateur. L'innovation devient, grâce à leur apport, le résultat d'une collaboration directe entre les producteurs et les consommateurs qui deviennent des coproducteurs. Dans le domaine scientifique, le phénomène des "lecteurs inattendus" ou "unexpected reader", les découvertes accidentelles et les heureuses coïncidences (sérendipité) sont bien connus et illustrent bien ce phénomène. Mais le *crowdsourcing* se distingue aussi de la logique de la *User Innovation* car dans ce dernier cas, l'entreprise n'est pas toujours à l'initiative et à l'origine des projets et des idées dont elle bénéficie via les suggestions des

consommateurs. Or, avec le *crowdsourcing*, l'entreprise reste à l'initiative des projets.

Le *crowdsourcing* se distingue également de l'open innovation car contrairement à cette dernière, il est une forme d'externalisation ("outsourcing") vers la foule des internautes via le web 2.0 et non une externalisation de l'innovation auprès d'autres sociétés.

La notion d'externalisation correspond toutefois bien au *crowdsourcing* car la démarche ressemble à celle menée dans le cadre d'un appel d'offres ouvert avec la publicité qui est faite autour de l'appel. Il s'agit d'externaliser certaines missions non pas auprès d'un prestataire défini, mais auprès d'une communauté indéfinie d'internautes bénévoles afin de pouvoir réaliser des projets ou des innovations qui auraient été impossibles sans eux. Le *crowdsourcing* pourrait ainsi être considéré à la fois comme une forme renouvelée d'externalisation, un modèle économique innovant et une alternative à la sous-traitance. Mais, contrairement à l'externalisation, le *crowdsourcing* ne nécessite pas de contrat entre le commanditaire et le prestataire, d'autant qu'il s'agit ici d'une quantité large et indéfinie de collaborateurs.

Enfin, le *crowdsourcing* pourrait être considéré comme l'application des méthodes Open Source à d'autres industries en dehors du logiciel. Néanmoins, les développements ne se font pas nécessairement sur mode exclusivement collaboratif et peuvent aussi être alimentés par l'esprit de compétition. Par ailleurs, si l'open source repose sur plusieurs contributeurs travaillant pour satisfaire les besoins de plusieurs usagers, le *crowdsourcing* repose sur l'idée que plusieurs contributeurs vont travailler au service d'une seule entité.

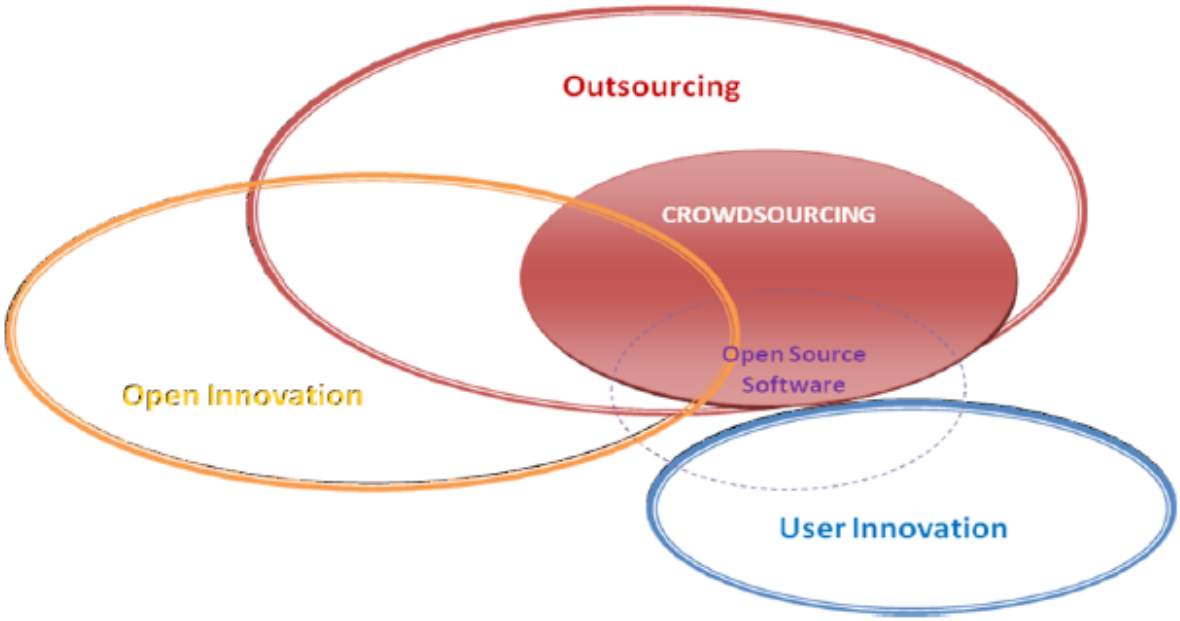


Figure 16. Positionnement du *crowdsourcing* parmi les domaines voisins (d'après Schenk, 2010)

Dans le cadre de ce travail de recherche, nous avons distingué cinq grandes familles de projets de *crowdsourcing* appliqué aux bibliothèques numériques et proposé une taxonomie originale comprenant :

Le *crowdsourcing* explicite : le recours aux bénévoles volontaires

Si le *crowdsourcing* explicite classique présente l'intérêt de collaborer avec le grand public et la société et être source d'opportunités par les ruptures innovantes que le public peut parfois susciter, le marché encore disponible pour ce recours revisité au bénévolat commence toutefois à se resserrer à cause de la multiplication des projets et de l'apparition de nouvelles formes de *crowdsourcing*. Par ailleurs, les bénéfices tirés de ces projets ne compensent que rarement les importants investissements nécessaires au développement des plateformes, à la communication, au recrutement, à la formation et au management des communautés de bénévoles.

Le *crowdsourcing* implicite : le recours au travail involontaire et inconscient

Le *crowdsourcing* implicite consiste à faire travailler les internautes sans qu'ils en aient conscience. Cette forme de *crowdsourcing* a permis d'obtenir d'excellents résultats mais peut poser des questions éthiques.

La *gamification* : le recours aux joueurs

Ces projets, qui consistent à obtenir du travail des internautes en les faisant jouer, peuvent être coûteux à développer et obtiennent également d'excellents résultats mais la collaboration sur le fond est moindre avec des internautes qui en bénéficient parfois moins sur le plan du développement personnel.

Le *crowdsourcing* rémunéré : le recours au micro-salariat

Cette forme de *crowdsourcing* popularisée par l'Amazon Mechanical Turk Marketplace et largement utilisée outre atlantique peut parfois être critiquée comme une forme d'exploitation du travail en dehors de tout cadre réglementaire. La marketplace met en relation des entreprises publiques ou privées qui proposent des microtâches (classification, indexation, identification, transcription, correction, rédaction) à plus de 700 000 travailleurs dans le monde et à un prix qu'elles fixent librement.

Le *crowdfunding* : la « mendicité » institutionnelle

Cette forme de *crowdsourcing* qui fait appel non plus au travail des bénévoles mais à leur argent a déjà été utilisé avec succès pour financer des projets. Le financement participatif (ou micro-mécénat ou encore mécénat à la demande) est bien une forme spécifique de *crowdsourcing* pour lequel l'apport des internautes est exclusivement financier. C'est cette forme particulière de *crowdsourcing* qui a fait l'objet de nos expérimentations.

Au delà de cette partie introductive destinée à définir le *crowdsourcing* afin de mieux délimiter le périmètre de cette thèse, nous reviendrons de manière bien plus approfondie sur la définition du *crowdsourcing* en l'appliquant spécifiquement au

domaine des bibliothèques numériques qui nous intéresse et en produisant une taxonomie originale plus détaillée du *crowdsourcing* dans les bibliothèques numériques. Ces développements trouveront leur place dans la partie destinée aux analyses du point de vue des sciences de l'information et de la communication.

1.3- Chronologie historique du *crowdsourcing*

Le *crowdsourcing* pourrait remonter à Hugues de Saint-Cher, un dominicain du 13è siècle, ayant coordonné de nombreux religieux afin d'indexer le contenu des textes saints (Le Deuff, 2015).

Mais la plupart des auteurs font remonter l'histoire du *crowdsourcing* au Longitude Act de 1714. Après l'accident de l'amiral anglais Cloudesley Shovell en 1707 aux îles Scilly, le gouvernement décida d'offrir 20 000 livres à celui qui serait capable de déterminer la longitude d'un navire en pleine mer et d'éviter de nouveaux accidents. (Dawson, 2011). Les célèbres scientifiques Cassini, Huygens, Halley et Newton échouèrent à trouver une solution et c'est John Harrison un charpentier et horloger qui remporta la récompense parmi plus d'une centaine de concurrents. (Lakhani, 2013)

Au début du 18^{e} siècle, Louis XIV proposa de récompenser les meilleures solutions proposées afin d'améliorer la production d'un produit chimique, l'Alkali (Chardonnens, 2015)

En 1726, une ordonnance de Louis XV demandait aux capitaines de navires de rapporter des plantes et des graines des pays étrangers qu'ils visitaient (Boeuf, 2012) et de contribuer ainsi à la recherche botanique.

Quelques décennies plus tard, en 1758, le mathématicien Alexis Clairaut parvint à calculer l'orbite de la comète de Halley en divisant les tâches de calculs entre trois astronomes. De son coté, l'astronome britannique Nevil Maskelyne calcula, en 1750, la position de la lune pour la navigation en mer grâce à la confrontation des calculs de deux astronomes ayant effectué les calculs deux fois chacun puis ayant été vérifiés par un tiers.

En 1775, Louis XVI offrait une récompense à qui permettrait d'optimiser la production de l'Alkali, un produit chimique. Le concours fût remporté par Nicolas Leblanc (Chardonnens, 2015).

En 1794, l'ingénieur français Gaspard de Prony organisa des microtâches d'additions et de soustractions auprès de quatre-vingt coiffeurs au chômage afin de développer des tables logarithmiques et trigonométriques détaillées.

En 1850, 600 volontaires d'Amérique du Nord et du Sud envoient des données météorologiques aux scientifiques de la Smithsonian Insititution au moyen de télégraphes (Steinbach, 2014).

En 1852, le magasin "Au bon marché" fondé par Aristide Boucicaut offre pour la première fois un magasin en libre service (ou "self service" en anglais), ancêtre des supermarchés contemporains. Une partie du travail du producteur est ainsi externalisée auprès du consommateur. Le modèle du libre service trouvera d'autres applications dans le commerce (caisses automatiques, par exemple) et des applications dans les banques (distributeurs de billets), la restauration (dans les fast food, par exemple, ce sont les consommateurs qui font le service et débarrassent la table), le mobilier d'intérieur (ce sont les consommateurs qui assemblent les pièces des meubles Ikea par exemple), les transports (VeLib, par exemple), les laveries automatiques de vêtements ou de véhicules et les bibliothèques (collections en libre accès).

En 1857, l'Oxford English Dictionnary bénéficia, suite à un appel à contributions bénévoles, de plus de 6 millions de documents contenant des propositions de mots et des citations d'usages.

En 1884, la statue de la Liberté est financée suite à une souscription publique de 125 000 personnes qui commença en 1875 en France.

En 1893, le scientifique statisticien anglais et père de l'eugénisme, Francis Galton constata, à l'occasion d'un concours lancé sur un marché de bétail et pour lequel il s'agissait de deviner le poids d'un bœuf, que la moyenne des estimations de la foule était plus proche de la vérité que l'estimation des experts, laissant supposer l'existence d'une sagesse des foules.

En 1894, le bibliothécaire James Duff Brown permet aux lecteurs de la bibliothèque publique de Clerkenwell d'avoir accès directement à une partie des collections, le libre accès en bibliothèque était né, il est l'adaptation du modèle du libre service aux bibliothèques.

En France, au 19e siècle, le gouvernement lança des appels à contributions. L'un d'entre eux, remporté par Nicolas Appert, permit de découvrir de nouvelles méthodes de conservation des aliments sous la forme de boîtes de conserve.

Au 19e siècle, dans le domaine de l'édition, se développa le système de souscriptions publiques pour financer la publication de livres. Ce modèle se développé dans bien d'autres domaines également (voir figure 17).

En 1900, la National Audubon Society (USA et Canada) organisa un décompte annuel des oiseaux, le "Christmas bird count".

En 1936, Toyota rassembla 27 000 personnes et sélectionna un dessin pour devenir le logo de sa marque. Beaucoup plus tard, le logo de Nike et de Twitter par exemple seront directement inspirés par des consommateurs.

En 1938, aux États-Unis, le Mathematical Tables Project mobilisa 450 chômeurs victimes de la dépression économique et pilotés par un groupe de mathématiciens et de physiciens afin de calculer des tables de fonctions mathématiques, bien avant l''invention de l'ordinateur.

Dans les années 1950, l'ingénieur industriel de Toyota, Taiichi Ōno, invente le modèle du "juste à temps" ancêtre du modèle "à la demande" et qui permettait de produire sans stocks ni invendus, en flux tendus, en fonction de la demande. Il s'agissait, d'une certaine manière, d'externaliser auprès du consommateur, la décision de mise en production elle-même. Ce modèle est, dans le domaine des bibliothèques, à l'origine de la numérisation à la demande par *crowdfunding* et de l'impression à la demande.

En 1954, le premier Téléthon, aux États-Unis permit de recueillir un financement pour lutter contre l'infirmité motrice cérébrale.

En 1955, l'Opéra de Sydney fut conçu et construit à la suite d'un concours public, qui encouragea les gens ordinaires de 32 pays à contribuer à ce projet de conception.

En 1979, le sondage Zagat, un guide de restaurants fonda ses évaluations sur une grande quantité de testeurs. Le projet fut racheté en septembre 2011 par Google.

En 1981, le guide de voyages Lonely Planet fut rédigé, pour sa 3ème édition, de manière participative par des voyageurs indépendants.

En 1997, le groupe de Rock Marillion financera sa tournée aux USA grâce aux dons pour un total de 60 000 $ de ses fans.

En 1998, l'annuaire Dmoz, propose un contenu généré par ses utilisateurs. Le web 2.0 était né.

En 2000, apparaît la plateforme philanthrope de *crowdfunding* justgiving.com et celle de financement participatif d'artistes artistshare.com qui seront suivies par de multiples initiatives jusqu'à aujourd'hui.

Le 23 novembre 2013, le jeu vidéo Star Citizen récolta une somme de 30 044 586 $.

Fin 2005, Amazon lança la plateforme de *crowdsourcing* Amazon Mechanical Turk Marketplace permettant de mettre en relation des sociétés et des institutions à la recherche et des travailleurs sur le web autour de microtâches.

Source gallica.bnf.fr / Bibliothèque nationale de France

Figure 17. Première forme de *crowdfunding* ? D'après http://gallica.bnf.fr/ark:/12148/btv1b8509563b (consulté le 23 juin 2016)

Précédemment, nous avons donné une définition synchronique du *crowdsourcing*, c'est à dire tel qu'il se présente à l'heure où ces lignes sont écrites. A la lumière de la chronologie que nous venons de présenter et qui est issue d'une analyse de la littérature, nous pouvons chercher à en donner une définition diachronique, c'est à dire tenant compte de son histoire. Nous pourrions ainsi dire que le *crowdsourcing* est l'expression dans l'économie du web 2.0 du mouvement d'externalisation progressive d'activités des producteurs auprès de foules de consommateurs.

1.4- Controverses philosophiques et politiques

Comme nous l'avons signalé dans l'avertissement de la thèse, le *crowdsourcing* est un sujet qui peut être source de forts clivages idéologiques. Dans les parties qui suivent, nous nous sommes cantonnés à rapporter, de la manière la plus équilibrée possible, les analyses et les avantages et inconvénients mis en avant par tel ou tel théoricien ou par telle ou telle idéologie. Nous déclinons donc toute responsabilité concernant les propos rapportés ici. Notre neutralité est aussi liée au fait que nous restons finalement très partagés face aux positions des uns et des autres. Nous avons veillé aussi à éviter tout jugement de valeur concernant certains faits et points de vues et à ne pas trop sortir du cadre d'une thèse en sciences de l'information et de la communication tout en prenant la liberté d'une certaine pluridisciplinarité.

L'origine philosophique et politique du *crowdsourcing* peut sembler très confuse de prime abord. Ce modèle économique semble, en effet, pouvoir faire écho à des idéologies aussi diamétralement opposées que sont le marxisme et le libéralisme. Nous verrons pourtant, à la fin de ce chapitre, qu'une certaine synthèse cohérente entre ces contraires peut être dessinée au travers d'une certaine « idéologie californienne ».

Il semble exister une parenté entre *crowdsourcing* et idéologies socialistes. N'a-t-on pas ainsi régulièrement accusé les sciences citoyennes de lyssenkisme[9], d'être en filiation avec la "science prolétarienne" et de représenter une volonté de contrôle populaire de la science, de représenter une « tentative d'intrusion idéologique et de prise de contrôle d'une partie de la production scientifique par des lobbys idéologiques »[10] ?

Les internautes qui participent à des projets de *crowdsourcing* semblent, en effet, accomplir le mot d'ordre socialiste « de chacun selon ses capacités à chacun

[9]Du nom de Trofim Lyssenko, un agronome russe des années 30 qui cherchait à appliquer le marxisme aux sciences naturelles, la notion de « lyssenkisme » est généralement utilisée pour évoquer l'intrusion des idéologies dans la recherche scientifique

[10] propos du blog "la faucille et le labo" rapportés par (Lipinski, 2014)

selon ses besoins ». En effet, chacun fait de son mieux pour contribuer à produire des contenus en fonction du temps, des forces et des compétences dont il dispose. Et les contenus produits bénéficieront à tous, ceux qui en ont beaucoup besoin, comme les autres, ceux qui ont beaucoup contribué comme les autres. Il n'existe pas d'impact proportionnel entre ce qui a été produit et ce qui sera consommé.

Dans nombre de projets de *crowdsourcing*, on voit apparaître, parmi les motivations des contributeurs, la volonté de sacrifier de leur temps au profit de l'intérêt général, le besoin de se sentir utile à une communauté, agir par altruisme et redevabilité, protéger le patrimoine culturel...

Certains auteurs, comme Jean-Pierre Gaudart dans son livre "la fin du salariat" annoncent même la disparition du salariat. Avec l'arrivée sur le marché du travail de la génération Y, en particulier, le développement du travail en free lance ou en tant qu'auto-entrepreneurs, le rapport au travail semble évoluer. L'engagement vis-à-vis de l'entreprise semble s'affaiblir avec l'émergence de salariés plus autonomes, individualistes et plus centrés sur leurs egos. La tension entre l'individu salarié et l'entreprise collective semble s'accroître avec l'arrivée sur le marché du travail de la génération Y. Les "digital natives" ne s'investissent plus dans ce cadre collectif, ils sont sans attaches et ne se sédentarisent plus. Considérés souvent comme des mercenaires sans foi ni loi, ils sont parfois aussi sans feu ni lieu, à la recherche d'une identité perdue et souffrent d'un manque de reconnaissance et d'une difficulté à se réaliser dans le cadre de l'entreprise traditionnelle. Dans le même temps, une "classe créative" semble émerger. On parle ainsi de "jobcrafting", c'est-à-dire de processus au cours duquel les employés révisent activement et progressivement leurs fiches de postes et leurs relations avec les autres (Deng, 2013). La notion de travail tend à disparaître au bénéfice de celle d'activité.

Avec le *crowdsourcing*, si la consommation devient productrice de valeur et si le loisir devient créateur de richesse, le travail devient un loisir. L'argent semble ne plus être le moteur principal de l'activité d'une masse importante d'amateurs, au sens noble, gravitant dans le domaine du logiciel libre au détriment de la motivation, de l'investissement de soi et de la passion. Sur le web, les modèles

économiques basés sur la gratuité semblent également l'emporter, laissant percevoir l'émergence de "communaux collaboratifs" (Rifkin, 2014), d'une "économie de la contribution" (Stiegler, 2015), d'une économie du partage, d'une "économie participatiste" ou d'une "économie collaborative". Selon (Stiegler, 2015), avec l'automatisation rendant le travail de moins en moins nécessaire, les salariés comme les consommateurs deviendront des contributeurs de l'entreprise, c'est-à-dire des amateurs motivés davantage par leurs centres d'intérêts que par leurs intérêts économiques. Il s'agira alors de les rémunérer selon un intéressement contributif. Déjà, certaines entreprises n'ont plus de salariés mais des contributeurs externes ou utilisent des travailleurs via l'Amazon Mechanical Turk Marketplace. Internet semble bien être le médium de l'abolition de la médiation. Ainsi, nombreux sont les sites web qui viennent disputer leurs places d'intermédiaires entre le producteur et le consommateur aux acteurs économiques traditionnels confortablement installés ou jouissant de monopoles (taxis, agences de location, agences de recrutement…). On parle même d'une "ubérisation" de l'économie. De plus en plus d'entreprises risquent ainsi de se faire évincer par des sociétés web ayant recours à des travailleurs indépendants plus compétitifs. Ce mouvement est loin d'être marginal. Ainsi, d'après une étude du PwC publiée en 2014 sous le titre "The Sharing Economy", l'économie collaborative devrait passer de 15 milliards en 2014 à 335 milliards d'euros en 2025.

Certains théoriciens du pair à pair ("P2P") comme Michel Bauwens, estiment que les humains peuvent désormais rentrer en contact, partager des données et collaborer sans permission, ni hiérarchie, chacun comblant les lacunes de l'autre et que cela va modifier en profondeur nos sociétés. Selon eux, le P2P est donc le socialisme du 21^{e} siècle. Les hiérarchies verticales étaient définies par le pouvoir. Avec les communautés P2P, c'est la réputation qui prédomine, le fonctionnement est plus horizontal. Cette réputation est mesurée, en fonction du trafic web généré par la production de telle ou telle personne, sur le même modèle que le taux de citation dans la recherche scientifique. On peut même parler d'économie de la réputation dans la mesure où la réputation peut être convertie en

argent via les publicités qui rapportent en fonction du trafic web généré, mais aussi en emplois, en opportunités de partenariats…

Quoi qu'il en soit, même s'il s'avérait moins révolutionnaire que ce que prétendent certains théoriciens, le *crowdsourcing* constitue « une innovation de rupture, qui va donc modifier profondément et durablement l'écosystème d'affaires » (Lebraty, 2015).

En cherchant à ré-humaniser Internet et en redonnant une place centrale à l'humain comme origine et finalité d'un web qui doit être créé par l'humain et pour l'humain, le *crowdsourcing* est incontestablement aussi un héritier des philosophies humanistes et eudémonistes. Le *crowdsourcing* a recours aux foules humaines dont les capacités et l'intelligence demeurent largement supérieures à celles des algorithmes. Face à l'intelligence artificielle et au big data, le *crowdsourcing* garde foi en la supériorité humaine. D'ailleurs le projet de *crowdsourcing* rémunéré de l'Amazon Mechanical Turk Marketplace prend malicieusement pour emblème un automate joueur d'échec très ancien qui était réputé doté d'une réelle intelligence artificielle alors qu'un humain était, en réalité, caché dans le mécanisme. Amazon affirme ainsi que l'intelligence humaine reste indépassable.

Le *crowdsourcing* participe également du mouvement des humanités numériques et on peut même, à juste titre, parler d'"humanisme numérique" à l'instar de Milad Doueihi, au sens où les nouvelles technologies ont une dimension universelle et qu'elles sont une culture car elles mettent en place un nouveau contexte (rapporté par Moatti, 2015).

Le *crowdsourcing* peut tout aussi bien être considéré comme une forme libérale, nouvelle et étendue d'externalisation et d'ouverture de l'organisation sur son environnement extérieur. En effet, dans un premier temps, la mondialisation de l'économie et la concurrence exacerbées entre les entreprises avait amené les industries, ne connaissant d'autres lois que celle de l'offre et de la demande, à externaliser vers des pays à main d'œuvre à bas coût. Mais, avec le développement d'Internet, il devient désormais possible de faire travailler

quiconque est simplement relié au réseau. Le *crowdsourcing* reste donc bien une forme d'externalisation du travail sur Internet, dans des domaines encore limités.

Sur Internet, les liens, les clics, les commentaires, les notations, les recommandations, les visites, les liens... fonctionnent comme les votes dans la démocratie. Les sites bien référencés et mis en avant par les moteurs de recherches sont des sites élus par les internautes. Il existe bien une hiérarchie entre eux puisque les pages les plus visibles sont les plus citées, les plus liées, les plus commentées. Le PageRank pourrait, d'une certaine manière, être considéré comme une forme de crowd voting (Renault, 2014) implicite. En ajoutant, sur le web, un lien vers un site web, l'internaute va ainsi inconsciemment voter pour que ce site soit mieux référencé par le moteur de recherche.

Le *crowdsourcing* a également largement recours à la notion de sagesse des foules qui est, elle-même très proche de la notion libérale de main invisible. Francis Galton, père de l'eugénisme et cousin de Charles Darwin, avait ainsi constaté à l'occasion d'un concours populaire consistant à deviner quel était le poids d'un bœuf, que la moyenne des estimations des participants était très proche de la vérité. On constate aujourd'hui, de la même manière, que si on demande à un amphithéâtre d'évaluer le nombre de billes contenues dans une bouteille ou la température d'une pièce, la vérité est très proche de la moyenne des réponses. C'est pour cette même raison que les participants du jeu "Qui veut gagner des millions ?" avaient beaucoup plus de chances de trouver la bonne réponse en sollicitant l'avis du public qu'en ayant recours à un ami. S'appuyant sur ce phénomène, l'Intelligence Advanced Research Projects Activity (IARPA), une agence américaine de renseignements a lancé le Good Judgement Project afin de tirer bénéfice de la sagesse des foules car susceptibles de mieux prévoir les événements géopolitiques que les experts et les analystes traditionnellement utilisés par les agences de renseignement. Ce projet fait écho, d'une certaine manière, à l'adage "vox populi vox dei" et à la citation suivante de Machiavel qui estimait que « *Ce n'est pas sans raison qu'on dit que la voix du peuple est la voix de Dieu. On voit l'opinion publique pronostiquer les événements d'une manière si merveilleuse, qu'on dirait que le peuple est doué de la faculté occulte de prévoir et*

les biens et les maux. Quant à la manière de juger, on le voit bien rarement se tromper. » (Machiavel, 1837).

Toujours dans le domaine du renseignement, l'analyse par *text mining* des lieux géographiques les plus co-occurrents avec le nom de Ben Laden a laissé apparaître, que ces lieux étaient les plus proches de l'endroit où il a effectivement été trouvé. Cela ne signifie pas que les journalistes savaient, cela signifie qu'une quantité importante de données peut se transformer en renseignement de qualité et que lorsqu'elles font foules, elles font science.

De ce point de vue, il semblerait donc bien qu'il existe une "main invisible" qui permette aux individus librement associés de trouver des solutions justes et harmonieuses sans intervention d'une autorité quelconque, que les intérêts particuliers non entravés soient naturellement bénéfiques à l'intérêt commun. Cette notion se rapproche aussi de celle d'ordre spontané proposée par Friedrich Hayek, c'est à dire un ordre auto-généré, auto-organisé, sans plan, ni autorité, comme celui qui règne sur les marchés mais aussi de Holacratie, une organisation fractale d'équipes auto-organisées de manière organique ou encore de sociocratie. On pourrait considérer que l'encyclopédie participative Wikipédia est aussi un ordre spontané car elle est exhaustive et structurée grâce à l'action autonome et non concertée d'individus et sans qu'un plan complet ait préexisté à son développement. Jimmy Wales, le fondateur de Wikipédia se revendique d'ailleurs de Friedrich Hayek, en particulier pour sa conception du projet Wikipédia. En effet, la croyance dans une correction spontanée des articles de Wikipédia est assez similaire à la croyance libérale de la main invisible du marché.

Les organisations qui ont recours au *crowdsourcing* ont conscience de leurs limites. Elles ont confiance dans la capacité des foules à trouver spontanément les meilleures solutions lorsqu'on rend la liberté d'initiative et l'autonomie aux individus qui les composent.

Avec le développement de la nouvelle économie, la différence entre vie privée et publique, bénévolat et travail semble devenir plus confuse. Les salariés travaillent de plus en plus dans les transports, le soir, pendant leurs week-ends, leurs congés… En contrepartie, ils consacrent aussi parfois du temps de travail à

des relations sociales, voir à des loisirs, avec la bénédiction de sociétés qui comprennent que leur épanouissement personnel sera source de créativité et d'innovation. On parle parfois de "weisure" selon l'expression de Dalton Conley, un mélange de travail (work) et de loisirs (leisure), ou encore de "playbor" ou de "playbour", mélange de jeux (play) et de travail (labour) ou enfin d'intrapreneurs, c'est à dire de personnes ayant l'esprit d'initiative et d'entreprise tout en étant salariés, d'entrepreneurs internes à l'entreprise. Les hiérarchies s'en trouvent bousculées, ce n'est plus la direction qui décide et les salariés qui exécutent, mais souvent les salariés qui sont directement à l'origine des projets. L'innovation ouverte remet en cause la division sociale du travail (Von Hippel, 2005). Avec le web 2.0 et tout particulièrement avec le *crowdsourcing*, la frontière entre producteurs et consommateurs est en train de disparaître car les consommateurs d'informations sur le web en sont également devenus les producteurs. Des millions de personnes produisent, pour le plaisir, des données, et travaillent ainsi gratuitement pour YouTube, Facebook, d'autres participent à améliorer les logiciels sans le savoir par l'utilisation gratuite qu'ils en font. Si Facebook a annoncé un chiffre d'affaire de 2,5 milliards de dollars en 2013, soit 6,81 $ par utilisateur actif, ces revenus demeurent néanmoins surtout liés à la publicité et non à la revente des données. Lorsque les internautes saisissent une recherche sur Google, rédigent un tweet, mettent du contenu sur Facebook, rédigent un commentaire de livre sur Amazon, postent une évaluation de vendeur sur ebay, évaluent la qualité d'un restaurant sur l'Internaute… ils produisent des données qui ont une valeur qui sera revendue par ces sociétés et travaillent gratuitement pour elles, en échange du service gratuit qu'elles leur rendent. (Fuchs, 2012) estime ainsi que Facebook a bénéficié de 60 milliards d'heures de travail non payé. Sur le web, on utilise de nombreuses applications gratuites en apparence. En réalité, en échange de la gratuité du service, on travaille sans le savoir à produire des données, lorsqu'on rédige sur Facebook, lorsqu'on recopie un CAPTCHA et même lorsqu'on fait une requête. Ce travail de production de données échappe à toute règle et à toute législation.

Ainsi, au lieu d'une participation des internautes, le *crowdsourcing* pourrait donc plutôt engendrer une exploitation du travail gratuit des usagers parfois qualifiée parfois de "servuction". Ainsi, comme le rapporte (Petersen, 2008), en 1999, 7 des 13 000 bénévoles d'AOL, qui travaillaient gratuitement à faire vivre et à dynamiser la communauté AOL avaient finalement revendiqué une rétribution de leur travail puis, deux d'entre eux avaient même été jusqu'à déposer une plainte contre AOL devant un tribunal fédéral à New York avant que l'enquête soit fermée en 2001.

Ce mode de travail qui modifie les frontières entre production et consommation a été conceptualisé sous le terme de « *digital labor* » ou « digital labour » qui pourrait être traduit en français par travail numérique. Il comprend le travail implicite et invisible de production de données par des internautes grâce à leurs activités sur le web et en bouleverse les limites (Cardon, 2016).

Quoi qu'il en soit, devant certaines dérives de sites réalisant leurs profits grâce au travail gratuit des internautes, les pouvoirs publics affichent parfois leur volonté de développer une fiscalité autour de la captation les données. Fiscaliser les données permettrait ainsi de rendre à la communauté une partie de la création qu'elle a fournie sous la forme d'un "travail invisible". Mais ce travail est d'autant plus invisible qu'il est de faible intensité et difficile à faire reconnaître.

Le "*digital labour*" pourrait aussi être rémunéré sous la forme de micro paiements individualisés, ou en échange d'actions en particulier pour le *crowdfunding* ("*equity crowdfunding*"), ou encore via la fiscalité collective sur les données. Avec le *crowdfunding* 2.0, les participants pourraient donc passer de consommateurs à actionnaires et les starts ups vendre des actions pour financer leurs projets. Un texte allant en ce sens a ainsi été voté aux USA par la Securities and Exchange Commission (SEC). Comme cela est expliqué sur le blog InternetActu.net [11] en particulier, l'usager pourrait ainsi être reconnu comme producteur de données, en reprendre le contrôle et être rétribué en tant que producteur de valeur.

[11]http://www.internetactu.net/2012/06/01/vers-un-nouveau-monde-de-donnees (consulté le 23 juin 2016)

Avec le *crowdsourcing*, on pourrait passer d'un mode de production dans lequel le prolétaire vend sa force de travail au capitaliste en échange de son salaire à une économie participative dans laquelle le contributeur offre sa participation dans l'intérêt d'une communauté d'internautes. L'Amazon Mechanical Turk Marketplace, par exemple, à l'instar d'autres plateformes de *crowdsourcing* rémunéré, permet une extension du travail indépendant de type freelancing, une nouvelle forme de travail, les employeurs proposant des tâches sur la plateforme et les travailleurs venant les exécuter librement comme des micro-entrepreneurs et en dehors de toute autre règle que la loi de l'offre et de la demande dans un marché totalement ouvert et libéral où les uns et les autres vendent et achètent librement du travail en ligne. Au lieu de risquer le « *burn out* » de ses employés, l'employeur peut ainsi recruter en quelques minutes des foules de travailleurs, avec des profils diversifiés, disponibles en permanence, généralement bon marché, accessibles sans autres démarches administratives et payés seulement une fois le travail accompli. L'employeur peut ainsi réaliser des tâches impossibles à imaginer auparavant. Il peut, en quelques minutes recruter des effectifs aussi importants et diversifiés que ceux de grandes entreprises et les mobiliser autour de projets.

Du point de vue des travailleurs, certains sont heureux de pouvoir travailler quand ils veulent, quand ils en ont besoin, autant qu'ils en ont besoin, pour qui ils souhaitent et de choisir les activités qu'ils vont accomplir. D'autres vivent, par exemples, des services qu'ils fournissent sur Uber en tant que chauffeur de voiture, de bricolage ou de jardinage sur TaskRabbit. Ils partagent les biens dont ils sont propriétaires mais dont ils ont un usage limité et sont dans des logiques de qualité et d'usages plutôt que de propriété afin de diminuer leurs dépenses et éviter les gaspillages grâce à une consommation collaborative (Peugeot, 2015).

Mais, d'un point de vue éthique, l'exploitation du travail bénévole ou du travail sous payé et échappant à toute législation dans le cadre de Amazon Mechanical Turk pose un problème à la fois juridique et même économique. On peut considérer qu'il s'agit aussi, à l'instar de toute externalisation, d'une forme de dumping social et de concurrence déloyale vis à vis de sociétés ou de corporations. On peut estimer que les travailleurs en réseau jouent le rôle d'une armée

industrielle de réserve qui pèse à la baisse sur les salaires et que la plateforme d'Amazon propose le même type de services que les prestataires traditionnels à un tarif sensiblement différent puisque non assujetti aux mêmes règles et aux mêmes impôts.

Avec le *crowdsourcing*, le risque reste important de faire de l'humain un simple moyen de parvenir à une fin marchande, de l'assimiler à un simple ordinateur (Sagot, 2011), de lui retirer tout caractère sacré, de le considérer comme une simple matière première et de se retrouver en contradiction avec la morale de Kant qui énonçait "Traite toujours autrui comme une fin et jamais seulement comme un moyen".

Le *crowdsourcing* peut être accusé d'être déloyal. Ainsi, une équipe qui participait au Shredder challenge organisé en 2011 par la Darpa (Pentagone) afin de reconstituer des documents passés dans une machine de type destructeur de documents a été victime de vandalisme car elle a été considérée comme utilisant des méthodes déloyales. Cette équipe avait fait appel au *crowdsourcing* sous la forme de puzzles tandis que ses concurrents utilisaient des algorithmes informatiques pour assembler les images. Ces derniers considérèrent cette méthode comme de la triche par rapport aux algorithmes qu'ils cherchaient à développer, et vandalisèrent rapidement le projet de *crowdsourcing*.

Comme le soulignent (Fort, 2011), l'Amazon Mechanical Turk Marketplace n'est ni un jeu ni un réseau social, mais un marché non réglementé qui ne paie aucun impôt et où les travailleurs, considérés comme des auto-entrepreneurs, qui vendent leur force de travail pour des tâches répétitives et peu qualifiées, qui sont sous-payés [12] , interchangeables, ne bénéficient d'aucune protection et sont doublement subordonnés au client et à la plateforme, bref une sorte de bagne numérique. Comme le prétend (Sagot, 2011), il est probable que ni les turkers, ni leurs employeurs ne déclarent leurs revenus, ne cotisent à une caisse de sécurité sociale, de retraite et ne sont pas inscrits au registre du commerce. Cette plateforme de travail au noir priverait ainsi les États de revenus légitimes et

[12] Le salaire horaire moyen serait de 2 $ d'après (Kittur, 2013)

remettrait directement en cause leurs législations du travail. Le fait de faire travailler des gens anonymes sans jamais les rencontrer encouragerait des comportements inhumains et une exploitation sans limite ni éthique de leur force de travail. Pour leur part, les travailleurs pourraient aussi, et pour les mêmes raisons que les employeurs, se sentir libérés de toutes obligations morales et développer des comportements cyniques (Kittur, 2013) ou des escroqueries.

Concernant les concours créatifs qui font appel à du “travail spéculatif”, c’est à dire à du travail produit gratuitement dans l’espoir d’être récompensé (Renault, 2014) par des foules de graphistes qui ont finalement peu de chances d’être rémunérés, ils avantagent grandement les entreprises qui bénéficient d’un nombre beaucoup plus grand de propositions de maquettes tout en ayant que quelques individus à récompenser pour un coût global bien inférieur à celui des agences traditionnelles. Il s’agit finalement plutôt de professionnels externalisés que de réels amateurs. Et, dans la mesure où aucun contrat ne lie le participant au concours à l’entreprise, le droit du travail ne saurait s’appliquer, d’autant que s’il s’agit d’un moyen de vivre pour certains candidats, ce n’est qu’un simple loisir pour d’autres. Le *crowdsourcing* pourrait également permettre la renaissance du salaire aux pièces et favoriser le désengagement de l’employeur qui ne serait plus contraint de « se lier avec un petit nombre de personnes lorsque l’on peut disposer d’une foule d’employés » (Lebraty, 2015)

Avec le *crowdsourcing*, la consommation de services gratuits sur les réseaux devient productive de données, d’informations et de valeur, rendant productifs tous les domaines de la vie sociale, le temps libre et la consommation devenant elle-même une production. De la même manière, Guy Debord prévoyait « une colonisation de l’ensemble des sphères de l’existence sociale par l’autorité de la marchandise dans l’organisation du Spectacle » (Sarrouy, 2014). Dans le prolongement de l’intérêt centré sur le consommateur à travers le modèle économique du “à la demande”, le *crowdsourcing* semble participer à réaliser cette colonisation, et finalement cette intégration du consommateur dans le procès de production en tant qu’auxiliaire de production non rémunéré. Comme le regrette Harald Staun, le temps mort, le temps libre disparaît avec l’arrivée du commerce et

la recherche de profit dans le temps libre. La vie même devient ainsi moteur de la productivité, le capitalisme un mode de production “bio-politique” (Aspe, 2013 rapporté par Sarrouy, 2014). Même nos relations humaines les plus profondes sont susceptibles d’être transcrites en algorithmes par les réseaux sociaux et valorisées commercialement. Les rapports marchands aussi se généralisent puisque, avec la consommation collaborative, chaque propriétaire d’un objet de consommation devient un commerçant qui peut en louer l’usage. La différence entre production et consommation, entre travail et loisir s’estompe donc, les internautes créent de la valeur par les contributions gratuites qu’ils fournissent et pourront être réutilisées et monétisées via le big data.

Comme l’affirment certains auteurs (Scholz, 2008), le web 2.0 aurait tous les attributs de l’idéologie, une idéologie totalitaire promettant des “lendemains qui chantent”, une idéologie qui ne se cantonne pas dans la sphère publique et politique, ne respecte aucune limite constitutionnelle à son pouvoir, mais s’immisce jusque dans le privé et l’intime, le rêve d’un société où tous seraient connectés, au delà des nations et des classes et dans le cadre d’un gouvernement mondial, bref, une tour de Babel. Le *crowdsourcing* pourrait finalement aussi présenter une parenté avec les idées libertaires et anti autoritaires puisqu’il substitue l’animation de communauté de volontaires qui s’auto-organisent de manière décentralisée au commandement hiérarchique et centralisé de salariés. Le sociologue Michel Lallement qui a étudié les hackers californiens considère ainsi qu'ils prolongent la contre-culture libertaire (Lallement, 2015). L’existence d’Internet semble démontrer la possibilité d’un fonctionnement harmonieux et sans hiérarchie. Dans l’encyclopédie participative Wikipédia, par exemple, un article rédigé par des scientifiques se retrouvera sur le même plan qu’un article rédigé par un collégien sur son héros préféré de bande dessinée. Linux est le résultat du travail agrégé de milliers de programmeurs, travaillant bénévolement et gratuitement à une œuvre commune de manière décentralisée. Christian Quest (OpenStreetMap) résume parfaitement cette idée : « Pour ajouter un commerce près de chez soi à une carte existante, on ne vous demandera jamais d’avoir un master en géographie ! ». Michel Bauwens qualifie d’ « anti-crédentalisme » ce type de position contre le

monopole des diplômes et déplore qu'on ne puisse être crédible en tant que scientifique sans titre de docteur, en tant que journaliste sans carte de presse (Bauwens, 2015). De la même manière, certains concours créatifs proposent à des graphistes, à des artistes, à des publicitaires amateurs, débutants, sans poste et sans références d'avoir autant de chances de réussir qu'un professionnel expérimenté et en poste ou qu'une personnalité de renom et de pouvoir ainsi plus facilement percer (Renault, 2014).

Le fait de donner d'avantage de responsabilités et de pouvoir d'agir au peuple et au consommateur, de capacité d'action aux internautes et de les émanciper rejoint la notion de "*empowerment*" notion qui pourrait être traduite en français par "empouvoir", "empouvoirement", "capacitation" ou encore appropriation, habilitation, pouvoir d'agir. Ainsi, le projet de sciences participatives fold.it affirme que son but ultime est que des gens ordinaires puissent éventuellement, grâce à son jeu de puzzle, gagner le prix Nobel (Good, 2011).

De la même manière que la frontière entre production et consommation semble s'effacer avec le *crowdsourcing*, la frontière entre les auteurs qui écrivent et les lecteurs qui lisent est en train d'être progressivement abolie puisque chacun est désormais à la fois lecteur et écrivain sur le web, réalisant ainsi encore d'avantage l'analyse de Walter Benjamin ("l'artiste comme producteur" 1934). Walter Benjamin estimait, en effet, que l'émergence des nouveaux médias allait remettre en question le paradigme de l'expert et que le progrès technique était à la base du progrès politique. (Deodato, 2014). On pourrait également parler de "écrilecture", c'est à dire d'une lecture active, d'une non séparation entre les actions de lire et d'écrire, en annotant, par exemple, au fil de sa lecture.

Comme nous l'avons vu dans le texte qui précède, le *crowdsourcing* est susceptible de séduire aussi bien les marxistes et les libéraux, pour des raisons diamétralement opposées. Comme le remarque, par exemple (Schultz, 2005), l'ambiguïté de l'info-communisme est l'une des principales ressources de l'économie de la connaissance néo-libérale et peut être décrit comme à la fois révolutionnaire et réactionnaire. Il combine à la fois les rêves de l'info-capitalisme et ceux du soviet constructivisme. Comme le remarque également Bastien Guerry,

« les “gauchistes” du web sont aussi des libéraux, voire des patriotes » (Benyayer, 2014). Elisabeth Grosdhomme Lulin estime aussi que « sur le plan des idées et des doctrines, [cette idée de co-production du service public par ses bénéficiaires] trouve des racines aussi bien à gauche qu’à droite : à gauche avec les utopies autogestionnaires, à droite avec les utopies libertariennes – d’un côté, dans le sillage de Pierre Joseph Proudhon, redonner le pouvoir au peuple, ouvrier ou citoyen, de l’autre, après Friedrich Hayek, limiter l’emprise de l’État sur l’économie et la société. » (Grosdhomme Lulin, 2013). Rachel Botsman et Roo Rogers estiment, dans un diaporama “what is mine is yours”, que la consommation collaborative répond à la fois aux idéologies socialistes et capitalistes sans être une idéologie. Enfin, (Nelson, 2012) constate quant à lui, qu’il existe finalement, dans toute cette confusion, une paradoxale parenté entre l’idéologie soviétique de l’émulation socialiste et les idées libérales américaines de *gamification*. En effet, les méthodes stakhanovistes visant à motiver et développer la productivité du travail en récompensant les meilleurs ouvriers par des points, des décorations, des médailles soviétiques, des titres de “Héros du travail socialiste”, des prix Staline et en organisant des compétitions entre ateliers, entreprises, usines, kolkhoz, sovkhoz, districts, villes, régions et républiques pour développer leur esprit d’initiative et d’entreprise n’est finalement pas si différente du gain de l’affichage de l’employé du mois dans le symbole du capitalisme américain que sont les restaurants Mc Donalds et qui leur offre parfois également des cadeaux. L’International Amateur Scanning League, un projet de numérisation du patrimoine par des bénévoles offre, par exemple, des médailles sur ce modèle, selon le nombre de DVD gravés.

Cette paradoxale proximité entre idées socialistes et libérales est bien représentée dans l’”idéologie californienne” qui combine l’esprit d’indépendance et d’autonomie hippie avec l’esprit d’entreprise Yuppie (Young Urban Professional). La silicon valley donnerait ainsi lieu à l’émergence d’une idéologie libertarienne, libérale (Yuppie) et libertaire (hippie). Richard Barbrook[13], à l’origine du terme estime même que internet pourrait être une forme moderne de l’économie du don

[13]Barbrook, R., Cameron, A. (2000). The Californian Ideology : Revised SaC Version, Borsook

à l'instar du sociologue Warren Hagstrom[14] qui estimait que la science était aussi une économie du don (Surowiecki, 2008). Chaque contributeur ajoute à la connaissance collective et reçoit des autres contributeurs beaucoup plus que n'importe quel individu pourrait lui donner. Le chercheur scientifique, le développeur ou, plus largement, le détenteur d'une information ou d'une connaissance ne la perd pas en la partageant. Selon (Barbrook, 2000), l'idéologie californienne serait néanmoins l'idéologie d'une sorte d'aristocratie de l'high tech Nietzschéenne, d'une sorte d'élite jacobine, d'une avant-garde cyber-communiste ou d'une sorte de technocratie du web qu'il nomme "digerati". Ces lettrés numériques ("digital literati" ou "digerati") sont convaincus que les nouvelles technologies vont révolutionner la société. Ils cherchent à éduquer les masses et à les conduire vers la modernité pour créer une civilisation utopique, une société de l'information. Les digerati seraient donc des modernistes réactionnaires cherchant à imposer une dictature du prolétariat revisitée et qui ne durerait, elle aussi, que le temps nécessaire à l'avènement de la nouvelle société. Ils ne sont pas sans rappeler les "anonymous" dont le slogan, "nous sommes légion" n'est pas non plus sans rappeler la puissance des foules d'internautes qui fonde le *crowdsourcing*.

Comme le souligne (Cardon, 2010), Internet est l'héritier de la contre culture américaine libertaire et égalitaire et de la méritocratie libérale du monde de la recherche et de l'informatique. Ils allient les idées de Marshall McLuhan avec certaines idées libertaires radicales. McLuhan pensait que le medium, c'est à dire l'intermédiaire entre l'émetteur d'information et son récepteur qui peut prendre la forme de l'oralité, de l'imprimé, du cinéma, le la radio, de la télévision et aujourd'hui d'Internet prime sur le contenu du message lui-même ("le message, c'est le medium"[15]). Il est également à l'origine de la notion de "village global". Les cyberlibertariens, ces « technofans », ces tenants des « mythes de la techno-utopie » (Chaudiron, 2013), ces partisans du déterminisme technologique et du "solutionnisme technologique" considèrent que les technologies apportent une contre-culture démocratique de manière inhérente, vont changer la société et

[14]Barbrook, R. (2000). L'économie du don high tech, Libres enfants du savoir numérique, Paris, Editions de l'Éclat , «Hors collection», 504 pages

[15]McLuhan, M. (1968). Pour comprendre les médias, Seuil, coll. Points, 404 p

résoudre les problèmes sociaux ou sociétaux de la même manière que les marxistes attendaient le paradis communiste du développement des forces productives et de la révolution qu'elles devaient fatalement provoquer.

1.5- Conséquences économiques, sociologiques et juridiques

1.5.1- Économie du *crowdsourcing*

Sur le plan strictement économique, le *crowdsourcing* pourrait représenter une réserve importante de marchés et de développements. En effet, le temps mondial cumulé de connexion sur Internet doit avoisiner les 160 000 000 d'heures pas jour. L'idée sous-jacente qui repose dans le *crowdsourcing* est que le temps libre passé sur le web à consommer des contenus pourrait être utilisée de manière productive pour l'économie. Ainsi, les données personnelles sur le web social sont converties en informations statistiques, donc en valeur. Les jeux sur le web en particulier pourraient avoir des finalités éducatives (*serious games*) mais aussi productives de données (*gamification*). Concernant le *crowdfunding* en particulier, d'après l'article "Global Crowdfunding Volumes Rise 81% In 2012" publié le 04/08/2013 dans The Huffington Post, les sites de *crowdfunding* auraient levés 0,89 milliards de dollars en 2010, 1,47 milliards de dollars en 2011, et 2,66 milliards de dollars en 2012.

En France, le nombre de bénévoles était estimé à 16 millions de personnes en 2011, c'est à dire 32 % de la population nationale[16]. En 2014, d'après une enquête CREDOC « Conditions de vie et aspirations », 47 % des français seraient adhérents d'une association (Daudey, 2014). Il existe donc un potentiel important pour le *crowdsourcing* dans ce pays en particulier.

[16] http://www.associations.gouv.fr/1121-le-benevolat-en-france-en-2011.html (consulté le 23 juin 2016)

1.5.1.1- La disparition du travail nécessaire ?

Les technologies évoluant, la productivité et la croissance augmentant, la part de travail nécessaire à la survie de l'humanité est devenue de plus en plus faible. En France, par exemple, l'agriculture ne représente plus que 3,9 % des emplois en 2005 et 1,8 % du PIB et l'industrie 24,3 % des emplois en 2005 et 18,7 % du PIB. Et cette industrie est loin d'être entièrement consacrée à la survie de l'humanité. En 1982, les États-Unis produisaient 75 millions de tonnes d'aciers avec 300 000 travailleurs. En 2002, 100 millions de tonnes étaient produites par seulement 74 000 travailleurs. Dans les services, on estime qu'une banque traditionnelle nécessite aujourd'hui 10 fois moins de salariés pour gérer les comptes d'un nombre identique de clients. On produit beaucoup plus avec beaucoup moins de travail. Jeremy Rifkin, prophétisait ainsi que « seuls 5 % de la population adulte suffirait à faire fonctionner les industries traditionnelles » et que « les usines, les bureaux et les exploitations agricoles, sans travailleurs ou presque, seront la norme dans le monde entier ». (Rifkin, 1996). L'amélioration de la productivité du travail, liée au développement des nouvelles technologies, détruirait de l'emploi. A titre d'exemple, les grandes industries traditionnelles embauchent un nombre bien plus importants de salariés que les plus grosses sociétés de l'Internet. Par exemple, Facebook n'aurait que 3976 employés en septembre 2012 pour 1 milliard d'usagers, soit 250 000 clients par employés, Twitter 900 employés pour 500 millions de clients, Google 54 604 employés pour 1 milliard de visiteurs uniques par mois en juillet 2012 (d'après Jean Paul Lafrance, 2013, rapporté par Sarrouy, 2014).

Le développement de ce mouvement marque-t-il la fin du capitalisme ? Si les technologies permettent de réduire les coûts marginaux des services jusqu'à la quasi gratuité et si ce mouvement atteint désormais aussi la production des biens équipés de capteurs produisant des données, ce sont les bases même de l'économie capitaliste qui vont s'effondrer selon (Rifkin, 2014). D'après ces théories, si les humains sont remplacés par des robots ou par des algorithmes, ils n'auront plus la capacité, faute de revenus, de consommer ce que les machines auront produit et on s'acheminera vers des crises catastrophiques de

surproduction remettant en cause le système capitaliste et son modèle fordiste qui espérait précisément éviter les crises de surproduction en indexant les salaires sur les gains en productivité et en permettant ainsi aux ouvriers de consommer d'avantage ce qu'ils ont produit.

Pour Michael Osborne et Carl Benedikt Frey, cités par le blog InternetActu.net, 47% parmi 702 métiers étudiés pourraient disparaître via l'automatisation[17]. Ce mouvement toucherait évidemment d'avantage les métiers moins qualifiés et les moins créatifs poussant les salariés vers des activités plus élevées. Ce mouvement s'accompagnerait aussi d'un développement des activités bénévoles et associatives, des hobbies et du *crowdsourcing* et conforte les partisans d'un revenu de base inconditionnel (ou "revenu universel" ou "revenu minimum d'existence" ou "allocation universelle") comme nous le verrons ultérieurement.

Au début du 19è siècle, le leader ouvrier John Ludd détruisit de nombreuses machines, puis, tout au long de cette période et jusqu'au 20è siècle, la classe ouvrière contesta l'automatisation du travail et le fordisme pour les mêmes raisons, sans supposer qu'un jour, la majeure partie des travailleurs ferait partie du secteur tertiaire des services. Aujourd'hui, Internet et l'ubérisation de l'économie peuvent provoquer le même type d'inquiétudes. La théorie de la destruction créatrice du célèbre économiste Joseph Schumpeter devrait pourtant inciter à l'optimisme. Cette théorie constate que dans l'économie, la disparition de secteurs d'activité va de pair avec l'apparition de nouvelles activités participant à l'évolution de l'économie.

La présente révolution du mode de production pourrait toutefois ne pas être "schumpetérienne" et pourrait détruire d'avantage d'emplois qu'elle n'en crée… Ainsi, selon Wendell Wallach, de l'Université de Yale, 47 % des emplois aux USA pourraient être remplacés par des algorithmes d'ici 10 à 20 ans. D'ailleurs, les géants de l'Internet n'embauchent que très peu d'employés, si on rapporte leur nombre à leurs chiffres d'affaires et au nombre de leurs clients. Sur ces sujets, on

[17] http://www.internetactu.net/2014/06/17/travail-et-automatisation-la-fin-du-travail-ne-touche-pas-que-les-emplois-les-moins-qualifies (consulté le 23 juin 2016)

trouvera une analyse relativement équilibrée entre croissance et décroissance dans la société postindustrielle chez l'économiste Daniel Cohen (Cohen, 2015).

1.5.1.2- *crowdsourcing*, revenu de base et théorie des biens communs

Le travail invisible des internautes pourrait être reconnu sous la forme d'un revenu inconditionnel afin de leur restituer la valeur qu'ils ont produite. Certains partisans d'un revenu de base inconditionnel qui serait financé par la TVA et versé à vie et sans conditions aux citoyens, estiment même que cette "contribution créative" permettrait de modifier leur rapport au temps libre en favorisant la création d'entreprises, mais aussi le travail non payé, le bénévolat et en leur permettant de mettre leur temps de travail contributif au service des autres sous la forme de *crowdsourcing*, par exemple. Pour cette raison, Bernard Stiegler préfère parler de "revenu contributif" (Stiegler, 2015). Au lieu de travailler afin d'en tirer un revenu et de craindre la perte de ce travail, on disposerait d'un revenu pour pouvoir se consacrer librement à l'activité de son choix. Motivés non plus par des besoins vitaux mais par des besoins plus élevés, cela permettrait aux individus d'être plus créatifs et innovants et de mieux coopérer. Ce mouvement répondrait à la destruction d'emplois par l'automatisation et engendrerait non pas une société de chômeurs, mais une société d'entrepreneurs libres et dépendants. Ce revenu serait conceptuellement équivalent à celui des ayants droits qui tirent profit de l'exploitation commerciale de l'œuvre de leur ancêtre jusqu'à 70 ans après sa mort. Les citoyens pourraient considérer qu'ils tirent profit de l'accumulation des connaissances accumulées par l'humanité comme d'un patrimoine immatériel. Il pourrait être considéré comme l'investissement, par la puissance publique, pour permettre aux contributeurs de poursuivre leur travail participatif. Avec l'automatisation, l'évolution du travail et le développement du travail bénévole invisible sous la forme de *crowdsourcing*, de nouvelles formes de rémunération pourraient émerger.

La notion de biens communs proviendrait de la campagne anglaise du 18è siècle qui était peu divisée en propriétés séparées et dont l'usage était partagé entre les communautés rurales. Michel Bauwens, théoricien du pair à pair,

remarque que "les entreprises fondent une partie de leur économie sur la rareté, ce qui est contradictoire avec la logique des biens communs" (Benyayer, 2014) et la contradiction travail / capital est en train d'être remplacée par la contradiction commun / capital. Ces théories ont toutefois été critiquées par la théorie de la tragédie des biens communs qui veut que l'accès libre et gratuit à une ressource engendre fatalement sa surexploitation et à sa destruction, l'usage étant individuel et le coût collectivement supporté, et l'intérêt individuel étant fatalement de la consommer au delà de ses besoins. Dans le domaine de la pêche, ce phénomène est illustré par des pêcheurs qui ont individuellement intérêt à prélever au maximum dans le stock commun au détriment de leur intérêt collectif et, à long terme, de leur intérêt individuel, une fois les stocks en ressources naturelles épuisés. Cette fatalité nécessiterait l'intervention de l'État pour empêcher qu'elle survienne. Mais, dans le cas du patrimoine numérique, le patrimoine est non "excluable", c'est un bien non rival, la ressource n'est pas limitée, son utilisation par un individu ne prive pas un autre individu de l'utiliser à son tour (Peugeot, 2012), son partage ne diminue pas, ne menace pas et ne divise pas la ressource, puisque, au contraire, il la multiplie à l'infini et pour un coût marginal quasiment inexistant.

L'émergence de monnaies virtuelles comme le BitCoin, conçu en 2009, pourrait favoriser le *crowdsourcing*. Ces monnaies virtuelles ont, en effet, tous les attributs d'une foule, la décentralisation et l'anonymat (Lebraty, 2015). Internet Archive, l'un des acteurs majeurs dans la numérisation participative, finance une partie du salaire de ses employés sous la forme de BitCoins. Cette monnaie virtuelle, convertible en dollars, permet de transférer de la valeur d'un internaute à l'autre et sans intermédiaires, mais aussi d'acheter des bons d'achats Amazon et des biens de consommation dans certains commerces.

Au 1er mai 2013, déjà près de 300 000 BTC étaient en circulation au prix unitaire de 94, 8 €. Le 13 mars 2013, le BTC se vendait à 47 $ et en 2012 à seulement 4,93 $

1.5.1.3- L'amateur, nouveau moteur de l'économie et du développement ?

Le terme d'amateur a été repris du français dans la langue anglaise. Aux 17è et 18è siècles, il désignait ceux qui pouvaient être élus à l'académie royale de peinture sans pour autant être peintres, mais en raison de leur passion pour l'art. Aujourd'hui, il désigne des personnes qui n'agissent que par amour de telle ou telle discipline mais est aussi employer de manière péjorative afin de discréditer des contributeurs pour leur manque de professionnalisme, qualifié d'ailleurs d'amateurisme.

La figure de l'amateur semble se décliner en 2 types distincts : celui qui va organiser sa vie professionnelle et sociale autour de sa passion de telle sorte qu'elles ne la contrarient pas ou mieux, concorde avec elle jusqu'à la professionnaliser. Celui qui considère cette activité comme étanche de sa vie professionnelle et sociale, allant même parfois jusqu'à en faire une activité secrète cachée de son entourage familial ou professionnel. Il existe ainsi des amateurs extravertis qui recherchent une reconnaissance sociale et obéissent à une logique de réseaux et des amateurs introvertis qui agissent d'avantage comme des bénévoles désintéressés et obéissent à une logique de communauté.

Avec le développement du *crowdsourcing*, nous pourrions passer d'un modèle d'innovation, tel que décrit par Joseph Schumpeter, partant du producteur actif vers le consommateur passif, l'entreprise étant à l'avant garde de la modernisation et cherchant à faire évoluer les utilisateurs, à un modèle d'innovation centrée sur les utilisateurs actifs qui feraient remonter leurs idées à des entreprises qui s'en inspireraient. La séparation entre la production et la consommation semble ainsi disparaître progressivement. Les utilisateurs ne veulent plus être de passifs consommateurs et considèrent qu'on ne possède pas vraiment quelque chose si on ne peut pas l'ouvrir, ils veulent agir et se regroupent de plus en plus en réseaux collaboratifs, ils échangent, bricolent, améliorent les objets de consommation sur le mode du "DIY" (do it yourself), innovent et influencent les entreprises sans attendre d'elle d'autre retour que la satisfaction de voir leur idées se concrétiser. Ils développent toute une culture "maker". Dans le domaine scientifique, on rencontre, par exemple, une "biologie de garage", portée

principalement par l'association DIYbio (Do-it-yourself biology) ou, en France, par l'association "La Paillasse", associations qui ouvrent la science et ses moyens aux amateurs. L'innovation devient le résultat de la collaboration entre les producteurs et les consommateurs qui en deviennent les coproducteurs et les coauteurs.

Ainsi, selon Eric Von Hippel qui parle d'innovations centrées sur l'usager, d'innovations par l'usage ou d'innovations ascendantes, 46 % des entreprises américaines dans des secteurs innovants ont pour origine un utilisateur, la plupart du temps, un homme jeune et diplômé bénéficiant d'une culture technique (Von Hippel, 2011) qui, ne trouvant pas le service ou le produit dont il a besoin sur le marché car les sociétés traditionnelles ne sont pas toujours organisées pour faire du sur mesure ou prêtes à risquer d'investir pour une demande incertaine. Ce passionné est généralement prêt à investir du temps et de l'argent pour le développer, à fabriquer plutôt qu'à acheter, à produire plutôt qu'à consommer et est prêt à partager gratuitement ses découvertes. Il a accès à des moyens informatiques et des technologies de plus en plus avancées et la production de son prototype lui est de moins en moins coûteuse. Ainsi, le skateboard a été inventé par des consommateurs, et 80 % des innovations dans les instruments scientifiques ont été développés par les usagers et on ne compte plus les développements qui sont le fruit des usagers dans le monde du logiciel libre (Von Hippel, 2005). Dans le domaine des bibliothèques, les fonctionnalités des SIGB et des OPAC produits par les prestataires ont ainsi, en grande partie, été inspirées par les clubs d'utilisateurs composés de bibliothécaires comme le signale toujours Von Hippel. Généralement, les "lead users" ou utilisateurs leaders bricolent un produit pour leurs besoins, ce produit est repris, copié et amélioré par d'autres consommateurs et le succès est tel que les entreprises finissent par s'y intéresser. Bien au delà des traditionnelles études de marché, les sociétés auraient donc tout intérêt à anticiper et à collaborer avec ces utilisateurs leaders en leur proposant des boîtes à outils, des forums, des réseaux sociaux, des plateformes. D'après (Von Hippel, 2011), les consommateurs-innovateurs représentaient quand même 6,1 % de la population de plus de 18 ans au Royaume Uni, 5,2 aux USA et 3,7 % au Japon.

Avec le développement du logiciel libre en particulier, les usagers sont de plus en plus reconnus comme une source possible d'innovations. La société Dell, par exemple, a lancé le site Ideastorm et a recueilli plus de 10 000 propositions d'idées pour améliorer ses produits et ses services. Avec son projet Techshop, largement ouvert aux idées des consommateurs, la société Ford a augmenté ses dépôts de brevets de 30 %. En France, les Fab Lab fonctionnent sur le même principe de "*user innovation*".

En comparant les idées provenant de professionnels avec celles venant d'usagers, (Poetz, 2012) rapporte, sans surprise, que d'après son étude, elles seraient plus innovantes (note moyenne de 2, 6 contre 2, 12) et plus avantageuses pour les consommateurs (note moyenne de 1,86 contre 2,44), mais également assez faibles en terme de faisabilité, les idées des professionnels ayant tendance à être beaucoup plus faciles à réaliser (note moyenne de 4,33 contre 3,91).

L'histoire des sciences est d'ailleurs remplie d'inventions venant de personnes extérieures au métier qui ne cherchent pas à reproduire les modèles établis avec lesquels ils ont été formés et qui sont susceptibles de provoquer des ruptures innovantes. Dans la nouvelle économie, il semble, en effet, que les entreprises doivent de plus en plus se connecter aux idées et aux énergies externes et intégrer le consommateur dans le processus de production (Ligeon, 2012).

Le *crowdsourcing* permet de créer un écosystème de l'innovation en faisant travailler des personnes de compétences et d'horizons très différents autour de projets communs et avec l'aide des nouvelles technologies.

1.5.2- Les usagers du *crowdsourcing*

Les “crowdsourcers”, “clickworkers” et autres “prolétaires du web” sont parfois comparés à des “Oompa-loopas”, une tribu travaillant en s’amusant pour le chocolatier Willy Wonka en échange de chocolat dans le roman de Roald Dahl, Charlie et la chocolaterie (Renault, 2014). Tous ces travailleurs forment-ils une catégorie socioprofessionnelle voir une classe sociale ? Va-t-on voir émerger une classe de “prosumers”, de “produser”, de “prosommateurs”, c’est à dire, une classe d’individus à la fois producteurs et consommateurs-usagers de leurs propres produits, plus attachée à l’usage partagé des biens qu’à leur appropriation privée, comme le prophétise Jérémy Rifkin ?

L’émergence de la génération Y ou digital natives dans l’entreprise qui représenterait déjà 40 % des actifs en France pourrait également avoir une influence sur le développement du *crowdsourcing*. Cette génération bouscule les hiérarchies, les autorités et les références, sa culture est d’avantage ouverte et participative. Il est donc probable qu’elle soit d’avantage perméable au *crowdsourcing* comme l’illustre le diagramme du pourcentage de contributeurs à Wikipédia par date de naissance :

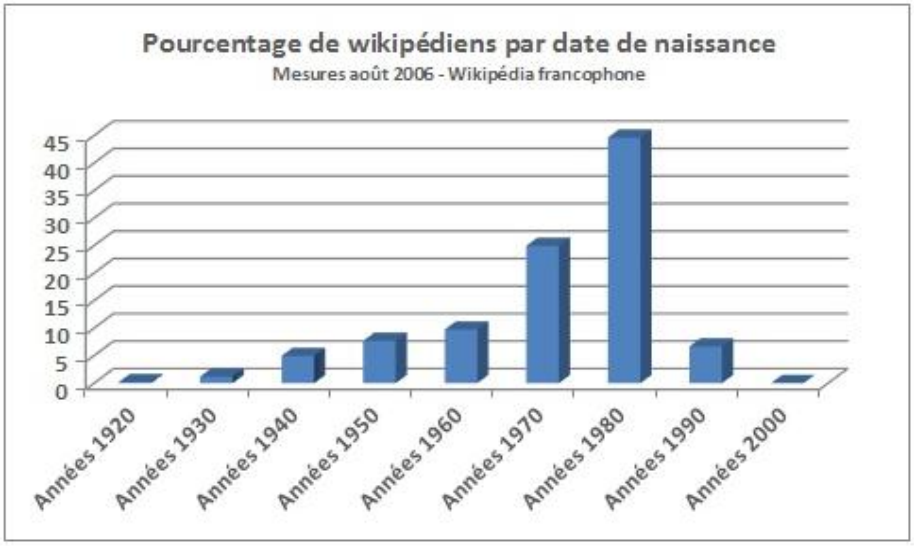


Figure 18. Pourcentage de wikipédiens par date de naissance, d’après Wikipédia

Les jeunes générations ont tendance à construire leur identité par leur participation sur le web, la rédaction de blogs, la mise en place ou la participation à des forums de fans. L'amateurisme et le *crowdsourcing* pourraient être un moyen pour eux de soigner leur *e-reputation* et de mettre en avant leur participation à des projets culturels, participation qui pourrait également être bénéfique pour leur CV et leur recherche d'emploi.

Mais les générations précédentes pourraient également être concernées par le *crowdsourcing*. Si le rapport (Beuth Hochschule für Technik, 2014) indique que 26 % des Wikipédiens ont entre 22 et 26 ans, il indique aussi que 28 % d'entre eux sont âgés de plus de 40 ans et qu'ils sont 36 % parmi cette catégorie plus âgée à faire partie des plus actifs. En effet, les retraités sont nombreux, ils disposent d'un temps disponible important et sont déjà assez actifs dans les associations et le bénévolat et, pour ce qui concerne le *crowdfunding*, ils disposent d'un capital disponible plus important que les générations qui les précèdent, qui ont souvent besoin de rembourser leur crédit et qui disposent moins souvent de revenus fonciers, et de placements financiers).

Le temps libre de toutes ces catégories de populations est un formidable réservoir de bonnes volontés pour le *crowdsourcing*. Chaque minute, 35 heures de vidéos sont ainsi mises en ligne sur YouTube, et chaque heure 38 400 photos sont postées sur Flickr, d'après le rapport "*crowdsourcing* in the cultural heritage domain : opportunities and challenges". (Paraschakis, 2013) évoque, quant à lui, les chiffres de 72 heures de vidéos qui seraient mises en ligne sur YouTube chaque minute et 2500 photos sur Flickr.

La théorie des communautés est un cadre conceptuel qui peut être mobilisé afin d'analyser le crowdsourcing d'un point de vue sociologique. Une communauté en ligne ou une communauté virtuelle de pratique est un groupe relativement homogène d'internautes qui travaillent ensemble au bénéfice d'une entreprise commune de manière relativement auto-organisée et informelle, qui s'entraident afin de résoudre des problèmes pratiques sous la forme d'un engagement mutuel et qui partagent un répertoire, c'est à dire, un patrimoine d'informations (Wenger,

1998). La communauté virtuelle est généralement structurée autour d'un noyau dur de membres actifs. Une culture de communauté est susceptible de se développer avec une identité commune, des références partagées et des règles implicites. Cette culture est transmise aux novices par les leaders ou par les seniors. (Daele, 2009). Lorsque cette communauté est plus hétérogène, créative, produit des connaissances nouvelles reconnues et faisant autorité auprès de la communauté scientifique et cherche à ce que cette connaissance ne consiste pas exclusivement à améliorer une pratique mais que ce soit une « connaissance utilisable » (Meyer, 2011), c'est à dire ayant une influence sur les politiques publiques, on parle alors plutôt de communauté épistémique (Millerand, 2011). Une communauté de pratique peut se transformer graduellement en communauté épistémique (Lièvre, 2014). Le type de communauté sera évidemment variable en fonction du type de crowdsourcing en présence. La communauté épistémique se rencontre toutefois d'avantage dans le cadre des sciences citoyennes ou de Wikipédia que dans les projets de bibliothèques numériques que nous avons identifiés.

1.6- Conséquences managériales, bibliothéconomiques et technologiques

1.6.1- L'exception française

De manière très générale, dans la culture anglo-saxonne, le partage de l'information par des communautés d'intérêts est relativement naturel. Ce n'est pas encore toujours le cas dans les institutions françaises. Ces dernières devront passer par une évolution culturelle majeure afin de s’adapter à ces nouveaux modèles. Une étude Deloitte Digital Collaboration commandée par Google en 2013 montre ainsi que si 62 % des salariés français considèrent que les outils collaboratifs améliorent la communication et la productivité, ils ne sont que 11 % à les utiliser dans le cadre de leur entreprise contre 18 % en Europe, 25 % aux Pays Bas, 21 % en Suède, 19 % en Allemagne, 18 % en Italie et 16 % au Royaume Uni.

Une autre étude du cabinet de conseil McKinsey place ainsi la France seulement en 8ème position sur les 13 pays étudiés pour ses pratiques numériques et l'explique notamment par des difficultés à opérer les changements de management et d'organisation dans les entreprises françaises et par le manque d'implication de ses dirigeants (McKinsey, 2014).

En attendant, il semble bien que le facteur culturel ait son importance dans l'adoption du *crowdsourcing* (Estermann, 2015). (Moirez, 2013) constate, par exemple, que les projets Web 2.0 et en particulier, les projets de *crowdsourcing* en bibliothèque fonctionnent moins bien en France que dans les pays anglo-saxons en raison de différences culturelles. (Boeuf, 2012) fait le même constat d'une difficulté de développement des sciences citoyennes en France en raison d'une plus faible implication des individus de culture latine par rapport aux individus de culture anglo-saxonne dans la vie collective et d'une plus grande méfiance, d'une crainte d'être instrumentalisé ou d'être inutile dans un projet qui ne le serait pas d'avantage.

1.6.2- Les bibliothèques françaises, exception dans l'exception ?

Les bibliothèques ont progressivement vu la remise en question des magasiniers avec le développement du libre accès, la remise en question des catalogueurs avec le développement d'un catalogage mutualisé à l'échelle mondiale, et enfin, la remise en question des acquisitions avec le développement des périodiques électroniques puis des ebooks. Comme le rapporte Clémence Just dans Archimag le 21 juillet 2015, des chercheurs de l'Université d'Oxford, estiment à 64,9 % la probabilité que le métier de bibliothécaire soit automatisé prochainement.

Les bibliothèques restent parfois éloignées du monde de l'entreprise et leurs cadres considèrent souvent l'intérêt public comme plus éthique que la recherche de profits. Le *crowdsourcing* pourrait donc y être considéré comme une forme de privatisation ou comme une forme renouvelée et alternative de partenariat public / privé (McShane, 2011).

La différence de conception entre les bibliothèques françaises et les bibliothèques anglo-saxonnes remontrait à la révolution française. En effet, la mission des bibliothèques françaises était singulièrement plutôt de protéger les confiscations révolutionnaires contre les foules dangereuses susceptibles d'en menacer l'intégrité et de les réserver à la nouvelle classe dominante, la bourgeoisie. Il s'agissait plus de « protéger le livre sacré des foules violentes que de diffuser les collections auprès d'un large public » comme le dit si bien (Breton, 2014). Dans ces conditions, les conceptions françaises et anglo-saxonnes des collections sont relativement différentes et l'implication du public comme l'externalisation de tâches auprès des foules, qui est la définition du *crowdsourcing*, aurait par la suite bien des difficultés à s'imposer en France. Dans la conception française, la collection est une œuvre intellectuelle et cohérente et la politique documentaire, la prérogative et la chasse gardée des bibliothécaires qui sont garants de la neutralité du service public au nom de l'intérêt général dans une vision verticale, jacobine et centralisée du monde avec la volonté de guider les lecteurs et d'encadrer les foules. Dans cette conception, la collection est plus importante que les lecteurs, l'offre prime sur la demande. Dans la conception anglo-saxonne, la collection ne doit répondre qu'à la nécessité de satisfaire les besoins en lecture d'une population dans une vision plus horizontale et décentralisée. Dans cette conception, les lecteurs sont plus importants que la collection, la demande prime sur l'offre. Il existe d'ailleurs dans les bibliothèques d'inspiration anglo-saxonnes, des postes de "*community librarians*" chargé de manager les relations avec la population et n'ayant aucun équivalent en France (Breton, 2014).

Il faut bien constater que la profession de bibliothécaire vit avec difficulté la remise en cause de son monopole. Les bibliothèques ne sont plus l'intermédiaire incontournable entre l'information et le public. Ce sentiment, qui n'est jamais clairement explicité, n'est pas sans rappeler le sentiment des *gate keepers*, les gardiens de l'ordre culturel et politique établi décrits par (Cardon, 2010). Selon ce chercheur, les médias cherchent à maintenir leurs privilèges, leur contrôle et leur monopole d'accès à l'information par un peuple jugé insuffisamment responsable

et éclairé pour se forger une opinion de manière autonome. Internet est donc, pour eux, une menace du modèle vertical et monopoliste de diffusion de l'information qu'ils ont forgé et une remise en question de leur autorité. De la même manière que les droits divins, les corporations et les ordres ont été balayés par le mouvement révolutionnaire du 18è siècle, l'information qui était produite par quelques uns (dont les auteurs, les journalistes, les éditeurs et les bibliothécaires) dans le web 1.0 est désormais produite par la multitude avec le web 2.0.

Pour une institution culturelle comme une bibliothèque, accepter d'ouvrir aux amateurs son indexation, son catalogage, son choix de documents à numériser... demande une évolution culturelle majeure. Il s'agit, en effet de passer d'une politique de l'offre centrée sur les collections et les activités des bibliothécaires à une politique de la demande centrée sur les services, les besoins et les activités des usagers puis directement déclenchée et conduite par l'initiative de l'usager individuel lui même et qui correspond bien aux modèles "à la demande". L'usager devient ainsi un acteur central des politiques de numérisation des bibliothèques, jusqu'ici réservé à ses professionnels (Klopp, 2014). Selon ce point de vue, de dépositaires du patrimoine imprimé, les bibliothèques devraient devenir des acteurs de la valorisation du patrimoine qui soient associés aux internautes.

Du coté des professionnels, la mise en place d'une démarche *crowdsourcing* dans une institution culturelle peut néanmoins, à juste titre, être ressentie comme dévalorisante pour le travail des conservateurs et des documentalistes qui pourrait être déprécié car réalisé gratuitement et par n'importe qui. Cette évolution nécessite donc un investissement important en conduite du changement et en communication interne. Comme Ben Brumfield le rapporte dans son blog, manuscripttranscription.blogspot.fr, dans le cadre du Manuscript Fragments Project développé par le Harry Ransom Center's, près de 20 % des commentaires (autour de la transcription, des sources, ou de l'identification de fragments) reçus à propos des manuscrits médiévaux proviennent de professionnels, mais ceux-ci ont tous préféré envoyer les courriels plutôt que de contribuer directement en ligne. Ceci peut probablement s'expliquer par la

nécessité de préserver leur réputation et par la peur de mettre leur contribution sur le même plan que celle du profane et de voir leurs compétences discutées par eux, leur autorité partagée avec eux. Le *crowdsourcing* est généralement ressenti par les professionnels à la fois, comme une perte de contrôle sur les choix des documents qui seront numérisés (en particulier avec la numérisation à la demande), sur la manière dont le patrimoine va être exposé et utilisé et, à la fois comme un engagement contraignant vis à vis des contributeurs, les résultats de leur travail devant être accessibles de manière pérenne.

Néanmoins, l'engagement des amateurs privés peut, sur certains sujets, apporter des contributions au bénéfice des institutions publiques et compléter utilement le travail des professionnels dont les effectifs, les moyens et les connaissances restent limités malgré toutes les bonnes volontés. Grâce au crowdsourcing, les bibliothèques peuvent puiser dans une foule illimitée d'internautes pouvant contenir de réels spécialiste de tels ou tels sujets connaissant bien mieux le contenu et l'intérêt de tel ou tel livre

Le *crowdsourcing* est un modèle centré sur l'utilisateur. Héritier du web 2.0, il est plus interactif, réciprocitaire et moins hiérarchique que les modèles « top down » de diffusion des connaissances. Néanmoins, nombre d'institutions ne sont pas suffisamment centrées sur leurs usagers et demeurent dans des logiques d'offres. Elles ne soucient qu'insuffisamment de la demande. Pourtant, comme le mentionne (Levi, 2014), les institutions culturelles pourraient dorénavant se concentrer sur l'agrégation et la livraison du patrimoine numérisé tandis que les internautes pourraient eux, se charger de l'enrichissement des métadonnées. Cela impliquerait une révolution culturelle dans la profession car les archivistes privilégient les collections et leur description perfectionniste par rapport à leurs utilisateurs. Il s'agirait, au contraire, de privilégier l'accès aux contenus par les usagers, même si ceux-ci ne sont pas encore décrits, et afin qu'ils le soient bénévolement. (Nguyen, 2012) invite ainsi les bibliothécaires à donner plus de pouvoir à leurs lecteurs, à encourager leur participation et à développer une véritable culture de la participation.

Avec le *crowdsourcing*, si l'internaute devient un bibliothécaire, le bibliothécaire et le conservateur de bibliothèque peuvent se sentir rabaissés au commun des internautes. Or, la société française se fonde sur l'idée aristocratique d'élection ou de délégation. Chaque domaine a ses spécialistes, ses experts dont la légitimité et l'autorité sont désormais remises en question. Les corps des conservateurs et des bibliothécaires ne font pas exception et le *crowdsourcing* pourrait être le nom donné à leur remise en cause par la masse des internautes anonymes et parfois incompétents, le nom donné à leur « ubérisation ».

1.6.3- Le règne de l'amateur : vers une médiocratie ?

Perte de monopole, de contrôle, de pouvoir, risque de mauvaise qualité, de malveillance, perméabilité aux lobbys et aux idéologies, risque de prise de contrôle par des minorités non représentatives, remise en cause de l'expertise des professionnels, perte de responsabilités... les raisons ne manquent pas de s'opposer à l'essor annoncé du *crowdsourcing* en bibliothèque.

En effet, les professionnels et les experts ayant produit des métadonnées dans un cadre formel, institutionnalisé, collectif et reconnu risquent de ne pas favoriser une diffusion sur le web engendrant une redocumentarisation participative. Celle-ci peut être synonyme d'appropriation personnelle et individuelle du patrimoine collectif par une poignée d'internautes se sentant autorisés à laisser leurs traces, à taguer ou à donner leur points de vues profanes, informels, personnels, intimes, banals, lambdas, triviaux et médiocres.

Les commentaires sur les images sont souvent du type "Excellent", "Superbe", "WOW !", "Sympa !", "Parfait !" etc. (Liew, 2014) et sont, par conséquent, parfaitement inexploitables.

Le terme d'amateur lui-même est ambivalent. Il peut désigner à la fois celui qui aime ou le non professionnel qui travail mal. L'amateur est un passionné qui consacre une part importante de son temps à sa passion et qui ne recherche pas d'autre récompense que la reconnaissance. Fondamentalement, ce qui distingue le professionnel de l'amateur est la connaissance des méthodologies et des normes de catalogage et de descriptions bibliographiques, des règles d'indexation

ou de transcription diplomatique, des standards d'encodage TEI ou EAD... Permettre au profane et au néophyte d'accéder à ces savoirs pourrait revenir à dévaloriser ces compétences et à exposer que ces savoirs ne reposent parfois sur aucune science particulière mais sur un ensemble de règles qui peuvent aussi être pour partie arbitraires. En acceptant que les amateurs puissent acquérir ces connaissances, les professionnels convertiraient pourtant les amateurs en semi-professionnels et donc en défenseurs de leurs intérêts professionnels.

Comme le souligne Rose Holey, lorsque les bibliothèques ne proposaient encore à leurs usagers que des documents imprimés, les lecteurs aimaient déjà échanger avec le bibliothécaire de référence ou avec d'autres lecteurs, partager des documents, mais il ne lui était pas possible d'annoter un livre sous peine d'exclusion. Avec le passage des bibliothèques de livre imprimé vers les bibliothèques électroniques, les lecteurs devraient pouvoir ne pas être de simples consommateurs d'information mais être également producteurs et surtout des collaborateurs pour les professionnels de l'information. Ainsi, ils peuvent désormais ajouter un résumé du livre ou de l'article qu'ils ont lu, le partager avec leurs réseaux sociaux, ajouter des informations, des métadonnées, des commentaires, des annotations, corriger des erreurs de métadonnées, converser avec d'autres usagers et même organiser avec eux un travail collaboratif. Les amateurs et les professionnels peuvent désormais collaborer, d'autant que les amateurs qui collaborent acquièrent assez rapidement un bon niveau d'expertise. En acceptant de recourir à ces foules d'amateurs, les bibliothèques trouveraient peut être ainsi plus facilement un expert sur tel ou tel sujet qu'au sein d'équipes réduites de conservateurs (Huvila, 2008) ? Quoi qu'il en soit, le *crowdsourcing* a déjà un énorme avantage par rapport à la traditionnelle externalisation déjà largement pratiquée par les bibliothèques qui ont recours à la main d'œuvre à bas coût comme à Madagascar, car un généalogiste érudit et passionné devient généralement rapidement plus compétent et connaît mieux le sujet qu'un sous-traitant d'un pays à bas coût dont la langue et la culture sont plus éloignées et qui risque de ne travailler sur le projet que pendant une durée très courte.

1.6.4- Le *crowdsourcing*, stade suprême de l'externalisation ?

Comme nous l'avons précédemment et largement évoqué, le *crowdsourcing* s'inscrit dans le mouvement économique de flexibilisation et d'externalisation qui a commencé avec la sous-traitance de pans entiers de la production dans des pays à main d'œuvre plus compétitive ou auprès de fournisseurs, de consultants, ou même parfois de salariés de l'entreprise devenus des travailleurs indépendants ou des auto-entrepreneurs. Avec le *crowdsourcing* on externalise désormais sur le web. Certains parlent même ainsi d'"externalisation ouverte". Au lieu d'externaliser auprès d'un sous-traitant déterminé dans un pays à bas coût, le *crowdsourcing* est une externalisation auprès d'une foule d'internautes anonymes de tous pays.

Dans une conjoncture difficile pour les bibliothèques, le *crowdsourcing* peut également s'avérer être un moyen de faire plus avec de moindres moyens. Dans le domaine de la numérisation, en particulier, on a assisté, ces dernières années à une externalisation du travail de correction de l'OCR ou de saisie des métadonnées vers des pays à bas coûts (Madagascar, Inde, Viêt-Nam...). Cette externalisation a permis à des prestataires de numérisation de diminuer les coûts et de proposer des services plus performants en se développant à l'étranger (Diadéis, Jouve, Aurexus), en faisant appel à des sociétés étrangères déjà constituées ou en bénéficiant de spécialistes dont disposent certains pays étrangers. L'externalisation est l'occasion aussi pour une société de s'ouvrir sur d'autres cultures de travail et d'enrichir ses propres procédures. Le *crowdsourcing* est une forme d'externalisation qui ne se soucie pas du lieu où travaille le contributeur, la seule condition requise étant d'être relié au réseau mondial du web. En effet, Jeff Howe dans l'article "the rise of *crowdsourcing*" publié dans Wired Magazine en 2006 et qui allait populariser le terme de *crowdsourcing*, indiquait que pendant les 10 dernières années les entreprises avaient cherché à délocaliser vers des pays où la main d'œuvre était meilleur marché comme l'Inde ou la Chine, mais que le lieu où se trouvent les employés pourrait avoir de moins en moins d'importance dans l'avenir, dans la mesure où ils sont connectés au réseau. En effet, pourquoi délocaliser dans des pays à bas coûts alors que via les réseaux, il est désormais possible de mobiliser, pour des coûts très faibles ou nuls, une main

d’œuvre plus diversifiée, motivée, qualifiée et compétente ? Pour les bibliothèques par exemple, cette diversité est un atout majeur car il devient possible de bénéficier des compétences de spécialistes de tel ou tel domaine bien au delà des limites d’équipes restreintes de conservateurs qui malgré une bonne culture générale, ne pourront jamais être spécialistes de toutes les disciplines. Elle permet, en outre, de développer la pluridisciplinarité. Comme l’affirme Nicolas Colin dans un propos rapporté par le blog Internet Actu, “il y a désormais plus de puissance à l’extérieur qu’à l’intérieur des organisations”. Le *crowdsourcing* pose ainsi aussi la question des frontières de l’organisation puisqu’il permet de créer de la valeur au delà de ses frontières (Renault, 2014bis et Lebraty, 2015).

La frontière entre ce qui peut être fait par l’intelligence artificielle des machines et ce qui doit être fait par celle de l’homme est en perpétuelle évolution. Pour le moment, le travail confié aux hommes ne l’a été que parce qu’il ne pouvait l’être auprès de machines.

Mais, cette externalisation pourrait aussi être une première phase vers la suppression de certains services publics culturels après avoir fait la démonstration de sa faisabilité et après les avoir réduits à une forme de mendicité sur le web. En effet, dans un contexte de désengagement de l’État, pourquoi continuer à payer des professionnels à réaliser un travail que des amateurs sont prêts à faire bénévolement ? Le crowdsourcing pourrait donc s’apparenter à une forme d’ubérisation de services publics. Avec le *crowdsourcing* en bibliothèques, on pourrait assister à une « ubérisation » des bibliothèques, c’est à dire à un remplacement du service produit par un professionnel par celui d’un amateur. A l’instar des autres formes d'ubérisation, il pourrait donc aussi provoquer des réactions d’hostilités.

Ce chapitre conceptuel a fait l’objet d’un article (Andro, 2014, 1)

Chapitre 2- Panorama de quelques projets de *crowdsourcing* appliqués à la numérisation des bibliothèques

Dans ce panorama, des analyses synthétiques sont données autour des grands types de tâches qui peuvent être confiés aux internautes. Afin d'illustrer chaque type, nous avons également sélectionné un seul projet représentatif du type. Un complément plus exhaustif à ce panorama des projets est accessible en annexes. Nous distinguons ainsi la mise en ligne et curation participatives, la numérisation à la demande par crowdfunding, l'impression à la demande, la correction participative de l'OCR et la folksonomie.

2.1- Mise en ligne et curation participatives : l'Oxford's great war archive et Europeana 1914-1918

La mise en ligne et la curation participative consistent à permettre à des internautes de compléter les collections numériques institutionnelles avec leurs exemplaires ou leurs propres sélections.

L'Université d'Oxford au Royaume-Uni a créé, en 2008, une archive qui contient 6500 images numérisées grâce à la contribution des citoyens anglais ayant apporté leurs archives personnelles sur la grande guerre, leurs lettres de famille, photographies, souvenirs de guerre à numériser, avec des notices rédigées par le grand public. Ces documents d'archives privées ont ainsi permis de compléter les collections publiques.

Le succès de ce projet a encouragé Europeana à mobiliser d'autres institutions nationales et locales à travers l'Europe dans un partenariat avec l'Université d'Oxford. Ainsi, "Adding your story to Europeana 1914-1918" s'est inspiré de cette initiative pour collecter des souvenirs de la Grande Guerre dans plusieurs pays européens.

En France, du 9 au 16 novembre 2014, plus de 70 points de collecte dans toute la France ont permis de mener une opération similaire avec l'aide

d’institutions volontaires qui ont pu participer à cette opération “grande collecte” en mettant à disposition du personnel et des ateliers de numérisation.

Nous n’évoquons ici que la co-construction de bibliothèques numériques et non la co-construction des collections physiques et des acquisitions participatives par des comités d’usagers de livres imprimés, notre périmètre s’arrêtant aux seules bibliothèques numériques.

La possibilité pour des internautes et des bénévoles de compléter les collections publiques avec la numérisation de leurs propres collections patrimoniales remet en question la notion de collections résultant du travail de sélection des bibliothécaires. L’ouverture des politiques documentaires des bibliothèques aux internautes représente donc une évolution majeure de leurs missions. La mise en ligne et la curation participatives rejoignent la numérisation à la demande par *crowdfunding* qui sera abordée dans le chapitre suivant, dans le sens où l’internaute devient un acteur de la politique documentaire et de la constitution de collections, mais, contrairement au *crowdfunding*, cette participation s’arrête à la sélection documentaire ou à la mise à disposition de documents et ne va pas jusqu’au financement de la numérisation elle-même.

2.2- la numérisation à la demande sous forme de *crowdfunding* appliquée aux bibliothèques numériques : le réseau européen ebooks en demand (EOD)

Le *crowdfunding* est généralement considéré comme une forme de *crowdsourcing* faisant appel non pas au travail et au génie des internautes mais aux ressources financières de ces "foules aux œufs d'or" (Onnée, 2014). Brabham considère, au contraire, plutôt le *crowdfunding* comme un mode de financement alternatif qui, à l'inverse du *crowdsourcing*, ne permet pas aux internautes de peser sur la politique du projet. Néanmoins, comme le montre (Onnée, 2014), les projets de *crowdfunding* font finalement souvent aussi appel à des formes *crowdsourcing* (votes, aide à la promotion sur les réseaux sociaux...) afin que le projet soit "emporté par la foule".

Selon (Onnée, 2013) qui en donne une définition, le *crowdfunding* « *consiste pour un porteur de projet (quel que soit son statut : particulier, organisation marchande ou non marchande, etc.) à avoir recours aux services d'une plateforme de financement (généraliste ou spécialisée) afin de proposer un projet (finalisé ou non) auprès d'une communauté (large ou ciblée) de contributeurs qualifiés de soutiens (backers) en échange éventuellement de contreparties préalablement définies* ». Cette communauté est généralement recrutée via les réseaux sociaux. En France, les plateformes Ulule [18] et KissKissBankBank occupent une position de leader sur ce marché.

Du coté des bibliothèques, la numérisation à la demande est avant tout un service rendu à l'usager, mais elle est aussi, pour les bibliothèques, un moyen d'externaliser le financement de leur numérisation et de compléter ainsi leurs programmes de numérisation. Le financement de la numérisation d'un document peut être individuel ou collectif. Elle peut être motivée par le besoin individuel d'accéder à un document difficilement accessible, par l'envie de soutenir

[18] La plateforme Ulule en particulier semble être beaucoup utilisée dans le domaine de la conservation du patrimoine http://www.club-innovation-culture.fr/crowdfunding-patrimoine-realise-2015 (consulté le 23 juin 2016)

financièrement une institution ou par la volonté de rendre plus accessible et de promouvoir une œuvre particulière.

C'est donc bien une forme de *crowdfunding* appliqué aux projets de numérisation des bibliothèques.

Le réseau européen ebooks on demand (EOD), lancé en 2006 dans le cadre du projet européen eTEN (2005-2008), et piloté par la Bibliothèque universitaire du Tyrol, permet aux bibliothèques qui y participent de disposer d'une plate-forme de paiement pour mettre en place leurs services de numérisation à la demande.

Chaque bibliothèque est invitée à ajouter dynamiquement des boutons EOD dans son catalogue en ligne les titres éligibles à une numérisation, c'est à dire, généralement antérieurs à une date déterminée, généralement 1900. Le catalogue national des bibliothèques de l'Enseignement Supérieur SUDOC propose également aux bibliothèques qui y participent d'ajouter un bouton vers Ebooks on Demand. Les usagers intéressés par tels ou tels titres du catalogue des bibliothèques peuvent appuyer sur ces boutons et être renvoyés vers une plateforme de paiement afin d'obtenir le PDF de l'exemplaire papier correspondant. Les documents numériques sont ensuite diffusés, après un délai moyen d'une semaine, via le Digital Object Generator qui génère un PDF multicouches avec OCR et une couverture présentant le service. Après un délai d'environ 2 mois d'embargo pendant lequel le commanditaire bénéficie seul du document numérique (6 mois à la Bibliothèque Nationale et Universitaire de Slovénie), la bibliothèque diffuse le document numérisé au sein de sa bibliothèque numérique et ajoute un lien sur la notice de son catalogue en ligne.

A la différence du projet Numalire qui a fait l'objet des expérimentations de cette thèse[19], la prestation est généralement effectuée par les services de numérisation des bibliothèques, non par un prestataire extérieur bien que ce ne soit pas obligatoirement le cas[20]. Cela a pour avantage de permettre des prix plus

[19] Voir le chapitre dédié aux expérimentations

[20] A l'exception notable de la bibliothèque bavaroise qui a déjà eu recours à un prestataire extérieur

bas, le temps de travail des agents publics n'étant généralement pas intégralement répercuté sur les prix, mais cela a aussi pour inconvénient de contraindre les bibliothèques qui souhaitent participer au projet à se doter d'ateliers de numérisations qui nécessitent un matériel coûteux et un personnel qualifié. Or, numériser par ses propres moyens est un métier et nécessiterait des moyens dont la plupart des institutions ne disposent pas et dont elles ne souhaitent pas forcément disposer dans la mesure où la numérisation est plus compétitive lorsqu'elle est prise en charge par un prestataire privé disposant d'une productivité, d'une intensité et d'une durée du travail souvent supérieure afin d'amortir de lourds investissements en machines sur des durées plus courtes.

En 2009, on comptait 13 institutions participantes dans le réseau. Aujourd'hui, 30 bibliothèques dans 12 pays européens y participent :

- Autriche (University of Innsbruck, Library (co-ordinator), University Libraries of Graz and Vienna, Library of the Medical University of Vienna, Vienna City Library, St. Pölten Diocese Archive
- Allemagne (University Libraries of Regensburg, Greifswald, Leipzig and Humboldt-Universität zu Berlin, Bavarian State Library, Saxon State and University Library (Dresden))
- Danemark (The Royal Library)
- Estonie (National Library, University Library of Tartu)
- France (Bibliothèque Inter Universitaire de Santé (BIUS), Bibliothèque Nationale Universitaire de Strasbourg (BNUS))
- Hongrie (National Széchényi Library of Hungary, Library of the Hungarian Academy of Sciences)
- Portugal (National Library)
- République tchèque (Moravian Library in Brno, Research Library in Olomouc, Library of the Academy of Sciences in Prague, National Technical Library)
- Slovaquie (University Library in Bratislava, Slovak Academy of Sciences)
- Slovénie (National and University Library)
- Suède (Umeå University Library)
- Suisse (The Swiss National Library)

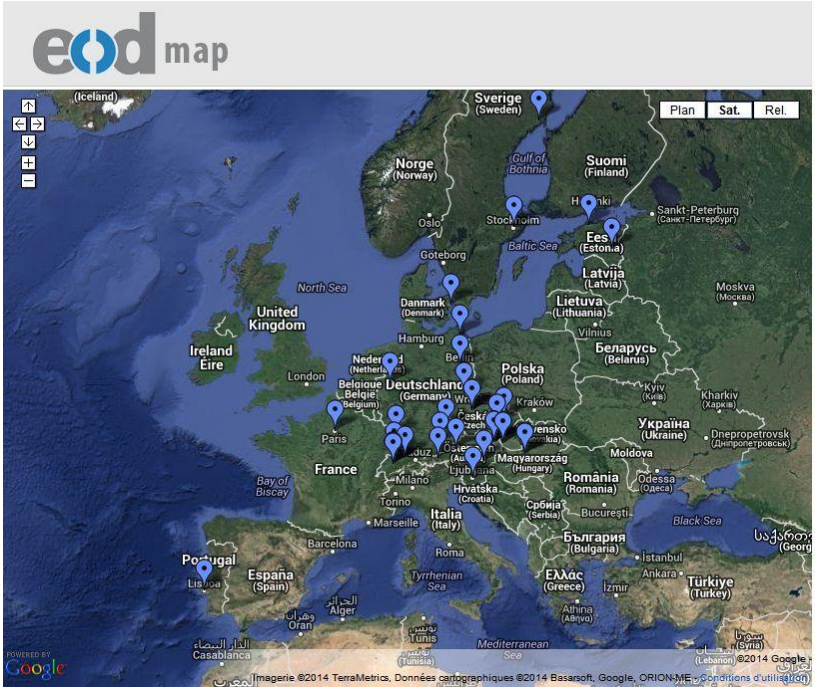


Figure 21. Localisation des membres du réseau Ebooks on Demand au 8 juillet 2014 d'après

https://www.facebook.com/eod.ebooks/app_402463363098062 (consulté le 23 juin 2016)

Au total, 3,5 millions de notices d'ouvrages sont proposées. Les membres du réseau paient une cotisation annuelle d'environ 1000 € à la bibliothèque chargée de la coordination. Ces frais recouvrent les frais réels d'administration, d'OCR, d'accès à la plateforme de paiement Order Data Management (ODM), sa maintenance, et son service d'assistance. Ces infrastructures étant mutualisées, leurs coûts sont ainsi partagés.

Entre 2007 et 2009, d'après (Gstrein, 2009), 3200 livres (840 000 pages) avaient été numérisés par 2000 usagers. Entre 2007 et 2011, d'après (Gstrein,

2011), près de 5000 livres avaient été numérisés pour environ 1 millions de pages scannées. Près de 2500 personnes auraient fait une commande sur cette période. Si on considère les trois bibliothèques ayant bénéficié des meilleures statistiques, chacune d'entre elles reçoit, en moyenne entre 250 et 350 numérisations de livres par an soit une par jour d'ouverture de la bibliothèque. En France, la Bibliothèque Inter-Universitaire de Médecine annonçait en 2009 avoir reçu 95 commandes pour 102 volumes et 18 932 pages numérisés, d'après une information informelle recueillie oralement.

Globalement, le nombre de commandes et les revenus générés par EOD sont croissants, comme l'illustre cet extrait d'un rapport d'activités :

eod eBooks 2008-2012

	2008 (6 months)	2009	2010	2011	2012 (projection)
> 1 orders	15 libraries	20 libraries	24 libraries	27 libraries	30 libraries
	€	€	€	€	
Revenue	25.107,01	51.582,93	63.607,28	78.512,30	?
Number of finished orders	700	1200	1480	1781	2200
Total enquiries			2550	4148	4700

Figure 22. Extrait d'un rapport d'activités EOD (d'après Klopp, 2014)

Dans la grande majorité des cas, la commande coûte 10 euros forfaitaires à l'usager pour les frais de gestion. A ces 10 euros s'ajoutent un coût à la page qui oscille entre 0,15 € et 0,30 €. Au final, le prix d'une numérisation de livre est généralement compris dans une fourchette qui va de 20 € à 49 €. Une minorité de livres numérisés (20%) coûtent au-delà de cette fourchette. Un document de 250 pages numérisé coûterait ainsi à l'internaute entre 30 € et 130 € (Mühlberger, 2009). Et en 2009, le coût moyen pour le lecteur serait plus précisément de 53 €.

Selon une enquête menée par EOD en 2009, ces tarifs sont considérés comme élevés ou très élevés par 30 % des répondants mais 95% restent satisfait du rapport qualité / prix. Mais, au delà de 50 €, les internautes deviendraient plus réticents à financer la numérisation d'un livre (Mühlberger, 2009).

A la Bibliothèque Nationale et Universitaire de Slovénie, par exemple, entre mai 2009 et 2011, les commandes semblent concerner principalement des tranches de prix relativement faibles :

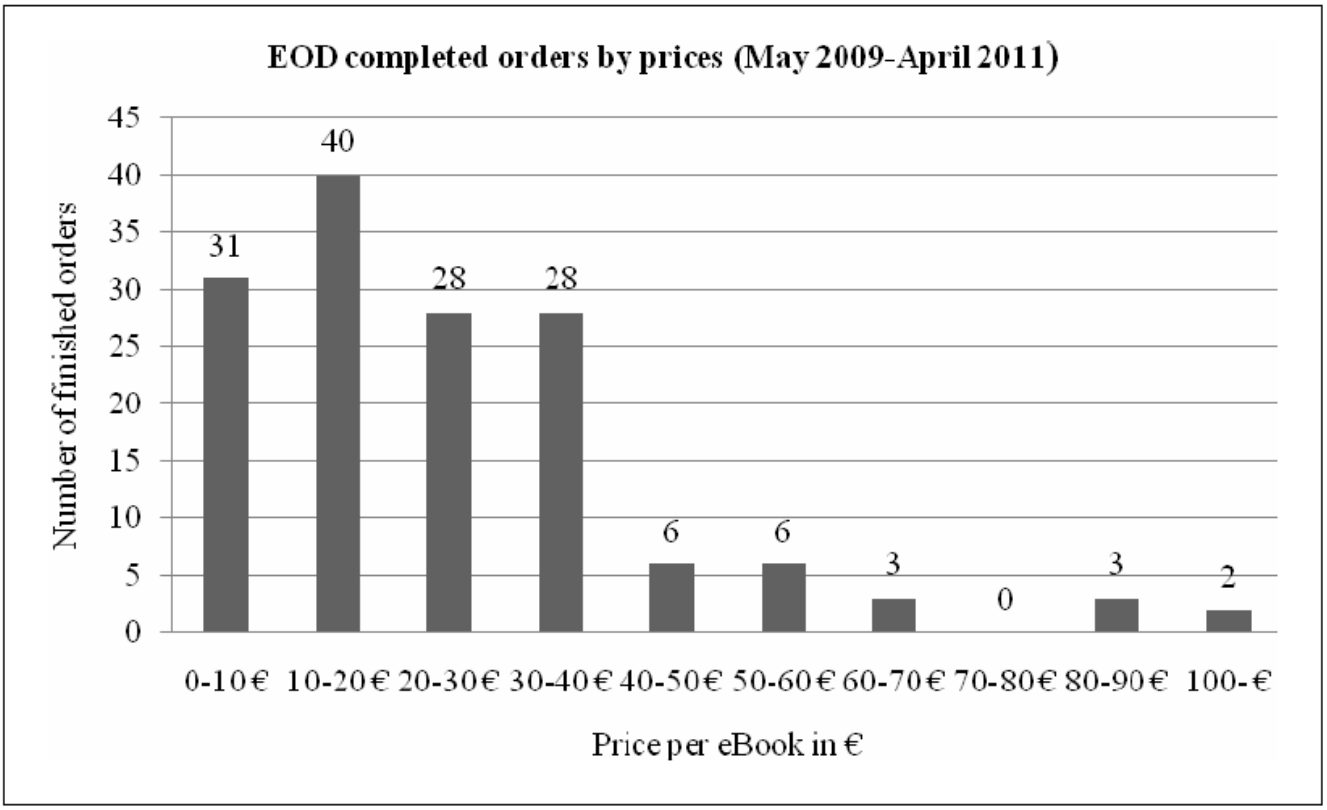


Figure 23. Commandes par tranches de prix pendant la période 2009-2001 à la Bibliothèque Nationale de Slovénie (d'après Brumen, 2012)

D'après nos calculs[21], le prix moyen pour la numérisation d'un livre de 200 pages (avec OCR) serait donc de 46,24 € en faisant la moyenne des tarifs des tarifs des 36 partenaires. L'enquête déjà mentionnée, nous apprend également

[21] D'après les tarifs EOD affiché sur http://books2ebooks.eu/fr/prices le 9 juillet 2014

que le PDF multicouches avec OCR est, avec la possession du livre original, la forme préférée des usagers, devant le PDF image, le texte océrisé en ligne, et la lecture de l'original imprimé en bibliothèque.

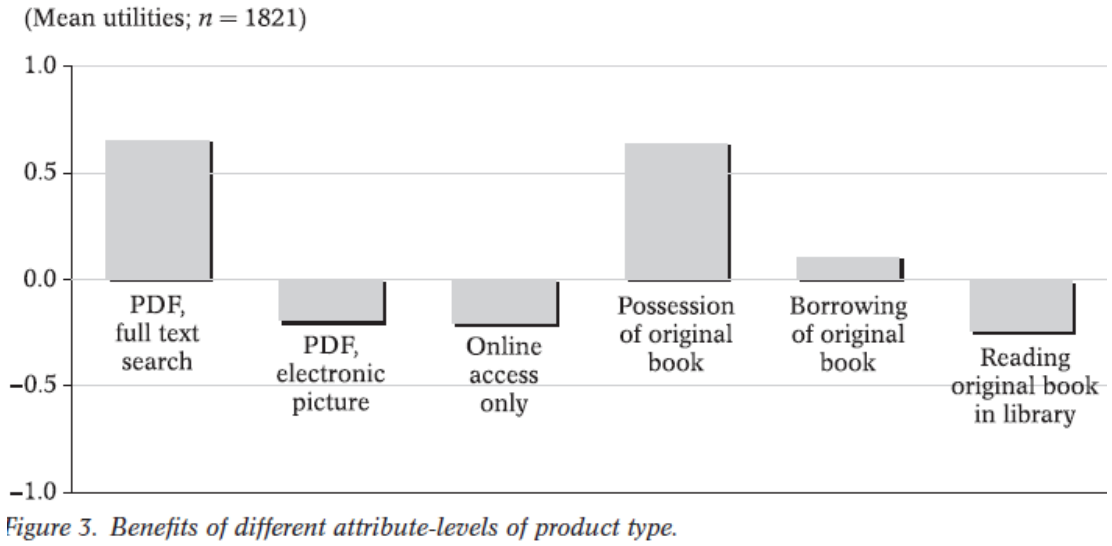


Figure 3. Benefits of different attribute-levels of product type.

Figure 24. La forme sous laquelle les usagers préfèrent consulter les documents
(d'après l'enquête rapportée par Mühlberger, 2009)

D'après l'enquête rapportée par (Mühlberger, 2009), un délai de livraison qui excéderait 3 semaines serait perçu très négativement par les usagers, comme l'illustre le diagramme suivant :

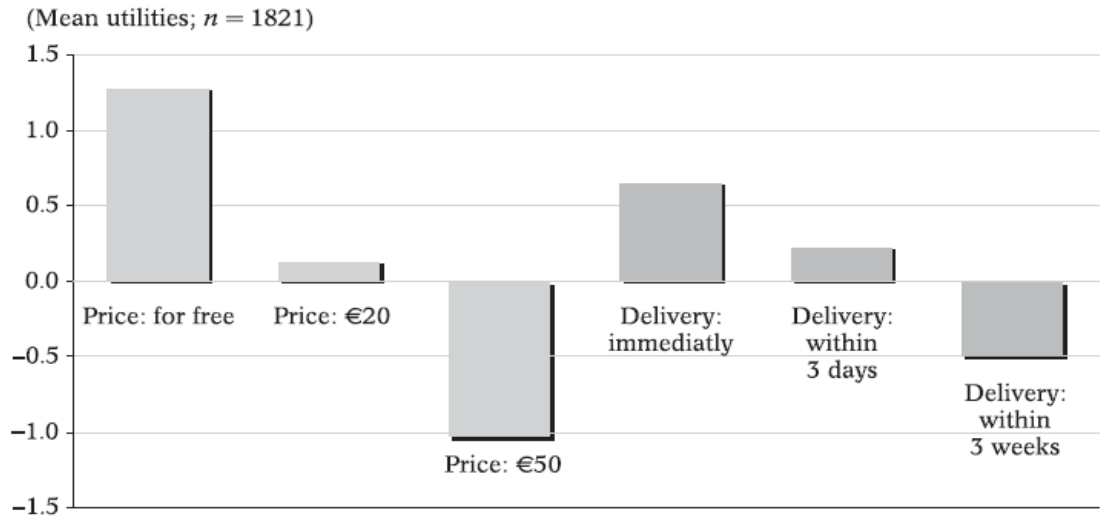


Figure 1. Benefits of different attribute levels of price and delivery time.

Figure 25. La perception positive / négative selon les prix et les délais (d'après l'enquête rapportée par Mühlberger, 2009)

A l'instar du projet Numalire, les usagers sont principalement motivés par des raisons d'ordre professionnel (plus de 60 %), mais 16 % d'entre eux sont des bibliophiles, des amateurs ou des collectionneurs :

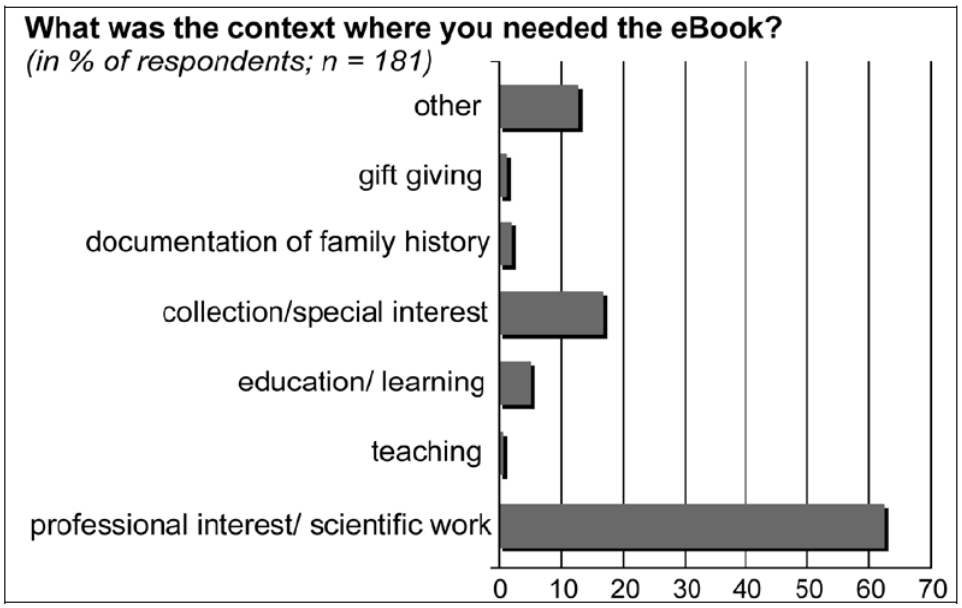


Figure 26. Centres d'intérêts des usagers d'après (Gstrein, 2011)

Les demandes seraient principalement justifiées par le fait que les usagers n'ont aucun autre moyen de se procurer de documents et par la difficulté d'accéder à certains livres anciens :

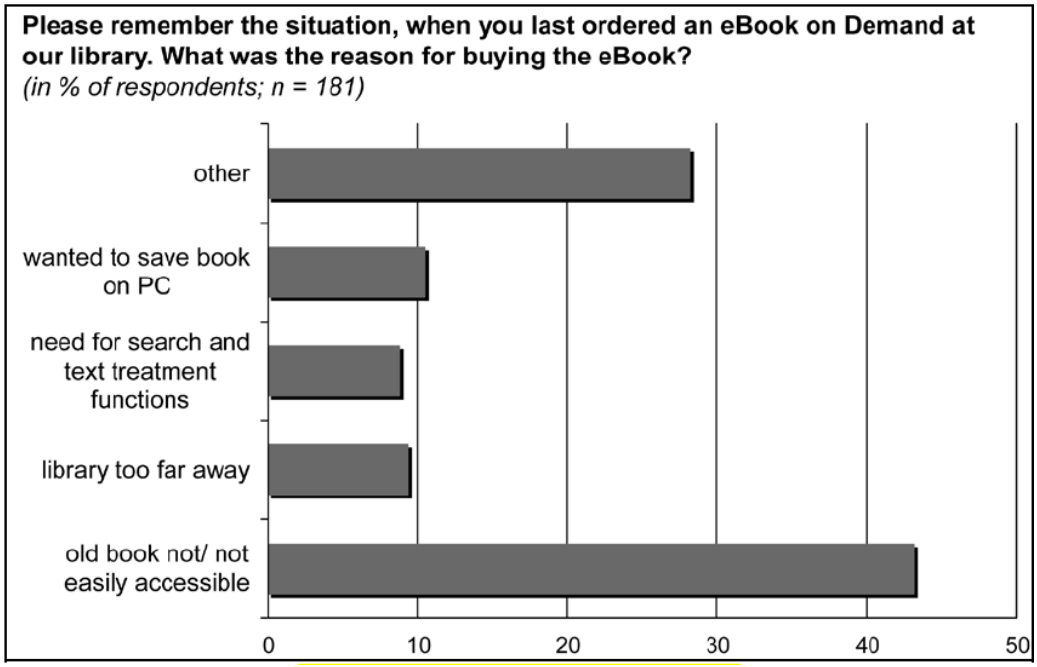


Figure 27. Raisons pour lesquelles les usagers ont commandé, d'après (Gstrein, 2011)

A l'instar des résultats obtenus avec l'expérimentation Numalire, ce sont majoritairement des hommes qui commandent. Les clients sont généralement originaires du même pays que la bibliothèque à qui ils commandent la numérisation. Ainsi, d'après les statistiques de la Bibliothèque Inter-Universitaire de Santé, la France reste majoritaire parmi les demandes :

Region	Number of customers	Number of orders
Europe	92	299
Own country	155	568
World wide	51	160
Total	**298**	**1,027**

Figure 28. Statistiques des commandes EOD de la Bibliothèque Inter-Universitaire de Santé d'après (Klopp, 2014)

Depuis 2009, et si l'institution a fait le choix de cette option, les livres numériques sont également envoyés vers Amazon Booksurge accompagnés de métadonnées et d'ISBN. Des reproductions en *Print on Demand* peuvent ainsi être commandées via Amazon Booksurge et les livres POD sont également accessibles directement à l'achat sur Amazon.

La numérisation à la demande de documents peut être proposée à des internautes à partir de plusieurs sources :

- Certains internautes, à la recherche d'un document, vont saisir son titre directement dans un moteur de recherche et arriver sur une plateforme bien référencée sur laquelle le financement de cette numérisation leur sera proposé.
- D'autres encore vont recevoir cette proposition à partir de boutons depuis les notices bibliographiques des catalogues en ligne de bibliothèques. Ces boutons pourraient ainsi proposer de "télécharger", de "numériser", de "réimprimer", et/ou d'"expédier" le document en fonction des services disponibles (Stambaugh, 2013).
- Enfin, certains contributeurs pourront être sollicités directement depuis les bibliothèques numériques. Celles-ci pourraient ainsi afficher non seulement les documents qu'elles ont numérisé, mais aussi ceux qu'elles souhaitent numériser et en proposer le financement aux internautes. Ce fonctionnement permettrait ainsi, en outre, aux bibliothèques, d'exposer leurs projets de numérisations futures. Or,

pour le moment, ces échanges d'informations se font par échanges de fichiers Excel ou, sans malheureusement rentrer dans le détail, à travers le site patrimoine numérique du Ministère de la Culture ou de NUMES pour celui de l'Enseignement Supérieur).

Ce type de service permettrait aux bibliothèques de proposer à leurs lecteurs des services de reprographie et de reproduction numérique, de qualité professionnelle, bien souvent inexistants, et de répondre ainsi aux demandes de reproduction de leurs usagers sans avoir à en supporter le coût. Il permettrait, de surcroît, de remplacer, de moderniser et de rendre plus efficaces, les services de Prêt entre Bibliothèques qui ont beaucoup vieillis. Il permettrait aussi aux bibliothèques de compléter leurs programmes de numérisation tout en partageant leur politique de sélection des documents à numériser avec le grand public et divers mécènes.

En effet, ce modèle économique pourrait également intéresser des institutions, des fondations et des mécènes, au delà des internautes à qui un service d'accès à des documents serait fourni. La mention « ce livre a été numérisé grâce au soutien de Madame X, de l'UMR CNRS Y, ou de la fondation Z », sur le modèle économique des publicités Google Adwords, permettrait d'encourager ce financement participatif. Un livre qui générerait, par exemple, 6000 visites offrira un trafic web dont la valeur pourrait être estimée au coût d'une numérisation et il pourrait être donc rentable pour un investisseur d'en financer la numérisation, d'autant que, à la différence d'une publicité de type Google Adwords, sa durée ne serait pas limitée et le prix à payer serait fixe. Une fois la numérisation financée, le nom de l'investisseur restera affiché et aucun coût supplémentaire ne pourra être demandé, quelque soit le nombre de visiteurs ou du nombre de clics générés. Pour certains documents, il est, en effet, probable, qu'au-delà d'un certain nombre de visiteurs, les coûts en numérisation soient compensés par un retour sur investissement, via la publicité visualisée et le trafic web généré par des liens. La numérisation du patrimoine pourrait ainsi, d'un certain point de vue, être considérée aussi comme un investissement, une action de communication et de marketing. En effet, une bibliothèque numérique permet de toucher, dans le monde entier, des internautes qui s'intéressent spécifiquement à tel ou tel sujet. Il suffirait

que les entreprises qui souhaitent toucher telle ou telle catégorie d'internautes financent la numérisation des livres qui les intéressent. En leur rendant ce service, ils amélioreraient leur image et pourraient aussi obtenir du trafic web via des liens pointant vers leurs sites.

En ouvrant les politiques documentaires et les politiques d'acquisitions de leurs bibliothèques numériques, les bibliothèques peuvent développer des collections numériques directement sélectionnées par leurs usagers. Tout en satisfaisant un besoin documentaire, l'internaute alimente ainsi la bibliothèque numérique qui devient une co-création. La politique documentaire et la politique d'acquisition de la bibliothèque numérique sont ainsi déterminées par ses usagers qui décident des titres qui rentreront dans la collection numérique, et non, par des professionnels qui risquent de ne pas maîtriser tous les domaines de la connaissance. Qui mieux que l'usager lui-même peut connaître les besoins de l'usager ? La bibliothèque numérique ainsi constituée au fil des années par la numérisation à la demande est donc le reflet des choix des internautes qui président ainsi à sa politique d'acquisition. Il s'agit d'une bibliothèque numérique enrichie par les internautes pour les internautes. Elle est l'œuvre des internautes eux-mêmes, dans une logique bottom-up radicalement différente de logiques d'offres déconnectées de la demande, du public et des usages. Derrière la notion de numérisation à la demande comme d'impression à la demande, figure la notion de déclenchement immédiat et en temps réel de l'offre par l'usager individuel auprès d'un fournisseur qui maintient un service potentiellement accessible. La numérisation à la demande est centrée sur l'usager et sur sa demande personnelle. Son modèle n'est pas "top-down", mais "bottom-up". Il répond particulièrement bien aux besoins d'étudiants et de professeur qui travaillent de plus en plus à distance et peuvent avoir besoin d'accéder de partout et à tout moment aux documents (Tafuri, 2011). La numérisation à la demande est aussi l'évolution d'un modèle économique du "just in case" ou "juste au cas où" (stocker pour anticiper les demandes) qui était celui des bibliothèques (constituer des collections pour anticiper les besoins des lecteurs) vers le modèle économique du "just in time" ou "juste à temps" (Reighart, 2014). Or, les bibliothèques tirent une bonne part de leur

expertise dans la constitution de collections et dans leurs politiques d'acquisitions, et pourraient donc se sentir directement remises en question par ce modèle économique qui ne pourra s'implanter sans une conduite du changement. Ce mouvement prolonge le mouvement qui a conduit les bibliothèques à adopter le modèle du "libre service", développé initialement dans les magasins, sous la forme de collections en "libre accès". Il s'agissait, en effet, déjà de sous-traiter le travail de magasinier auprès du lecteur lui-même et, d'une certaine manière, d'intégrer progressivement le consommateur dans le procès de production.

En s'inscrivant dans ce type de démarche sous la forme de numérisation à la demande, les bibliothèques externalisent, auprès du grand public, le laborieux travail de sélection des documents qui méritent encore d'être numérisés et le travail ingrat de vérification que les documents n'ont pas déjà été numérisés car il est peu probable qu'un internaute soit prêt à financer la numérisation d'un livre déjà numérisé pour un coût qui reste non négligeable. Ce travail d'identification peut difficilement être automatisé et, comme le signale (Pignal, 2013), « en coût complet, la sélection peut être plus onéreuse que la numérisation d'un fonds entier ».

Grâce à la numérisation à la demande par *crowdfunding*, l'argent public pourrait se concentrer sur les documents non susceptibles d'intéresser le privé et qui présentent un intérêt patrimonial, scientifique ou historique. L'argent public pourrait ainsi être mieux utilisé et laisser l'argent privé prendre en charge la numérisation de livres qui intéressent des particuliers ou, s'ils sont susceptibles de générer du trafic web, d'intéresser des investisseurs ou des mécènes. La numérisation à la demande pourrait présenter un modèle alternatif à celui de numérisation de masse proposé par Google Books (Chamberlain, 2012). La numérisation de masse avec de l'argent public et la numérisation individuelle à la demande avec de l'argent privé du *crowdfunding* pourraient ainsi se compléter de manière harmonieuse.

De la même manière, la numérisation à la demande pourrait permettre de trouver une harmonie entre différents acteurs qui y trouveraient chacun leurs intérêts :

- Les particuliers qui ont la possibilité d'accéder à des documents difficilement accessibles et de disposer de services de reproduction numérique
- Les bibliothèques qui peuvent compléter leurs bibliothèques numériques et offrir un nouveau service sans en supporter le coût
- Les mécènes qui peuvent valoriser leurs noms en finançant des livres sur des thématiques proches de leurs préoccupations
- Des investisseurs qui pourraient investir dans la numérisation d'un livre en espérant qu'il génère un trafic suffisant pour que leur société ou leur site web en bénéficie.

Ce type de service semble répondre à un besoin réel. Une étude sur la faisabilité d'un service de numérisation à la demande va largement dans ce sens (Chamberlain, 2010). Parmi 61 universitaires et 16 bibliothécaires : 91,8% des universitaires de Cambridge sondés, seraient intéressés par un tel service. Et 65,5 % d'entre eux seraient également intéressés par le développement d'un service d'impression à la demande. Selon les répondants, le coût adéquat d'une impression à la demande serait de 10 à 15 livres (soit 11,81 € à 17, 71 €) pour 42, 9 % d'entre eux et de 15 à 25 livres (soit 17, 71 € à 29,53 €) pour 33,3 % d'entre eux, le coût adéquat de numérisation à la demande devrait être, quant à lui, de 10 à 15 livres (soit 11,81 € à 17, 71 €) pour 66,7 % des répondants et de 15 à 25 livres (soit 17, 71 € à 29,53 €) pour 35 % d'entre eux, alors que le coût réel, selon (Chamberlain, 2010) serait plutôt de 40 livres (soit 47,25 €). 44 % des répondants seraient prêts à attendre la livraison du service après une semaine ou plus de délais et seuls 10 % estiment que ce temps devrait être de 24 heures seulement.

D’après (Chamberlain, 2010), les tarifs pratiqués par les bibliothèques sont les suivants :

Bibliothèque	Numérisation à la demande	Impression à la demande
Université de l’Utah	0,05 $ (ou 0,04 €) par page 20 $ (ou 14,77 €) pour un livre de 400 pages	0,05 $ (ou 0,04 €) par page 20 $ (ou 14,77 €) pour un livre de 400 pages
McGill Libraries (Canada)	10 $ (ou 6,94 €) forfaitaires le PDF (numérisation avec Kirtas)	29 $ (ou 20,14 €) forfaitaires le livre imprimé avec Espresso Book Machine
National Library of Australia		13,20 $A (ou 8,9 €) par tranche de 50 pages 52,8 $A (ou 35,6 €) pour un livre de 400 pages
National Archives	3 £ 50 (ou 4,23 €) pour la plupart des documents	
Cambridge University Library	Après 1900 et pour 400 pages : 265 £ (ou 320,58 €) (scan) Avant 1900 et pour 400 pages : 1 298, 50 £ (ou 1570,87 €)	Après 1900 et pour 400 pages : 265 £ (ou 320,58 €) (photocopie)

Tableau 10. Tarifs pratiqués par diverses institutions pratiquant la numérisation et l'impression à la demande

Le coût d’une numérisation à la demande, donc à l’unité reste bien supérieur à celui dont sont plus coutumières les bibliothèques. Contrairement à la numérisation de corpus plus importants ou à la numérisation dite de masse, avec la numérisation à la demande, il est impossible de classer, par trains, les documents en fonction de leurs caractéristiques physiques et il est donc nécessaire de choisir quel matériel utiliser, de paramétrer les scanners et de les

régler à chaque document à numériser. Ce temps de paramétrage ne peut être lissé sur la numérisation de plusieurs livres et il devient plus difficile de rentabiliser l'utilisation de tel ou tel type de machine, ce qui engendre un coût à l'unité bien plus important.

La mise en place de services de numérisation à la demande pourrait se faire via des délégations de services publics. Ainsi, sans avoir à passer par les bibliothèques et les administrations de l'État et générer des coûts de gestions pour elles, ce pourrait être directement le particulier ou le mécène qui pourrait ainsi commander puis payer le délégataire.

Mais, la mise en place de ce type de service pourrait se concrétiser trop tardivement. En effet, la numérisation pourrait devenir de l'histoire ancienne. Google Books a dépassé son objectif initial et a dépassé le seuil des 30 millions de livres numérisés. D'autres projets ont également beaucoup numérisés et plusieurs milliers de livres sont numérisés chaque jour. Leonid Taycher, un ingénieur travaillant pour Google a estimé, en 2010, à près de 130 millions le nombre total d'imprimés produits dans le monde depuis Gutenberg. D'après (Mühleberger, 2009), environ 1 millions de livres ont été publiés en Europe entre 1500 et 1800, 5 millions entre 1800 et 1900. Les bibliothèques peinent donc désormais à identifier des livres qui n'ont pas déjà été numérisés. A la Bibliothèque Sainte-Geneviève, par exemple, en partant d'une extraction du SUDOC contenant 16 000 notices de documents exemplarisés à la seule Bibliothèque Sainte-Geneviève, il n'en est resté que 400 après élimination des doublons, des tirés à part, de certaines monographies en plusieurs volumes décrites tantôt sur plusieurs, tantôt sur une seule notice, des documents déjà numérisés, des documents sans aucun intérêt… De son coté, Google a déjà ralentit le rythme de sa numérisation. Bien que l'évolution juridique du code de la propriété intellectuelle et l'émergence de projets comme le Registre des Livres Indisponibles en Réédition Electronique (ReLIRE) pourraient permettre de numériser des œuvres épuisées et des œuvres orphelines, le marché de la numérisation semble atteindre progressivement ses limites et se rétrécir.

Dans ces conditions, il devient de plus en plus difficile pour les chefs de projets de numérisation d'identifier des documents qui méritent encore d'être numérisés et le marché pour un service de numérisation à la demande par *crowdfunding* pourrait également se restreindre. Mais, ces conditions nouvelles pourraient également et au contraire être perçues comme une possibilité pour rendre encore plus pertinent ce type de service. Si la numérisation de masse a atteint ses limites et ne peut plus être maintenue, seule une numérisation "en dentelles", de "niches", à l'unité et donc à la demande pourrait avoir un intérêt, d'autant qu'elle permettrait de numériser des documents qui ont échappés au dépôt légal, aux logiques de collections, aux grands programmes de numérisation ou des documents en langues rares ou encore sur des sujets très spécifiques et, ce faisant, permettre de mieux satisfaire les besoins des usagers.

Les plateformes générales de *crowdfunding* déjà existantes pourraient également être utilisées par les bibliothèques afin de faire financer par des internautes ou des mécènes la numérisation de leurs livres sans avoir à développer des plateformes spécifiques.

Le modèle Gold Open Access, utilisé par le projet RevealDigital, pourrait également être utilisé afin d'obtenir un retour d'investissement sur les numérisations déjà effectuées. Les documents numérisés ne seraient qu'accessibles aux abonnés, par pay per view ou dans les bibliothèques mais pourraient être "libérés" par souscriptions *crowdfunding* et mécénats à la demande. Mais rien ne garantit qu'un tel modèle puisse encore souffrir la concurrence de bibliothèques numériques gigantesques et dont le contenu est accessible gratuitement.

2.3- L’impression à la demande (*Print on Demand*, POD) : l’Espresso Book Machine

Numérisation à la demande et impression à la demande obéissent à des logiques très voisines. Bien que l’impression à la demande ne soit pas, à proprement parler, et contrairement à la numérisation à la demande, une forme de *crowdsourcing*, il est impossible d’évoquer la numérisation à la demande séparément de l’impression à la demande d’abord parce que les services de numérisation à la demande proposent généralement aussi de l’impression à la demande, ensuite parce que, historiquement, l’impression à la demande a parfois précédé la numérisation à la demande, et enfin car le modèle économique du « à la demande » est le même. Ainsi, au lieu de convertir, par la numérisation, un imprimé sur support papier en document électronique suite à la demande d’un usager, on va, au contraire, et toujours à la demande d'un usager, rétroconvertir par son impression, un document électronique en nouveau document sur support papier et ainsi, « ressusciter » l'imprimé d'origine.

Depuis plusieurs années, on a constaté, que le nombre de tirages des livres avait tendance à diminuer dans le secteur de l'édition. Pour s'adapter à cette situation, est apparu, depuis 2002, le modèle du *Print on Demand* (POD). Il s'agit d'imprimer les livres en flux tendu sur des imprimantes jets d'encres plutôt que sur des machines offset comme c'était le cas auparavant, en quasi temps réel, et selon la demande des consommateurs. Ainsi, ces derniers influent directement sur la production. Le *Print on Demand* permet ainsi :

- de ne plus surproduire et avoir de trop nombreux invendus qui représentent une perte sèche pour les entreprises ;
- de ne plus avoir à administrer, à gérer et à conserver des stocks qui peuvent être coûteux en hébergement et en magasiniers ;
- de restreindre les coûts liés à la logistique de la chaîne du livre et en particulier pour ce qui concerne les transports ;
- de ne plus avoir à anticiper et à prévoir à l’avance de nombre d’exemplaires qui pourraient être vendus et de pouvoir ainsi prendre davantage de risques ;

- de produire des tirages au plus près des besoins ;
- de permettre la publication d'ouvrages destinés à des communautés très restreintes, des livres très spécialisés avec de plus petits tirages ;
- de dépasser le problème des ouvrages épuisés (Blummer, 2006) ;
- dans des sociétés devenues multiculturelles, de mieux satisfaire les besoins de populations lisant des langues diverses ;

Le *Print on Demand* est l'application, dans les domaines de l'imprimerie et de l'édition, du modèle économique "juste à temps". Traditionnellement, les éditeurs, comme les bibliothèques, fondaient leurs modes de fonctionnements sur un modèle très différent, celui du "juste au cas où" qui consistait à produire et à stocker pour anticiper la consommation et les demandes ou encore, à acheter des livres au cas où un lecteur en ait un jour besoin.

Comme nous l'avons succinctement évoqué dans notre chronologie du crowdsourcing, le modèle économique dit du "juste à temps" aurait été inventé, dès les années 1950, au Japon et développé, en particulier, dans l'entreprise Toyota. Au Japon, l'espace disponible dans les commerces et les échoppes étant très limité, à cause de contraintes liées à la géographie et à l'urbanisme nippons, il n'était pas possible de disposer d'un stock important d'exemplaires de la même marchandise et il était donc nécessaire de trouver des moyens de remplacer rapidement les marchandises vendues sans, pour autant, pouvoir disposer de stocks. Ensuite, ce modèle a largement été développé et conceptualisé par le toyotisme. Il s'agissait alors surtout de diminuer les coûts, d'éviter les invendus et les stocks de marchandises susceptibles de perdre graduellement de la valeur. Avec ce modèle, l'offre est plus directement déterminée par la demande, la production est poussée, en flux tendus, par la consommation.

En 2007, la production de livres imprimés, via des machines Offset, ne se serait accrue que de 1 % alors que celle d'imprimés à la demande aurait été multipliée par 6 entre 2006 et 2007. Sur la période 2002-2007, le nombre de titres produits sur Offset aurait augmenté de 29 % quand il aurait augmenté de 313 % pour le *print on demand* (Dougherty, 2009). Un an plus tard, en 2008, d'après

Bowker, la production de livres imprimés aux USA avec le modèle traditionnel aurait connu une croissance de 3% tandis que la production sous la forme de d'impression à la demande a augmenté de 132 %. Aux États-Unis, le nombre de livres imprimés grâce au *print on demand* serait désormais supérieur aux autres, grâce à l'autoédition, en particulier.

Le coût de production avec des imprimantes de type jet d'encre qui sont utilisées pour le modèle *print on demand* demeure supérieur par rapport au modèle traditionnel qui a recours à des machines de type Offset. Ainsi, le prix d'un livre serait 20 % à 30 % plus cher, d'après Luc Spooren, que nous avons personnellement rencontré et dont les propos ont été rapportés par le numéro du 9 juin 2009 du Nouvel Observateur. Malgré cet inconvénient, 30 000 exemplaires pourraient être obtenus en seulement 2 jours sur un mode *print on demand* quand la même quantité d'imprimé nécessiterait 2 semaines pour être produits sur un mode traditionnel (Dougherty, 2009).

Après avoir rencontré un vif intérêt auprès du Government Printing Office, de Internet Archive et de Google Books, ce modèle économique devait nécessairement rencontrer également le monde de la numérisation des bibliothèques. Grâce à l'impression à la demande, les textes tombés dans le domaine public et qui ont été numérisés vont pouvoir « ressusciter » sur support imprimé, sous la forme de fac-similés. Mais ce mode de fonctionnement « à la demande » est très différent du mode de fonctionnement traditionnel des bibliothèques obéit au modèle du "juste au cas où". Elles achètent des livres et constituent des collections en anticipation par rapport à la demande. Mais, avec l'impression à la demande, comme avec la numérisation à la demande, c'est un modèle diamétralement opposé qui s'applique, celui du "juste à temps". Son application remettrait donc directement en cause la politique d'acquisition des bibliothèques par les professionnels.

Ainsi, (Lewis, 2010) va même jusqu'à imaginer une bibliothèque classique qui achète actuellement 10 000 livres par an. Chaque titre lui coûte 35 $ en moyenne à l'achat, 25 $ pour sa commande et son catalogage et 40 $ pour son équipement, son stockage et sa circulation soit 100 $ par livre, soit 1000 000 $ par

an pour 50 000 consultations ou emprunts par an, l'utilisation du fonds ne portant que sur une mineure partie des documents acquis. Alors que, à l'instar de la plupart des bibliothèques, son lectorat diminue dramatiquement, plutôt que de maintenir l'essentiel de ses moyens financiers et humains à acheter, à cataloguer, à équiper et à conserver des livres de moins en moins consultés et dont seule une minorité sera un jour consultée, la bibliothèque en question ne pourrait-elle pas plutôt produire, via une Espresso Book Machine, un livre que si un usager en a besoin et libérer la masse salariale ainsi libérée pour se consacrer à de nouvelles missions plus innovantes, plus utiles et plus enrichissantes (archives ouvertes, bibliométrie, veille, formation, numérisation, *text mining*..) ? Dans cette hypothèse, elle aurait à dépenser 60 000 $ par an pour louer une Espresso Book Machine, 40 000 $ par an pour payer un opérateur (sauf si elle parvient à reconvertir un des ses employés) et la production de chaque livre imprimé lui reviendrait à 3 $ tandis que les droits de l'éditeur s'élèvent à 15 $ par livre imprimé en moyenne. Dans ces conditions, à budget annuel identique de 1000 000 $, la bibliothèque pourrait produire près de 40 000 livres par an (Lewis, 2010), soit 4 fois plus qu'avant. Elle pourrait faire constituer sa collection par ses usagers et être sûre que chacun des livres qu'elle conserve a été consulté au moins une fois. Le temps d'attente pour l'usager avant d'avoir accès au livre n'excéderait pas 5 minutes.

Conçue en 2006, vendue par la société OnDemandBooks et, dans un premier temps, diffusée par Xerox, l'Espresso Book Machine résulte de l'intégration d'un copieur, d'une imprimante, et d'un massicoteur-plieur et brocheur dans une même machine. Ce photocopieur avec son module capable d'ajouter une couverture brochée au livre permettrait d'imprimer des livres du format 11,4 x 12,7 cm au format 21 x 27,3 cm.

Installée dans des librairies, des bibliothèques, des gares, des aéroports ou des lieux à forte fréquentation, l'Espresso Book Machine permet d'acheter in situ, une version imprimée des livres électroniques disponibles sur EspressNet qui propose plus de 8 millions de titres dont une partie de Google Books (1 million de livres), archive.org (plus de 2 millions de livres), HathiTrust, Lightning Source et Gallica. La Bibliothèque nationale de France elle-même propose plusieurs milliers

de livres de son catalogue Gallica en *Print on Demand* sur l'Espresso Book Machine. Cette machine permet aussi à ses usagers d'imprimer leurs propres productions. Ce type de demande est d'ailleurs bien souvent majoritaire d'après les retours d'expériences. Les bibliothèques peuvent également intégrer leurs bibliothèques numériques au sein du catalogue de l'Espresso Book Machine.

Figure 29. Photographie d'une Espresso Book Machine (d'après http://ondemandbooks.com)

Pour un livre de 300 pages, la durée serait de 5 minutes d'après (Anderson, 2010), mais pour les documents les plus complexes, cette durée peut aller jusqu'à 20 minutes. D'après (Dougherty, 2009) le coût serait d'environ 10 $ (soit 7,36 €). Selon (Geitgey, 2011), ce prix varierait plus précisément de 6 $ pour un livre de 150 pages maximum à 10 $ pour un livre de 151 à 450 pages. D'après (Chamberlain, 2010), un livre de 400 pages coûterait 8 livres (9, 44 €) et il faudrait en imprimer plus de 1000 par an pour que l'opération soit rentable. D'après (Wilson-Higgins, 2011), ce coût serait de 0,01 $ par page et, pour l'Université du Michigan, le prix pour un exemplaire avec couverture rigide serait en moyenne de 39,95 $ avec un frais de port et de manutention de 7 $ pour les USA et de 15 $ pour le reste du monde. La librairie Blackwell de Londres, quant à elle, propose une autoédition à 35 livres pour le premier exemplaire. Les livres supplémentaires coûtent 5p par page avec un minimum de 5 livres par livre. Pour un livre de 300 pages, le coût serait donc de 15 livres, à condition qu'un premier livre de test ait déjà été créé. Ces coûts sont relativement faibles si on les compare avec ceux des

services de prêts entre bibliothèques qui seraient proches de 30 $. Par contre, l'achat d'une de ces machines est assez coûteux et nécessite une maintenance technique (bourrages éventuels, papier, encre, colle, cartons de couvertures…). Une location, une franchise ou une délégation de service public pourraient être les cadres adéquats. D'après (Wilson-Higgins, 2011), le coût d'installation serait d'environ 92 000 $ (soit près de 68 000 €). Le même auteur indique qu'il faudrait imprimer environ 60 000 livres chaque année pour que le coût à l'exemplaire soit suffisant. Ce prix pourrait néanmoins diminuer. Ainsi, l'Université de Toronto a annoncé avoir acheté une presse Asquith pour moins de 46 000 euros. Dans les bibliothèques, la première Espresso Book Machine a été achetée par la bibliothèque publique de New York le 21 juin 2007. En juillet 2012, des machines ont été installées dans de multiples bibliothèques et librairies aux USA (27), au Canada (12), en Angleterre (2), en Australie (2)...

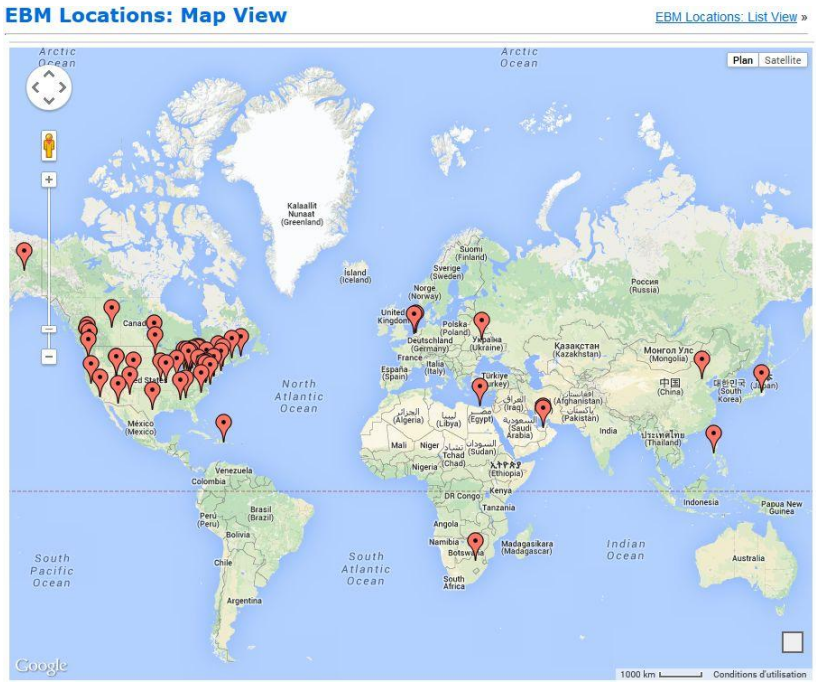


Figure 30. Répartition des EBM dans le monde d'après http://www.ondemandbooks.com/ebm_locations.php **le 9 juillet 2014**

En France, aucune EBM n'était installée jusqu'à la récente acquisition par les Preses Universitaires de France (PUF) en mars 2016. Une Espresso Book

Machine aurait pu être installée gratuitement à la Bibliothèque Sainte-Geneviève, en raison de sa proximité géographique avec la plupart des éditeurs parisiens, dont le marché est à conquérir pour la société, et dans le cadre d'une délégation de services publics et sous la forme d'une concession de 2 ans pendant lesquelles, la Bibliothèque aurait même pu bénéficier d'une marge sur les ventes. La BnF n'a, pour sa part, pas souhaité en installer une. Cette machine pourrait néanmoins offrir des services nouveaux aux usagers des bibliothèques tout en étant une source de revenus, en particulier dans le cadre d'un partenariat avec une société privée comme une librairie, par exemple.

Parmi les autres usages de l'Espresso Book Machine, on trouve aussi :

- l'auto-publication (de livres, de congrès, de mémoires ou de thèses) ;
- la possibilité de compléter les collections des institutions (remplacer des exemplaires disparus ou multiplier les exemplaires des documents très demandés) ;
- le prêt entre bibliothèques (Geitgey, 2011).

Le prêt entre bibliothèques d'un livre reviendrait à 30 $ quand l'impression à la demande d'un livre via une Espresso Book Machine reviendrait en moyenne à seulement 10 $, d'après une étude de l'Université Virginia Tech rapportée par (Dougherty, 2009)

Mais, en s'engageant dans cette voie, les bibliothèques cessent d'être de simples bibliothèques pour devenir aussi des libraires et des centres d'autoédition. Le succès de l'opération dépendra surtout du public qu'elles parviennent déjà à drainer.

Ces éléments relatifs à l'impression à la demande ont fait l'objet d'un article (Andro, 2015, 3)

Les documents numérisés par les bibliothèques pourraient donc être vendus, sous la forme d'impressions à la demande, par correspondance, par des sociétés, ou sur place via une Espresso Book Machine, dans le cadre de délégations de service public. Les bibliothèques confieraient ainsi cette nouvelle

mission de service public à un délégataire rémunéré par l'exploitation de ce service sans avoir à en assurer la gestion. Ainsi, tout en offrant de nouveaux services à leurs usagers sans avoir à en supporter le coût, les bibliothèques pourraient même bénéficier d'un retour sur investissement et participer à créer de l'activité économique. Elles pourraient également accroître la visibilité de leurs bibliothèques numériques en disséminant leurs contenus sur des sites de librairies en ligne, mais aussi compléter leurs collections d'imprimés grâce au *print on demand*, multiplier le nombre d'exemplaires de certains titres beaucoup demandés, remplacer les exemplaires disparus, offrir des service d'auto-publication de mémoires, de thèses ou de tout autres types d'écrits, éditer des livres d'événements pour les institutions publiques ou privés et moderniser leurs services de prêt entre bibliothèque qui pourraient parallèlement devenir moins coûteux pour les usagers.

Malgré cela, ce type de projet peine à se développer en France. Comme le signale très justement (Klopp, 2014), le *Print on Demand* n'intéresse guère les bibliothèques en France qui ne l'évoquent presque jamais dans leur littérature professionnelle alors qu'il s'agit d'un sujet qui fait couler beaucoup d'encre à l'étranger, en particulier, outre atlantique. Le partenariat avec des sociétés privées cherchant à faire du bénéfice est probablement perçu de manière péjorative par les professionnels des bibliothèques qui préfèrent peut être un financement de leurs projets et de leurs services exclusivement avec de l'argent public.

Le possible développement de ce type de service en bibliothèque en modifie aussi la définition et le périmètre. La frontière entre bibliothèque et librairie pourrait être remise en cause avec ce modèle économique. Comme le rapporte (Arlitsch, 2011), une étude de Allen Kent de 1979 (Use of Library Materials: The University of Pittsburgh Study. M. Dekker, New York.), un livre acheté par une bibliothèque universitaire a moins d'une chance sur deux d'être consulté un jour. Comme le suggère cet auteur, c'est vraisemblablement toujours le cas. Le *Print on Demand* pourrait remettre en question le traditionnel modèle "use it or lose it" sur lequel est fondée la politique d'acquisition des bibliothèques et qui repose sur

l’anticipation des besoins de leurs lecteurs en lui substituant un modèle à la demande plus centré sur l’usager.

Ainsi, de la même manière que la numérisation à la demande permettrait de créer des bibliothèques numériques dont la politique documentaire et la constitution de collections numériques seraient l’œuvres des internautes eux-mêmes, avec l’impression à la demande, pourraient être édifiées des bibliothèques physiques de documents imprimés à la demande de lecteurs, des collections qui seraient directement l’œuvre des usagers.

2.4- La correction participative de l'OCR et la transcription participative de manuscrits

La numérisation de la page d'un livre va générer une simple photographie de cette page. A partir de cette simple image numérique, il est impossible de rechercher (« en texte intégral ») un mot en son sein et de faire indexer son contenu par des moteurs de recherche. Il est également impossible d'en copier-coller un paragraphe, de générer des fichiers EPUB pour être lus sur des tablettes et des liseuses de livres électroniques. Pour rendre possibles ces usages, il va être nécessaire d'océriser l'image du texte, c'est-à-dire, de la soumettre à un traitement de reconnaissance optique de caractères (OCR) avec l'aide d'un logiciel dédié. Ce logiciel va déterminer les zones de textes, les colonnes, les tableaux, les images (segmentation), puis chercher à identifier à quel caractère correspond l'image de tel caractère. A la fin du traitement, le logiciel aura produit un fichier texte à partir du fichier image, en identifiant chacun de ses caractères, comme si on s'était chargé de le saisir sur clavier.

Malheureusement, ce type de traitement de reconnaissance de caractères génère encore parfois de nombreuses erreurs. La qualité de l'OCR dépendra de la qualité de la numérisation comme de la qualité du texte imprimé. Au niveau du document d'origine, les éléments suivants peuvent, par exemple, être sources d'erreurs :

- Sur le support papier : trou, décoloration, tâche, pli, déformation, disparité…
- Annotations manuscrites
- Typographies : irrégulières (incunables, par exemple), mal imprimées, typographies anciennes, oubliées, ou très originales...

Voici quelques exemples d'erreurs d'interprétations courantes :

Caractère sur l'imprimé	Erreur courante d'interprétation du logiciel OCR
H	li
M	in

Museum	inuseuim
Théologie	tliéologie

Et voici un exemple de texte d'OCR brut :

Avec quel plaisir nous eussions lu vos réeits et écouté les vieilles légendes du bon veux temps, que vous eussiez su nous raconter si bien! C'est con amer, que uous eussions feuilleté l'album dans lequel on retrouverait, l'antique cité romaine, avec ses mors à triples bandeaux de briques, que nos pères regardaient comme trois cercles d'or et qu'ils ont mis dans les armoiries de la ville. Qui de nous n'eût vu avec pla,sir le Chalon du moyen âge, avec les tours et les flèches de ses nombreuses eghses paroissiales et conventuellas, avec ses pignons sur rue, ses vieilles boutiques avec leurs auvents saillants, sons lesquels nos mères se plaisaient à jaser. N'an' nons-nons pas été heureux aussi de connaître tout ce que la Renaissance, à son tour, ava.t élevé dans notre ville, - car elle y avait aussi prodigué ses œuvres,- et ne regre,tera.t.o„ pas toujours, entre autres, ces tombes de marbre et de bronze qu'elle ava,t engees ans plusieurs de nos chapelles , Cette histoire est encore à faire. Le Père Berthand a bien composé son indigeste OrUniaU; Saint-Julien de Balleure a la.sse nne meilleure histoire; Pierre Naturel a écrit celle de nos évéqnes, demeurée s. longtemps enfouie sous la poussière de la bibliothèque de Lyon, oh j'ai eu la Donne chance de retrouver le manuscrit de la main d'Enoch Virey, que notre docte ami Henn Batault va publier. Le P. Perry a été aussi un excellent annaliste, en pmsant aux vra.es sources. Courtépée a écrit également d'excellentes pages mais A n a être que succinct. Ni les uns ni les antres n'ont eu, comme vouf, ava - âge de savo.r marner le pinceau, le crayon et le burin. ,1 manque dans leurs livre, les vues et les plans des lieux et des monuments dont ils ont parlé. C'est donc" vous, q„. savez être historien et artiste, à nous donner bientôt une histoire complète on sommaire, de ce passé déjà bien loin de nous

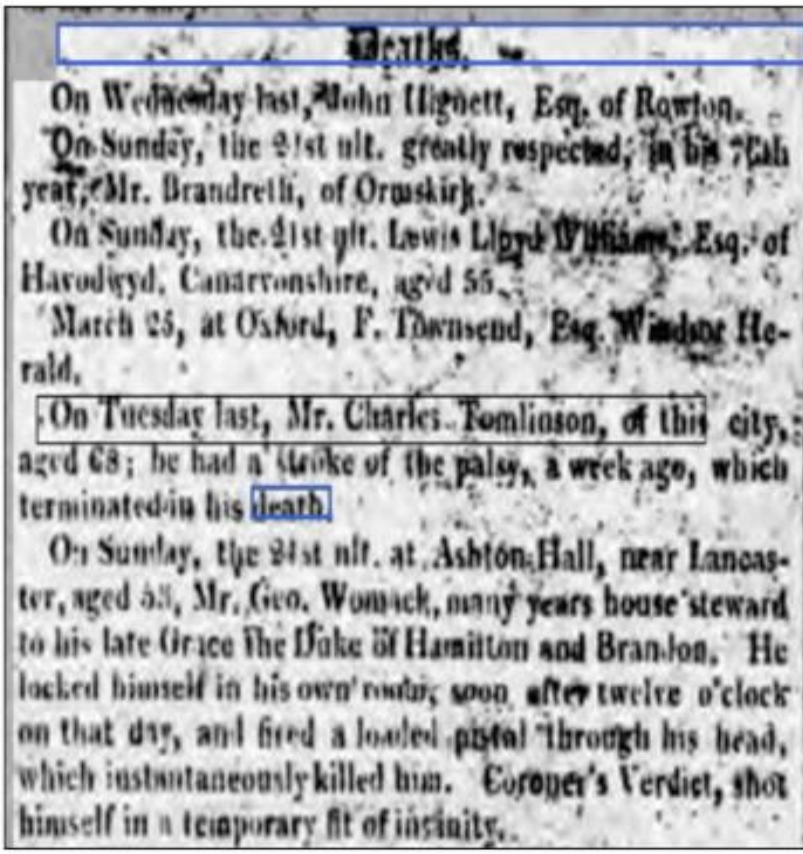
Figure 19. Capture d'écran d'un texte en OCR brute

En fonction de la qualité du document original et de la performance de numérisation, le texte obtenu comportera plus ou moins d'erreurs d'interprétations. Voici, par exemple, le résultat d'une reconnaissance optique de caractères sur un original d'encore plus mauvaise qualité issue de la bibliothèque numérique des journaux australiens TROVE :

raw OCR text

Deaths. **lln»rieff**, Esq. of <c .. Qn. Sunday, the till. greatly **Drandrellt**, of **Orms4\irJi**.- ~ ; ;√ ' • * On **ijfr r inn ljjjil F iij '11 f Havodivyd**, **Carnarvonshire**, S ; ** *- ' « ' March **Oxford**, **F. Tfovmeud**, **Uerald.** » • V . •On Tnesdav last, Mr. **Charles. IWilinson**, this 8 ; had vf thesis#„ a week ago, which terminate<i'iu his death. . / ' ▪ O'i Sunday, dJst nit. at. **AsbtCnvHall**, mar **Lancaster**, Mr.,**Geo. Worn ick**, many years house'steward hit late Once The **Hamilton** and **Brandon**. He locked himself h»oWn'r«wte<: soon. twelve o'clock" that dny, and fii»-d a loaded pistol "through Ins bead, 1 which instantaneously killed him. Coronet's Verdict, shot himself in a temporary fit of Friday week,

newspaper image



Deaths.

On Wednesday last, John Liggett, Esq. of Rowton.

On Sunday, the 21st ult. greatly respected, in his 74th year, Mr. Brandreth, of Orrnskirk.

On Sunday, the 21st ult. Lewis Lloyd Williams, Esq. of Havodivyd, Carnarvonshire, aged 56.

March 25, at Oxford, F. Townsend, Esq. Windsor Herald.

On Tuesday last, Mr. Charles Tomlinson, of this city, aged 68; he had a stroke of the palsy, a week ago, which terminated in his death.

On Sunday, the 21st ult. at Ashton Hall, near Lancaster, aged 53, Mr. Geo. Womack, many years house steward to his late Grace the Duke of Hamilton and Brandon. He locked himself in his own room, soon after twelve o'clock on that day, and fired a loaded pistol through his head, which instantaneously killed him. Coroner's Verdict, shot himself in a temporary fit of insanity.

Figure 20. Capture d'écran d'un journal numérisé et de son OCR

Selon (Conteh, 2009) rapportant l'expérience de la British Library sur les journaux du 19è siècle, en moyenne 20 % du texte d'une page n'est pas correctement océrisé. Comme le constate également (Holley, 2009), le taux d'OCR peut varier de 71 % à 98,02 % d'un périodique numérisé à l'autre. Il est possible d'améliorer ce taux en recherchant des exemplaires originaux imprimés de meilleure qualité, en augmentant la résolution de la numérisation, en utilisant des formats de conservation (TIFF ou JPEG 2000), en utilisant des fichiers en niveaux de gris ou en couleur, en effectuant divers traitements au niveau de l'horizontalité

des lignes de textes ou de la géométrie des pages. Ensuite, bon nombre d'erreurs pourront être corrigées en ayant recours à des dictionnaires de mots. Un contrôle non automatique et une correction par l'humain resteront parfois nécessaires.

S'agissant des écritures manuscrites, en particulier, la reconnaissance de caractères n'en est encore qu'à ses balbutiements (Brokfeld, 2012). Une OCR brute, c'est à dire non corrigée par l'homme, pourra rendre difficile ou parfois même impossible la lecture d'un texte sur tablette par exemple, son interrogation par des recherches en texte intégral, son indexation par les moteurs de recherche ou son annotation par *text mining*.

Pour toutes ces raisons, les bibliothèques qui disposent de moyens financiers suffisants, externalisent ce travail de correction manuelle de l'OCR auprès de prestataires faisant appel à de la main d'œuvre à bas coût, à Madagascar, en Inde ou au Viêt Nam. Une alternative serait d'externaliser ces opérations auprès de la foule des internautes en permettant aux internautes de corriger les textes obtenus, afin d'en améliorer la qualité, de permettre de meilleures recherches en texte intégral, une meilleure indexation par les moteurs de recherche, de produire des fichiers EPUB susceptibles d'être lus sur tablettes, de permettre une réutilisation des données dans le *linked open data*, et de rendre possible des exploitations sémantiques, culturomiques, ou *text mining* des textes.

2.4.1- Le *crowdsourcing* explicite : la correction / transcription volontaire

2.4.1.1- La correction participative et volontaire de l'OCR : l'Australian Newspapers Digitisation Program (TROVE)

Initié en mars 2007 et lancé en août 2008 par la Bibliothèque Nationale d'Australie, ce projet est l'un des premiers et des plus importants projets de correction participative de l'OCR en bibliothèque. Il propose principalement la correction de textes de journaux numérisés depuis le 19è siècle. La bibliothèque numérique de Trove propose toutes sortes de documents, mais c'est la partie journaux qui attire le plus d'audience et de contributions.

Les données statistiques disponibles dans la littérature sont nombreuses et témoignent de la réussite du projet qui est l'un des meilleurs benchmarks dans le domaine. Voici les principales données recueillies dans la littérature :

Date	**Nombre de documents proposés**	**Nombre de contributeurs**	**Nombre d'articles corrigés**	**Nombre de lignes / pages corrigées**	**Nombre de tags**	**Nombre de commentaires**	**Nombre de visiteurs uniques sur le site**
Août 2008			12 000 articles	200 000 lignes			
Octobre 2008		868 utilisateurs enregistrés dont 390 actifs	50 000 articles	700 000 lignes			
Novembre 2008	3,5 millions d'articles, 367 000 pages	1488 utilisateurs enregistrés	60 000 articles	1 million de lignes	18 000 tags	800 commentaires	94 000 visiteurs uniques
Février 2009	toujours 3,5 millions d'articles, 367 000 pages	2994 utilisateurs enregistrés	104 000 articles	2,2 millions de lignes	43 000 tags	1806 commentaires	205 000 visiteurs uniques
Août		5000		6 millions	102 000		

2009		utilisateurs		de lignes	tags (dont 38 000 différents) par 500 utilisateurs		
Novembre 2009	8,4 millions d'articles, 830 000 pages	Plus de 6000 utilisateurs	318 000 articles	7 millions de lignes	200 000 tags		
Mai 2010		9000 utilisateurs		12,5 millions de lignes	424 335 tags	9 192 commentaires	987 147 visiteurs uniques sur janvier-mai 2010
Mai 2012				66 527 535 lignes			
Novembre 2012		77 042 comptes créés dont 6 739 actifs		Plus de 80 millions de lignes		47 450 commentaires	
8 février 2013 d'après http://trov		83 152 comptes créés dont 8		86 221 902 lignes / 8 millions	1 904 630 tags	50 926 commentaires	

e.nla.gov.au/system/stats?env=prod		186 actifs en janvier 2013		de pages dont 2 558 631 sur le seul mois de janvier 2013			
Mai 2014				129 046 297 lignes			
24 mai 2016 (échange de courriels avec Marie-Louise Ayres, responsable de TROVE)				Près de 200 millions de lignes			

Tableau 3. Statistiques collectées dans la littérature à propos du projet TROVE

Ainsi, en moyenne, 2 682 119 lignes de textes sont corrigées chaque mois par près de 30 000 volontaires, en faisant la moyenne sur les 5 premiers mois de 2014 (Zarndt, 2014). (Ayres, 2013) évalue à 425 000 heures de bénévolat, 270 années de travail, 12 millions d'euros la valeur que les 100 millions de lignes de textes corrigés ont rapporté au projet. Sur la base de coûts moyens de correction de l'OCR auprès de prestataires de 0,50 $ pour 1000 caractères et d'une moyenne

de 40 caractères par ligne, Brian Geiger évaluait, en 2012, à 1 378 175 $ le gain, ou plutôt l'argent non dépensé, pour Trove (68 908 757 ligne corrigées) (Geiger, 2012). En mai 2014, nous pourrions évaluer ce coût à 2 580 926 $ en reprenant ce mode de calcul : 129 046 297 / (1000 / 40) x 0,5. Ce calcul a été confirmé par (Zarndt, 2014).

Si on observe le diagramme des statistiques du nombre de ligne corrigées :

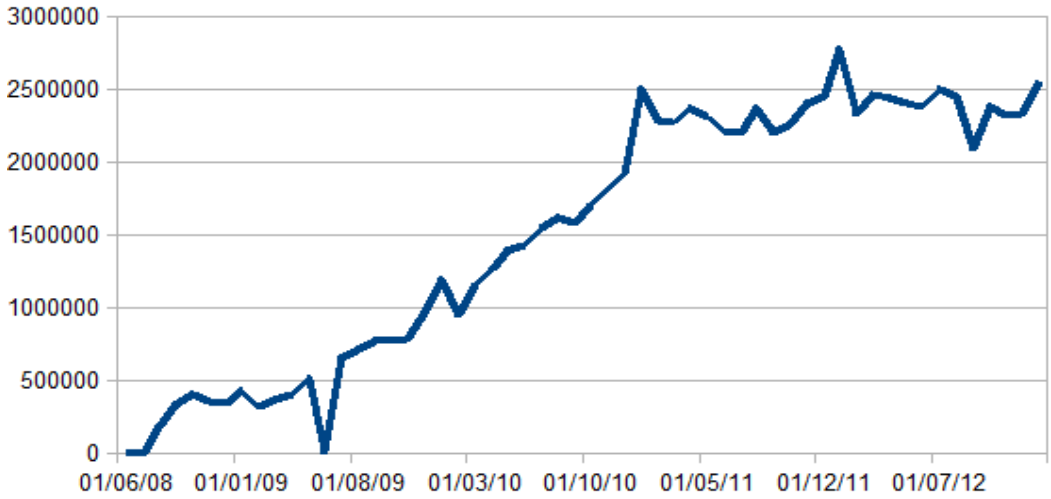


Figure 31. Évolution du nombre de corrections de lignes sur TROVE d'après les statistiques obtenues sur le site lui-même (http://trove.nla.gov.au/system/stats?env=prod)

Il semble qu'un seuil semble néanmoins avoir été atteint et que la croissance du nombre de contributions ne soit plus proportionnelle à celle du contenu mis en ligne. Le "marché" de la correction participative de l'OCR pourrait donc avoir atteint ses limites d'après (Ayres, 2013)

Par ailleurs, d'après (Holley, 2009) 29 % du travail a été réalisé par les 10 plus gros contributeurs qui peuvent consacrer près de 40 heures par semaine à ce travail. Plus récemment, Paul Hagon, senior web designer à la Bibliothèque nationale d'Australie, indiquait que 43 % des corrections (41 millions de lignes de textes) ont été effectuées par les 100 plus gros contributeurs du projet. Le même constat a également été fait pour l'activité tagging du projet, puisque 57 % des tags ont été ajoutés par seulement 10 supers taggers (Holley, 2010).

Comme l'indique (Holley, 2009), au début du projet, la moitié des contributions était le fait de volontaires anonymes et l'autre moitié seulement était le fait de volontaires identifiés. Mais, 6 mois plus tard, 80 % des contributions étaient désormais le fait d'internautes avec login. Cette statistique est confirmée par Paul Hagon, qui mesure à 85 % la proportion de corrections réalisées par des utilisateurs enregistrés. Rose Holley estime que ceci peut s'expliquer par le fait que les internautes ont besoin que leurs contributions soient reconnues et qu'ils soient nommés. D'après (Alam, 2012), la communauté de bénévoles de Trove serait toutefois également intéressée par des motivations intrinsèques (recherches personnelles, altruisme, distraction) plutôt que extrinsèque (reconnaissance, récompenses).

De nouvelles fonctionnalités de curation permettent aux internautes d'ajouter leurs propres journaux numérisés et de créer leurs propres collections publiques ou privées de documents. Ainsi, 40 000 collections (privées ou publiques) ont été créées par des internautes, d'après (Ayres, 2013)

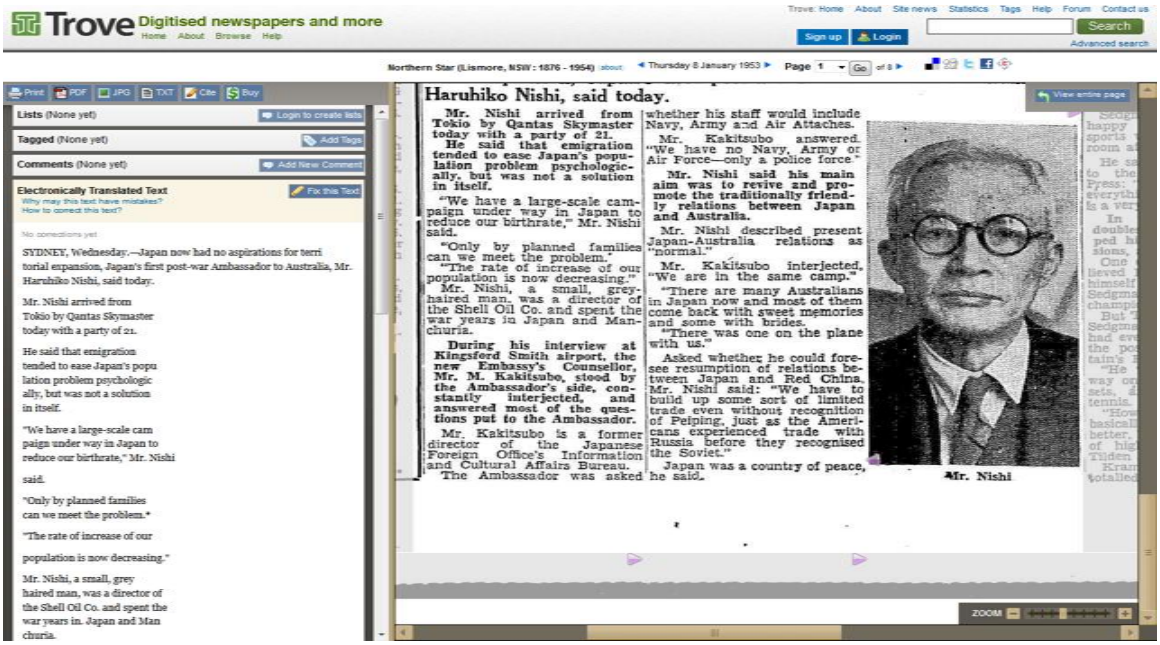


Figure 32. Capture d'écran de TROVE[22]

[22] **On trouve à droite de l'écran la page de journal d'origine au format photo et**

La philosophie de TROVE est fondée sur une confiance importante vis à vis des contributions des internautes et sur l'auto régulation. Les tags ne sont pas modérés et il est possible de corriger un document sans avoir de compte et sans être authentifié. Un système anti-spam reCAPTCHA permet toutefois de s'assurer que les saisies de ces internautes anonymes ne soient pas le résultat de robots. Par ailleurs, toutes les modifications sont enregistrées et peuvent donc être annulées par un administrateur en cas de contribution malveillante afin de restaurer la version antérieure. Néanmoins, comme l'indique (Holley, 2009), il n'y a eu aucun vandalisme détecté sur les 6 premiers mois du projet.

Comme le souligne (Ayres, 2013) 62 % des visites sur les différents sites de la Bibliothèque nationale d'Australie (y compris site web, catalogue, services aux bibliothèques...) viennent de TROVE, et 75 % d'entre ces visites sur TROVE proviennent directement du moteur de recherche Google et accèdent à un document particulier de TROVE, sans passer par une recherche depuis le site de TROVE. Cela accrédite l'idée qu'un bon référencement et une bonne visibilité sur le web sont bien plus importants que les fonctionnalités bibliothéconomiques traditionnelles. En moyenne, 60 000 visiteurs uniques par jour consultent ce site. Ce nombre est en forte croissance et était de 1,8 millions de visites en juin de l'année 2013. Les visiteurs restent en moyenne neuf minutes sur le site (contre trois minutes seulement sur le catalogue de la Bibliothèque Nationale et une seule minute sur son site institutionnel). 40 % des visiteurs sont extérieurs à l'Australie mais proviennent de pays anglophones, 70 % sont des femmes, 65 % ont plus de 50 ans, une proportion forte d'entre eux est plus diplômée et plus riche que la moyenne nationale. Ainsi, TROVE semble finalement surtout intéresser des retraités passionnés d'histoire locale ou de généalogie. Cela pose un problème de représentation de la population que la bibliothèque doit servir et pose la question du devenir du service car rien n'indique que les générations futures de retraités s'intéresseront aussi aux mêmes sujets.

Enfin, bien que nous ne l'ayons pas évoquée dans le chapitre consacré à la numérisation à la demande, la Bibliothèque nationale d'Australie a également été

à gauche, l'OCR à corriger

l'une des premières à proposer un service de numérisation à la demande (Holley, 2011).

2.4.1.2- La transcription participative et volontaire de manuscrits : Transcribe Bentham

Jeremy Bentham est un philosophe et jurisconsulte anglais de la fin du 18è siècle et du début du 19è siècle reconnu comme le père de l'utilitarisme avec John Stuart Mill. Dès 1958, le Bentham project a consisté à publier les œuvres de Jeremy Bentham ("collected works of Jeremy Bentham") conservées jusqu'alors sous la forme de manuscrits. Le projet Transcribe Bentham a, quant à lui, été lancé le 8 septembre 2010 par l'University College London's Bentham Project, en partenariat avec l'UCL Centre for Digital Humanities, l'UCL Library Services, l'UCL Learning and Media Services, et l'University of London Computer Centre. Le projet a reçu le prix Ars Electronica en mai 2011 pour la catégorie Digital Communities (5000 €), et a été financé par une subvention de la Mellon Foundation.

Le projet a bénéficié à compter d'avril 2010 et sur une durée d'un an d'un financement de 262 673 livres sterling du Arts and Humanities Research Council, mais aussi de 2 Research Associates à plein temps chargés de développer, de tester, de recruter, de communiquer, de coordonner et de modérer les contributions, du personnel de l'UCL Library et d'un consultant du Centre for Digital Humanities de l'UCL.

Ce projet est l'un des rares à avoir communiqué sur ses coûts de mise en œuvre :

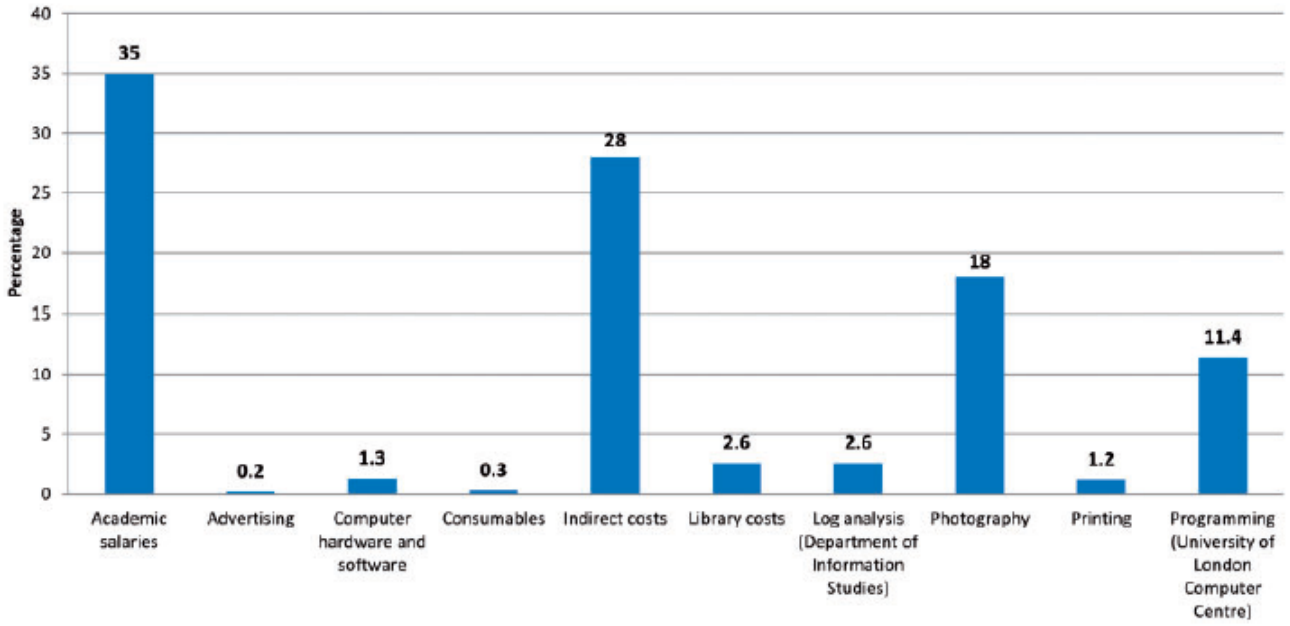


Figure 33. Budget du projet Transcribe Bentham d'après (Causer, 2012)

Sur les 60 000 volumes de manuscrits (30 millions de mots) conservés à l'University College of London, 12 400 ont ainsi été proposé à la transcription participative et à l'encodage TEI sous la forme d'un wiki. (Moyle, 2011). Voici les données statistiques récoltées dans la littérature relative à ce projet :

Date	**Nombre de contributeurs**	**Nombre de manuscrits transcrits**	**Nombre de visiteurs**
Du 8 septembre 2010 à avril 2011 (6 premiers mois)	253 contributeurs (21 % des visiteurs)	500 manuscrits	
Au 8 mars 2011	1207 inscrits	1009 manuscrits dont 569 validés et publiés	15 354 visites de 7 441 visiteurs uniques soit une moyenne de 84 visites par jour

Au 3 août 2012	1726 inscrits	4014 manuscrits dont 3728 validés et publiés	
Au 9 mars 2012	1550 inscrits	2975 manuscrits dont 2758 complets	
Du 28 janvier 2012 au 2 novembre 2012,		51 manuscrits (25 500 mots) par semaine	
Au 2 novembre 2012		4612 manuscrits dont 94 % publiables soit 41 manuscrits (20 000 mots) par semaine	
Du 1er octobre 2012 au 22 février 2013	25 volontaires	639 manuscrits soumis (212 521 mots) qui ont nécessité 8 jours de travail (63 h et 14 min) pour être contrôlés et validés	
Au 15 mars 2013		5243 manuscrits, soit 2080 par an	
Au 28 juin 2013		5279 manuscrits (2 800 000 mots)	

Tableau 4. Statistiques collectées dans la littérature à propos du projet Transcribe Bentham

Des statistiques sous forme de diagramme ont également été publiées par (Causer, 2012) :

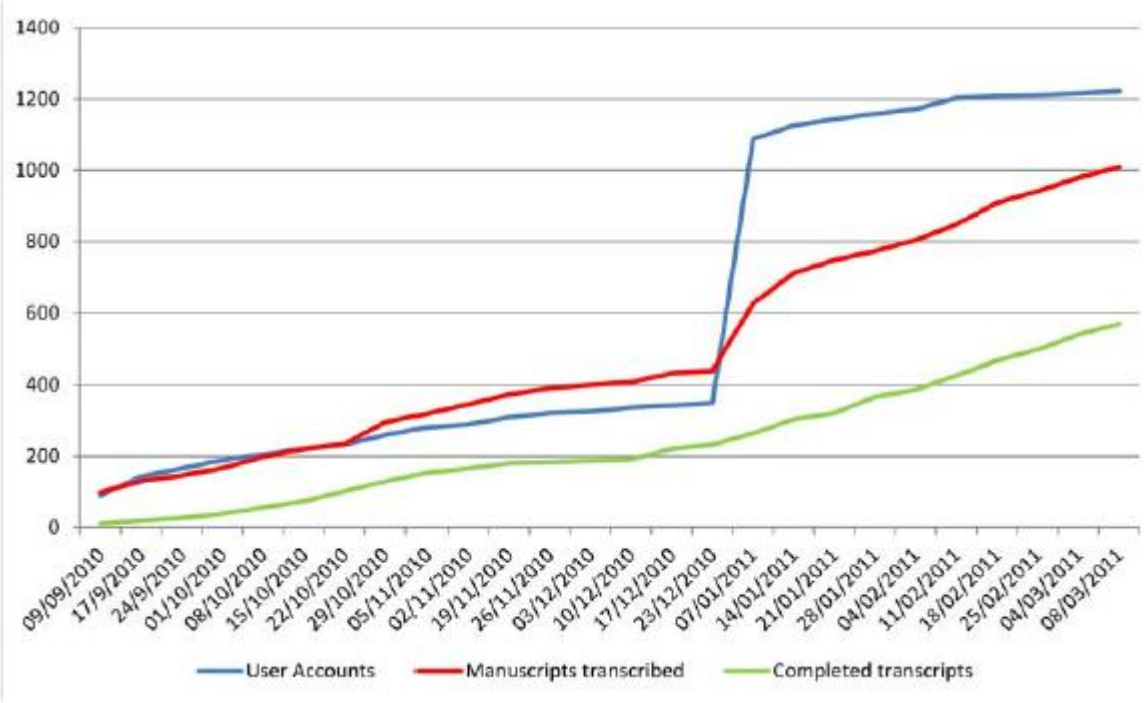


Figure 34. Évolution du nombre de comptes, de manuscrits transcrits et complétés
entre le 8 septembre 2010 et le 8 mars 2011, d'après (Causer, 2012)

32 ans seraient donc nécessaires, au 16 mai 2013, pour transcrire les 20 000 manuscrits de Bentham. Mais, sans l'aide des bénévoles, à un rythme de 549 manuscrits transcrits par an, ce sont 131 ans qui auraient été nécessaires.

Au cours de la période pilote, les 7 plus gros contributeurs ont effectué 70 % des contributions (McKinley, 2012). 15 "super volontaires" ont transcrits chacun entre 6 et 30 manuscrits (Causer, 2012). Cet auteur évoque la notion de *communitysourcing* pour décrire le projet Transcribe Bentham plutôt que celle de *crowdsourcing* dans la mesure où, à l'instar de nombreux autres projets dans le domaine culturel, on ne peut pas véritablement dire que la masse indifférenciée des internautes contribue alors qu'il s'agit plutôt d'une petite minorité de bénévoles.

L'outil MediaWiki, familier aux wikipédiens, a été utilisé pour la transcription. Les développements ajoutés dans le cadre du projet ont donné lieu à un outil qui peut être téléchargé gratuitement. Les textes sont classé par sujet, date, mais

aussi par difficulté de transcription. Afin d'encoder en Text Encoding Initiative (TEI), sans rebuter les néophytes, les transcripteurs disposent d'une barre d'outils sous la forme d'une interface WYSIWYG à partir de laquelle chaque balise XML TEI apparaît sous la forme d'une icône.

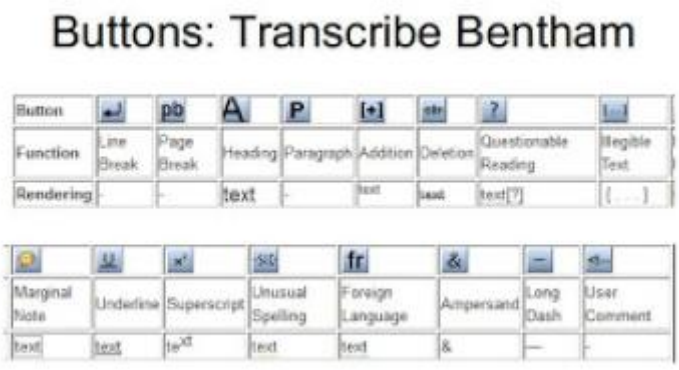


Figure 35. Boutons utilisés par Transcribe Bentham

Pour contribuer, l'authentification est obligatoire. Lorsqu'un transcripteur valide sa production, celle-ci est soumise à l'éditeur du projet pour validation par des experts, puis diffusion. Certains manuscrits sont difficiles à transcrire et un équilibre entre la quantité de textes transcrits et leur qualité a du être trouvé de manière pragmatique. Les manuscrits les plus difficiles à transcrire, rédigé à la fin de la vie de Bentham restent transcrits de manière plus traditionnelle, par des spécialistes.

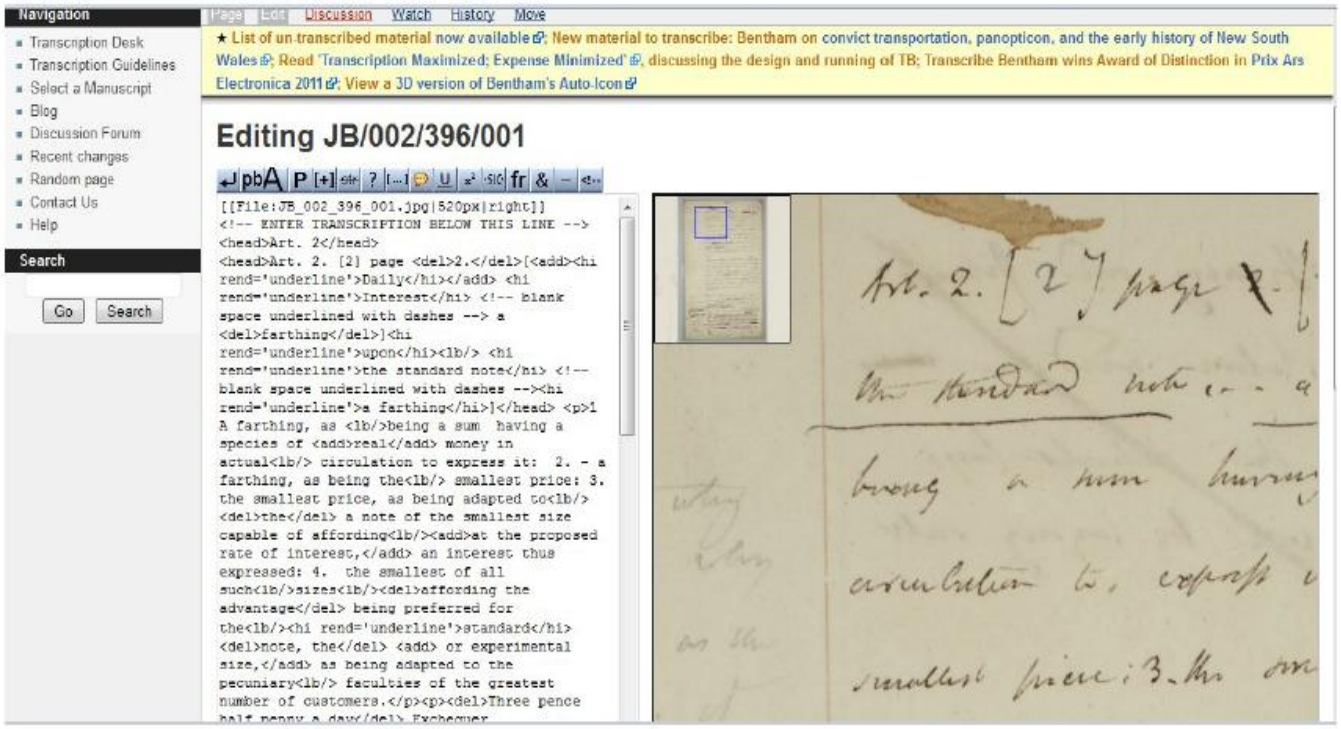


Figure 36. Interface de transcription de Transcribe Bentham d'après (Brokfeld, 2012)

Afin de former les bénévoles, des vidéos screencast[23] leur permettent de visualiser des démonstrations en ligne. Afin de les stimuler, un système de points a été mis en place avec un classement de type *top ten*, ainsi qu'un tableau de bord permettant de suivre l'évolution du projet en temps réel ("Benthamometer"). En fonction de leur niveau, les internautes sont classés, à l'image de ce qui se pratique sur les forums, de "stagiaire" à "génie". Des cadeaux virtuels sont destinés aux meilleurs contributeurs. Les contributeurs sont également nommément remerciés dans les publications des œuvres de Bentham.

Le projet a beaucoup investi dans la communication (communiqués de presse, interventions à la radio, publicités, listes de diffusion, forums, réseaux sociaux comme Facebook et Twitter, blog officiel, vidéos[24], blogs, congrès...). La

[23] http://boinc.cs.uct.ac.za/transcribe_bushman (consulté le 23 juin 2016 mais la page n'était plus accessible)

[24] https://youtu.be/CtEqW4WwMHU (consulté le 23 juin 2016)

publication, en particulier, d'un article dans le New York Times le 27 décembre 2010 a considérablement accru le trafic web. D'autres communiqués ont également été diffusés dans le Sunday Times, la Chronicle of Higher Education (juillet 2010), la Deutsche Welle World radio, et l'Austria's ORF radio. L'achat de publicité Google Adwords (pour 60 £) a également été expérimenté, mais sans résultat notable. L'annonce a été visualisée 648 995 fois et a abouti à 452 clics, mais n'a pas ramené de contributeur vers l'espace de transcription. Une action de communication a également été entreprise auprès des écoles, des universités et des érudits. Des écoliers sont ainsi venus participer au projet avec leur enseignant. Au total, la communication du projet a coûté £800 et l'existence du projet a été signalée et mentionnée dans plus de 70 blogs, 2 émissions de radio et 13 articles de presse. Au 3 août 2012, il y avait 853 abonnés au compte Twitter et 339 fans sur Facebook mais ces réseaux sociaux semblent n'avoir qu'un faible impact sur le trafic vers le site. Ils ont donc plutôt été utilisés pour intégrer la communauté. En conclusion, il semble donc que ce soient les médias traditionnels qui ont permis le mieux de recruter.

Les visiteurs du site viennent principalement des USA et du Royaume Uni en seconde place. Seulement 6 % d'entre eux ont créé un compte (Causer, 2012). Une enquête a également été menée sur 101 personnes et a permis de mieux connaître le profil des contributeurs du projet. Contrairement au projet TROVE, les internautes semblent provenir de manière réduite d'une recherche sur un moteur, ce qui justifie probablement mieux les dépenses en communication consenties dans le cadre du projet.

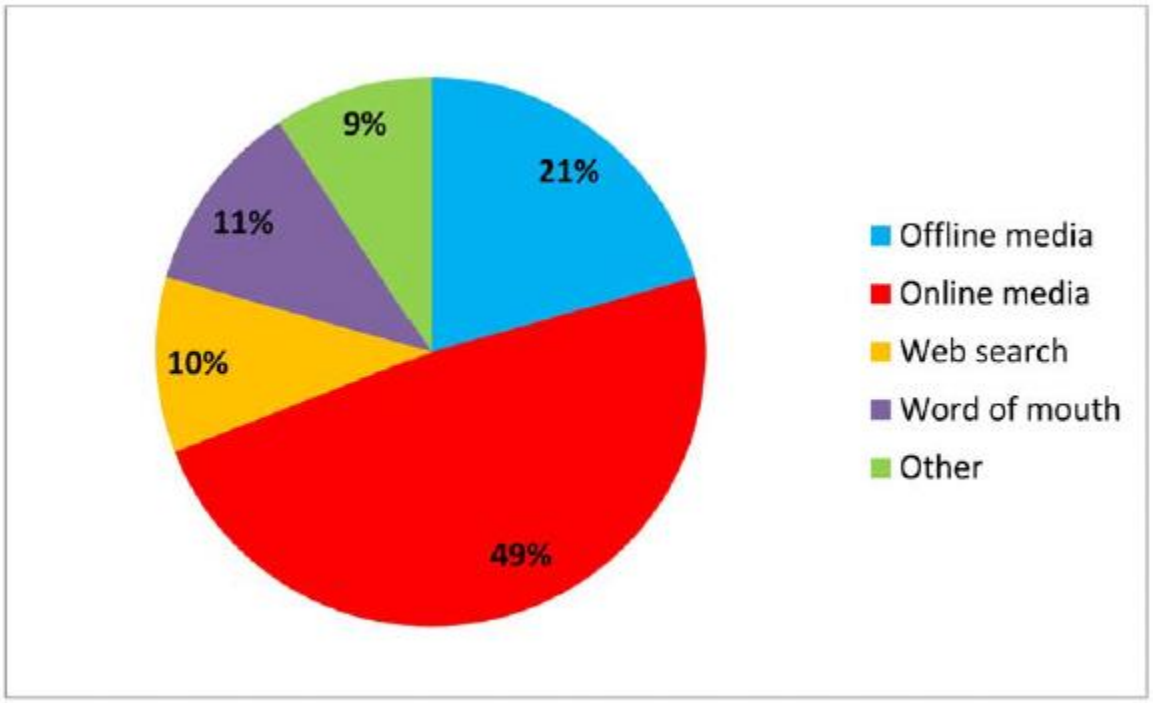


Figure 37. Diagramme représentant comment les internautes ont découvert le projet Transcribe Bentham (d'après Causer, 2012)

97 % des répondants au sondage ont étudié au moins en premier cycle et près d'un quart a un doctorat. Près des deux tiers sont des femmes. Les retraités et les jeunes diplômés sont sur-représentés. La sociologie de contributeurs correspond donc globalement à celle observée pour le projet australien TROVE.

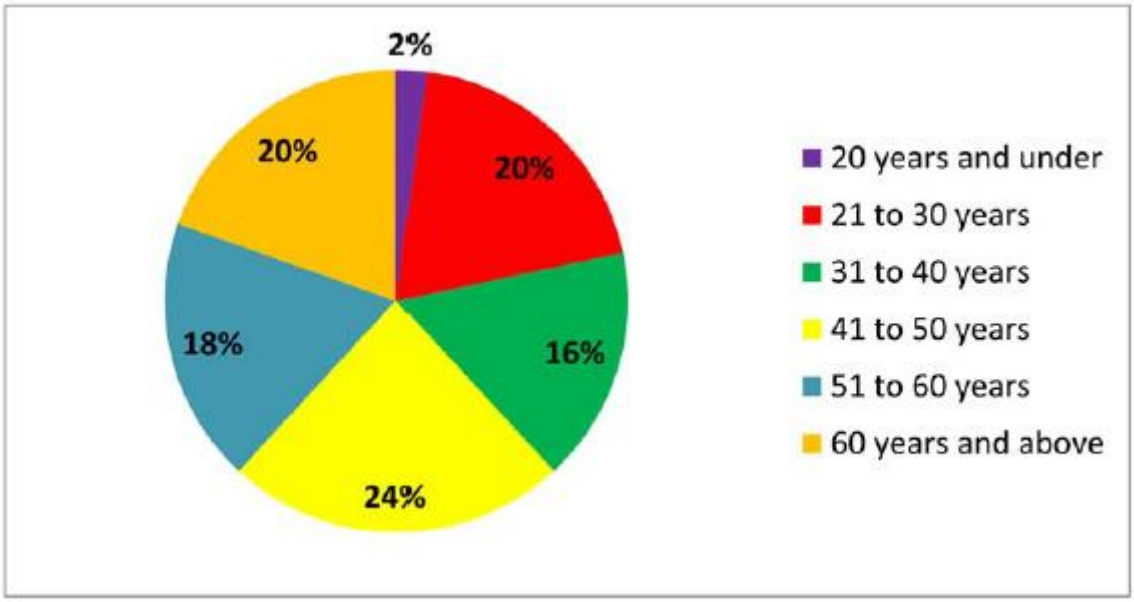


Figure 38. Diagramme représentant la répartition des contributeurs de Transcribe Bentham selon leur âge (d'après Causer, 2012)

Les motivations des contributeurs sont surtout intrinsèques et liées à la fois à l'intérêt pour l'œuvre de Bentham et l'intérêt technologique pour le *crowdsourcing*. Certains volontaires ont notamment évoqué être excités par le fait d'être les premiers à lire des manuscrits encore inédits de Bentham et d'avoir ainsi la sensation de faire de l'"archéologie littéraire".

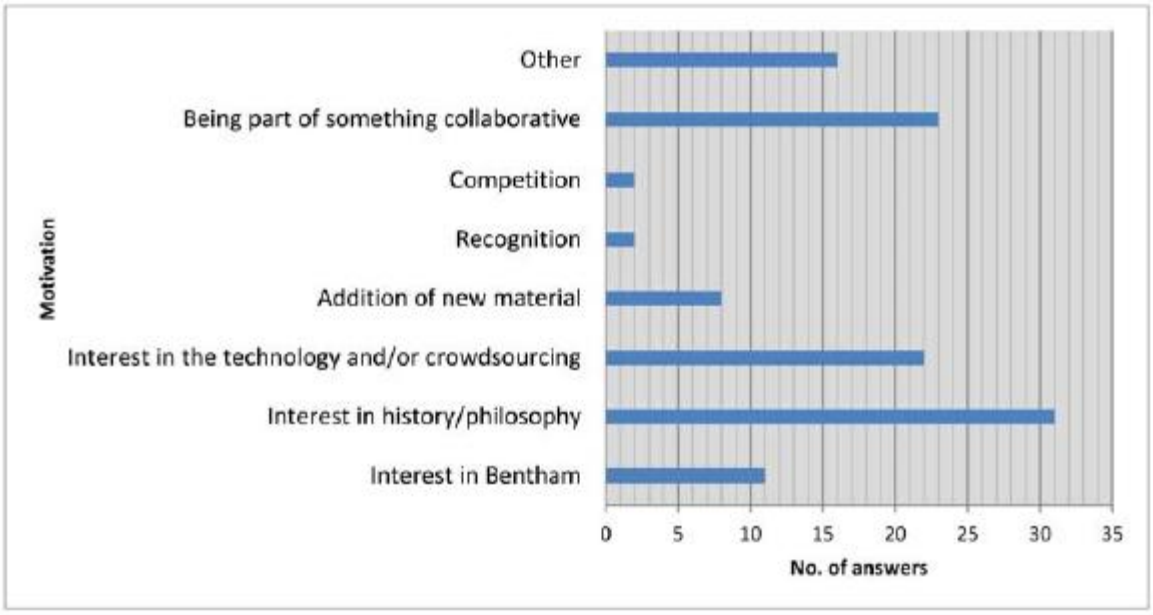


Figure 39. Motivations des volontaires du projet Transcribe Bentham d'après (Causer, 2012)

2.4.2- La *gamification*, la correction de l'OCR en jouant : Digitalkoot (Bibliothèque Nationale de Finlande)

Développé par la société Microtask dont le slogan est "Microtask adore le travail que vous détestez", Digitalkoot est une application de la *gamification* à la correction de l'OCR. La *gamification* consiste à diviser les tâches répétitives comme la correction de l'OCR en microtâches susceptibles d'être proposées aux internautes sous la forme de jeux. L'objectif de ces internautes est donc de se divertir tout en contribuant à un projet culturel ou de contribuer à un projet culturel tout en se divertissant. Contrairement à la plupart des projets de correction participative de l'OCR sous la forme de bénévolat explicite, la correction de l'OCR se fait ici hors contexte, sans prendre connaissance, de manière linéaire, du contenu intellectuel du document.

Le projet, lancé le 8 février 2011, utilise la plateforme développée par IBM, dans le cadre du projet Impact, sous le nom de Microtask. Le nom de Digitalkoot

est inspiré du « talkoot », un mode de construction des habitations finlandaises très ancien qui repose sur l'entraide collective.

Le premier jeu disponible sur Digitalkoot s'appelle “Mole Hunt”. Il s'agit de chasser les taupes qui sortent de terre en validant le plus vite possible si les images de mots qu'elles affichent correspondent bien au texte proposé par l'OCR. Ce jeu permet ainsi de faire valider l'OCR brute en confrontant les validations des joueurs.

Figure 40. Capture d'écran du jeu Mole Hunt

Un deuxième jeu est proposé par Digitalkoot sous le nom de “Mole Bridge”. Au cours de ce jeu, l'internaute doit transcrire le plus rapidement les mots qui sont affichés sous la forme d'images. A chaque fois, sa saisie permet d'ajouter une nouvelle brique au pont qui doit permettre aux taupes de traverser la rivière sans se noyer. A chaque erreur une brique du pont explose. Les décors, la vitesse et la difficulté évoluent à chaque changement de niveaux.

Figure 41. Capture d'écran du jeu Mole Bridge

Sans qu'ils le sachent les joueurs sont évalués grâce à une phase de test et tout vandalisme peut ainsi rapidement être détecté. La qualité globale des données saisies est obtenue grâce à la confrontation des transcriptions des joueurs. Néanmoins, cette méthode est coûteuse et nécessite une participation importante et simultanée difficile à obtenir.

La qualité de l'OCR corrigée obtenue avec le jeu est de 99 % (sur seulement 2 articles comportant 1467 mots pour le premier et 516 mots pour le second, 14 erreurs et 1 erreur ont respectivement été trouvées) alors que l'OCR brute de départ n'avait une qualité que de 85 % en moyenne.

Les statistiques collectées dans la littérature au sujet de Digitalkoot sont les suivantes :

Date	Nombre de visiteurs	Nombre de contributeurs	Nombre de mots corrigés	Temps de travail consacré
31 mars 2011	31 816 visiteurs	4 768 contributeurs (soit 15 % des visiteurs)	2,5 millions de microtâches (soit 118 microtâches par internaute)	2740 heures (soit 9,3 minutes par internaute)
15 septembre 2011	80 000 visiteurs		5 millions de saisies de mots	4000 heures
février 2012 (1 an)		101 614 contributeurs	6 461 659 corrections	5473 heures (=328 376 minutes)
octobre 2012		109 321 contributeurs	8 024 530 microtâches	

Tableau 5. Statistiques collectées dans la littérature à propos du projet Digitalkoot

Si on part du postulat qu'une page compte de 220 à 260 mots en moyenne, cela signifie que ce sont de 30 000 à 37 000 pages qui ont été corrigées par les joueurs en octobre 2012. En moyenne, on peut considérer qu'ils ont ainsi corrigé de 154 à 182 livres de 250 pages. D'après notre expérience de chef de projets de numérisation, la correction de l'OCR d'une page varie entre 1 € et 1,5 €. Nous estimons donc que le projet a rapporté, en octobre 2012, l'équivalent d'une valeur de travail de 31 000 € à 55 000 €.

Le résultat le plus impressionnant de cette expérimentation a été que près d'un finlandais sur 46 a joué à Digitalkoot, soit 109 321 joueurs. Cette forte participation est le résultat d'une campagne de communication dans la presse

internationale (New York Times, Wired), la télévision et les réseaux sociaux ayant habilement joué sur le sentiment patriotique des finlandais (“Start saving … Finnish culture here”). La possibilité de se connecter avec l'aide de son compte Facebook, qui a été utilisée dans 98 % des cas, a également permis de faire connaître le projet de manière virale. Les responsables du projet ont estimés ainsi qu'un bon tiers des joueurs ont fait venir jusqu'au site certaines de leurs relations sur Facebook et que ce réseau social aurait drainé 99 % du trafic web.

Concernant la sociologie des joueurs, ils seraient plutôt jeunes (entre 25 et 44 ans) par rapport à des projets de correction bénévole de l'OCR sans *gamification*. Les femmes représenteraient la moitié des joueurs mais auraient réalisé 54 % du travail et joueraient plus longtemps (13 minutes en moyenne contre 6 minutes concernant les hommes) (Chrons, 2011). Mais les 4 meilleurs participants seraient des hommes, le meilleur d'entre eux ayant joué 101 heures pour 75 000 transcriptions de mots. On constate, là encore, que, comme c'est le cas pour tous les projets de *crowdsourcing*, la majeure partie du travail est l’œuvre d'une infime minorité. Ici un tiers du travail est réalisé par 1 % des contributeurs.

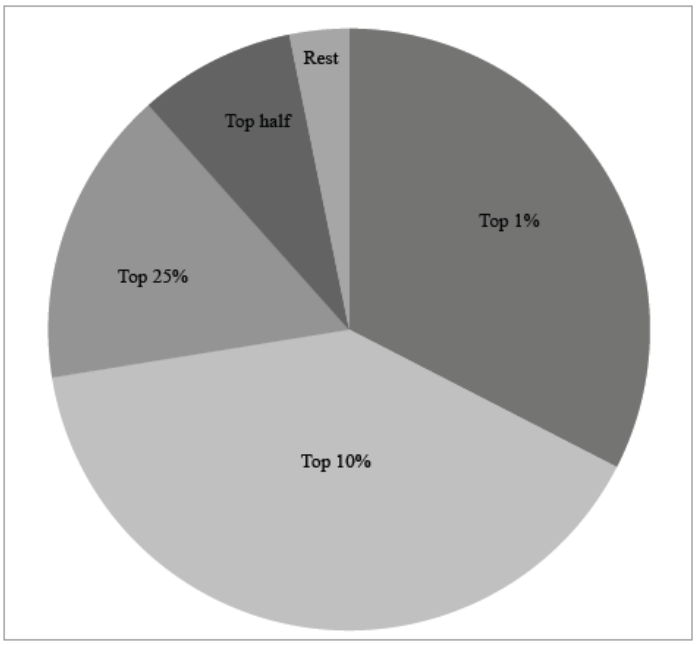


Figure 42. Proportion du travail réalisé par les 1 %, 10 %, 25 %, des meilleurs contributeurs (d’après Chrons, 2011)

Le projet devrait évoluer vers la possibilité pour les internautes de travailler de manière préférentielle sur des thématiques qu'ils affectionnent et vers l'ouverture aux classes scolaires. Une nouvelle application, Kuvatalkoot, devrait aussi permettre, prochainement, aux internautes d'annoter des articles de journaux.

2.4.3- Le *crowdsourcing* implicite : la correction involontaire de l'OCR via reCAPTCHA au service de Google Books

CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) est une application gratuite, mise au point par des chercheurs de l'Université Carnegie-Mellon (Luis Von Ahn, Ben Maurer, Colin McMillen, David Abraham et Manuel Blum). Cette application est destinée à éviter que des robots malveillants comme Googlebot puissent soumettre en masse des requêtes sur des sites web (gestionnaires de courriels comme Gmail ou Yahoo, réseaux sociaux, wikis, blogs...) et paralyser les serveurs ou encore que des comptes e-mail non encore blacklistés puissent être créés automatiquement et en multitude et générer du SPAM. On parle, pour ce type d'application, de Human Interactive Proof (HIP) car système exige que l'internaute, pour créer un compte, saisisse un ou quelques mots déformés et prouve ainsi qu'il est bien un humain et non un robot. Ce système est inspiré du test de Turing. Le test de Turing, décrit par l'informaticien Alan Mathison Turing en 1950, partait de l'hypothèse qu'un ordinateur pourrait être considéré comme intelligent s'il parvenait à avoir une conversation avec un humain sans que ce dernier s'en aperçoive et sans qu'il puisse distinguer la conversation d'un humain et d'un ordinateur. Sur le même principe, le test CAPTCHA permet de reconnaître un humain d'un ordinateur, ce dernier étant à ce jour encore incapable de lire des mots déformés.

Mais, comme cela est expliqué dans un article publié dans la revue Science (Von Ahn, 2008), reCAPTCHA a aussi été utilisé, au delà de sa vocation première, afin de permettre une correction par les internautes du texte océrisé du New York Times numérisé. Depuis le 17 septembre 2009 et son rachat par Google, ce programme est, à présent, utilisé dans le cadre du programme Google Books afin de faire corriger aux internautes, par *crowdsourcing* implicite, le texte des millions de livres numérisés par Google. Son slogan est ainsi devenu "Stop spam, read

books". Depuis mars 2012, Google utilise aussi reCAPTCHA afin de faire corriger par les internautes les photos de numéros de rues tirées de Google Street View afin d'affiner la géo-localisation de Google Maps. D'après un article de Actualitté[25] publié en octobre 2013, la Quadrature du Net, une "association de défense des droits et libertés des citoyens sur Internet", serait en train de développer, un logiciel reCAPTCHA-like pour Internet Archive. Cette information doit pourtant être prise au conditionnel, car le développement d'un tel système exige des moyens suffisants.

Avec reCAPTCHA, les internautes ne sont pas toujours conscient que leur saisies, pour des raisons de sécurité, sont utilisée afin de corriger les contenus de Google Books et de Google Maps. Dans ces conditions, on pourrait qualifier ce type de *crowdsourcing* de « *crowdsourcing* inconscient », de « *crowdsourcing* involontaire » ou encore de « *crowdsourcing* implicite ». C'est ce dernier qualificatif, bien que très peu répandu dans la littérature francophone, que nous retiendrons. En effet, la participation des internautes n'est pas, non plus, nécessairement inconsciente ou involontaire alors qu'elle reste bien, dans tous les cas, implicite.

Le *crowdsourcing* implicite prend en considération le fait qu'une toute petite minorité de bénévoles contribuent à des projets de *crowdsourcing* bénévole (qu'il conviendrait donc d'ailleurs de qualifier de « *communitysourcing* »). Tenant compte des limites de ce *crowdsourcing* explicite, du risque de saturation et des difficultés à recruter de nouveaux bénévoles, le *crowdsourcing* implicite consiste à utiliser de manière astucieuse, écologique et économique les activités courantes des internautes sur Internet, de les détourner pour d'autres fins. Ainsi, l'énergie produite par les internautes afin de ressaisir des mots et créer des comptes sur des sites web est recyclée au bénéfice de la bibliothèque numérique de Google.

Les livres numérisés par le projet Google Books ont systématiquement fait l'objet d'une reconnaissance de caractères par deux différents logiciels d'OCR. Les différences entre les textes obtenues sont comparées avec des dictionnaires de langues. Environ 25 % des mots océrisés sont considérés comme susceptibles

[25] « Exclusif : Un Captcha pour Internet Archive, concurrent de Google Books ». Actualitté, le 28 octobre 2013

d'être des erreurs de reconnaissances. Ils sont alors envoyés sur reCAPTCHA pour être corrigés par des internautes. De leur coté, les internautes, pour créer des comptes sur des sites web sont contraints de saisir deux mots déformés pour prouver qu'ils ne sont pas des robots. L'un des deux mots, déjà corrigé et validé, sert bien un objectif de sécurité. L'autre mot est utilisé, quant à lui, pour faire corriger l'OCR brute de Google Books par les internautes.

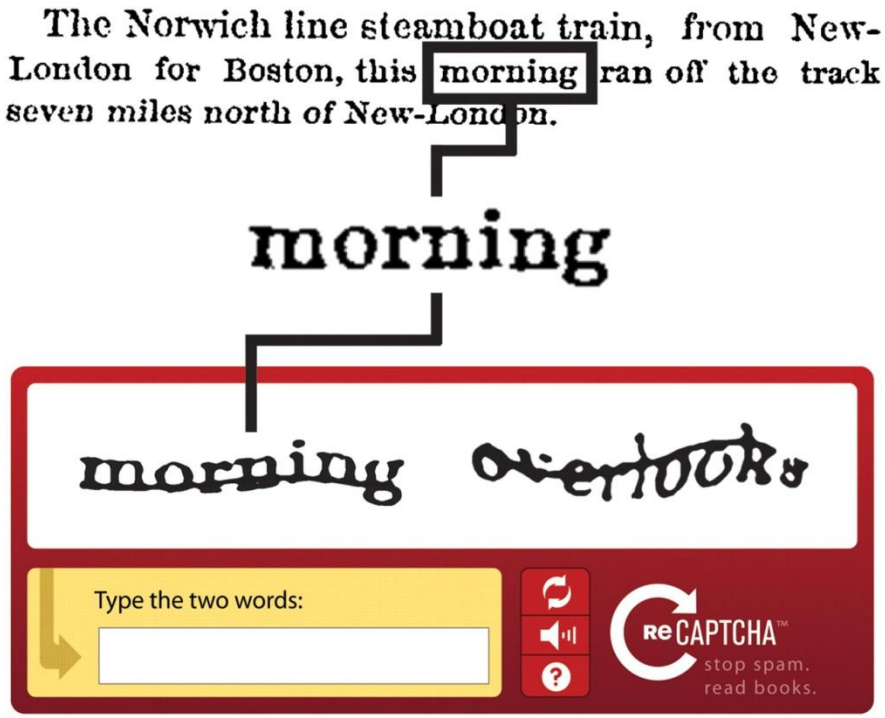


Figure 43. Schéma expliquant comment fonctionne reCAPTCHA d'après le site Google.com

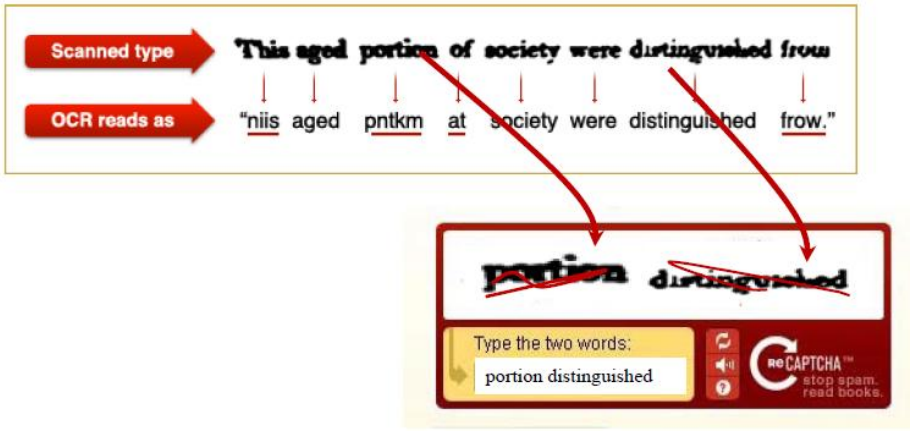


Figure 44. Autre schéma expliquant comment fonctionne reCAPTCHA, d'après (Ipeirotis, 2011)

La méthode classique de confrontation des saisies, bien connue des sociétés qui font de la correction d'OCR brute est ensuite utilisée. Le même mot est envoyé à trois personnes différentes sur le web avec des déformations différentes afin d'éviter que les mêmes distorsions puissent générer les mêmes erreurs de saisies. Si les trois saisies des trois internautes sont strictement identiques, la correction est validée. Dans le cas contraire, le mot à corriger est proposé à des internautes supplémentaires jusqu'à ce que l'une des propositions totalise 2,5 votes (sachant qu'un internaute compte pour 1 vote et qu'une de ses propositions pour 0,5 vote). Parfois, les internautes peuvent être dans l'incapacité d'identifier le mot à ressaisir. Ils peuvent alors demander à ce qu'un nouveau mot soit affiché. Si un mot est ainsi proposé six fois et signalé comme illisible à chaque fois par les internautes, le système le considérera comme impossible à identifier.

Les responsables du projet ReCAPTCHA ont communiqué les statistiques suivantes :

Nombre d'internautes nécessaires pour obtenir une correction valide	Part que cette situation représente sur le total des mots corrigés
2 internautes	68 %
3 internautes	18 %
4 internautes	7 %
5 internautes	3 %
6 ou plus	4 %

Au total, ce système permettrait d'obtenir en moyenne, une OCR corrigée à 99,1 %.

En 2008, reCAPTCHA était installé sur 40 000 sites web (dont Facebook et Twitter) et avait permis de valider plus 440 millions de mots. 17 600 livres et 1,2 billion de mots étaient proposés sur l'année 2008. D'après des statistiques de 2012, près de 200 millions de mots sont ainsi saisis tous les jours par les internautes qui y consacreraient 12 000 heures de travail par jour et ce rythme serait croissant. En 2007-2008, plus d'un milliard de reCAPTCHAs auraient ainsi été résolus (Conteh, 2009).

(Ipeirotis, 2011) fait le calcul suivant. Il y avait 40 millions de ReCAPTCHA saisis chaque jour en 2008, soit 40 000 livres corrigés par jour. Dans ces conditions, il ne faudrait que 12 ans pour corriger 100 millions de livres. Néanmoins, ce calcul confond probablement 40 millions de saisies avec 40 millions de validation. Or, il faut confronter plusieurs saisies pour obtenir une validation. Par ailleurs, si 40 millions de mots font 40 000 livres, un livre ne comporterait que 1000 mots. Or, il ne semble pas que Google Books se soit concentré sur les tirés à parts.

Afin de connaître le nombre de mots validés produits chaque jours et d'évaluer la valeur produite par ce travail, nous avons donc produit une nouvelle estimation :

Nombre d'internautes nécessaires pour obtenir une correction valide	Part que cette situation représente sur le total des mots corrigés	Nombre de mots saisis par jour dans cette situation (sur la base de 200 millions de mots saisis chaque jour)	Nombre de mots validés par jour dans cette situation
2 internautes	68 %	200 millions x 0,68 = 136 millions	136 / 2 internautes = 68 millions
3 internautes	18 %	200 millions x 0,18 = 36 millions	36 / 3 internautes= 12 millions
4 internautes	7 %	200 millions x 0,07 = 14 millions	14 / 4 internautes= 3,5 millions
5 internautes	3 %	200 millions x 0,03 = 6 millions	6 / 5 internautes= 1,2 millions
6 ou plus	4 %	200 millions x 0,04 = 8 millions	8 / 6 internautes= 1,33 millions maximum
		Total :	**86 millions de mots validés chaque jour**

Tableau 6. Statistiques collectées dans la littérature à propos du projet reCAPTCHA

Afin d'évaluer approximativement combien de livres sont ainsi corrigés grâce aux internautes, nous partons du postulat qu'il y a, en moyenne, 75 000 mots dans un livre de 300 pages et comptant 250 mots en moyenne sur chaque page. Si 86 millions de mots sont corrigés et validés tous les jours, on peut

considérer que 86 000 000 / 75 000 = 1147 livres sont corrigés tous les jours, soit 1147 x 30 = 34 410 livres tous les mois et 1147 x 365 = **418 655 livres par an**.

A l'occasion d'une conférence Ted de 2011[26], le fondateur du projet, Luis Von Ahn parle d'un demi-million de livres par an. C'est un chiffre assez proche de nos calculs et qui nous semble plus fiable que les autres estimations précédemment mentionnées.

Il semble néanmoins que le reCAPTCHA pourrait être prochainement abandonné par Google qui ralentirait également son programme de numérisation. Début 2015, lorsque Google avait annoncé le lancement de ce projet, l'objectif de 15 millions de livres numérisés avait été annoncé et avait laissé sceptiques bon nombre de professionnels. Cet objectif a pourtant été atteint et plus que doublé, la barre des 30 millions de livres ayant dès 2013 été dépassée. Avec une vitesse de 418 655 livres corrigés par an, il faudrait plus de 70 ans pour corriger les textes océrisés de ces 30 millions de livres. Le nombre de livres qui ont été imprimés depuis l'invention de l'imprimerie par Gutenberg est estimé à 129 864 880 par Leonid Taycher, un ingénieur de Google d'après un article publié sur son blog le 5 août 2010. La correction des textes océrisés de ces 129 864 880 livres prendrait donc 310 ans, d'après nos calculs. Néanmoins, le nombre d'internautes et donc le nombre de saisies de reCAPTCHA pourraient être croissants.

Mais ces 70 ans pour corriger le contenu actuel de Google Books ou ces 310 ans pour obtenir des textes corrigés de tous les imprimés jamais produits sont des durées brèves en comparaison avec le rythme des bibliothèques publiques pour corriger l'OCR des textes. En effet, ces corrections sont rarement effectuées par les bibliothèques qui proposent des OCR de mauvaises qualité dont le contenu est généralement impossible à consulter sur liseuses, et le plus souvent difficile à indexer, à rechercher, à extraire et à réutiliser. Lorsqu'une correction de l'OCR est effectuée, les bibliothèques ont recours à des prestataires qui exploitent de la main d'œuvre bon marché à Madagascar, en Inde ou au Viêt-Nam.

[26]https://www.ted.com/talks/luis_von_ahn_massive_scale_online_collaboration (consulté le 23 juin 2016)

Si Google Books n'utilisait pas le système reCAPTCHA et faisait appel à ce même type de prestations, il lui faudrait financer entre 1 € et 1,5 € par page de texte corrigée, c'est à dire, pour un livre de 300 pages, de 300 € à 450 €. C'est une somme importante qui explique pourquoi les bibliothèques ne corrigent que très peu l'OCR des textes qu'elles numérisent. D'après nos calculs, nous avons estimé que Google fait corriger l'OCR d'environ 1147 livres par les internautes chaque jour soit 418 655 livres par an. Nous pouvons donc estimer que le *crowdsourcing* implicite lui évite de dépenser entre 1147 livres x 300 € = 344 100 € et 1147 livres x 450 € = 516 150 € chaque jour. Nous estimons donc que Google bénéficie de l'équivalent de **146 millions d'euros par an** de travail gratuit grâce au *crowdsourcing* implicite. C'est un budget bien supérieur à celui que les bibliothèques pourraient consacrer à la correction de l'OCR. En temps de travail, si on considère que les employés corrigent à un rythme de 60 mots par minute, cela représente l'effort de plus de 1700 personnes travaillant 35 heures par semaine.

Si les internautes travaillent ainsi gratuitement plusieurs centaines de milliers d'heures chaque jour et, dans la plupart des cas, sans le savoir, pour le projet Google Books, cette énergie aurait, dans tous les cas, été utilisées pour des raisons de sécurité, conformément à la vocation première de reCAPTCHA. Google, conscient de la valeur de cette énergie, l'a très astucieusement et écologiquement réutilisée aussi pour son programme de numérisation, un programme dont chacun peut, par ailleurs bénéficier. C'est peut être d'ailleurs parce qu'il joue parfois ainsi quasiment un rôle de service public que Google s'est attiré les foudres de l'administration française par exemple. L'idée de réutiliser reCAPTCHA afin d'améliorer la qualité des textes océrisés reste un magnifique exemple d'innovation issue de l'ouverture d'esprit et de la transdisciplinarité. Deux domaines de compétences aussi différents que la sécurité informatique et la numérisation du patrimoine ont ainsi trouvé une application commune innovante.

Néanmoins, Google est parvenu récemment à créer un algorithme capable de contourner le reCAPTCHA. Un nouveau dispositif "No Captcha reCAPTCHA" avec des questions logiques pour prouver que c'est bien un humain qui y a répondu, devraient donc remplacer la méthode précédente.

2.4.4- Le *crowdsourcing* rémunéré : l'Amazon mechanical turk marketplace (AMT)

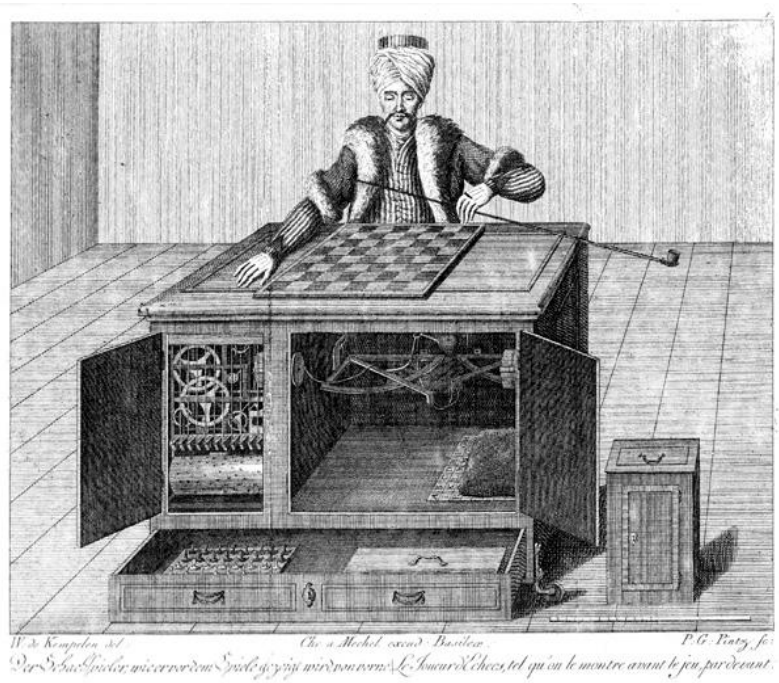


Figure 45. Le joueur d'échec turc. « Tuerkischer schachspieler windisch4 » par Karl Gottlieb von Windisch. 1783.
Public domain via Wikimédia Commons

Au 18è siècle, le turc mécanique (ou automate joueur d'échecs) était une machine automate ayant appartenu au baron von Kampelen et sensée être dotée d'une intelligence artificielle lui permettant de jouer aux échecs. Cette machine aurait notamment vaincu ultérieurement Napoléon Bonaparte et Benjamin Franklin. En réalité, la machine était finalement actionnée par l'intelligence humaine d'une personne simplement cachée à l'intérieur. Amazon s'est inspiré de cette histoire d'humain caché dans la machine afin d'illustrer la nécessité d'utiliser l'intelligence humaine pour réaliser des objectifs encore impossibles aux machines et de montrer que certains travaux, que l'on croit réalisés par des machines, le sont, en réalité, par des humains cachés. Ainsi, la société Amazon a lancé, en grande partie pour ses propres besoins, le 2 novembre 2005, l'Amazon turk marketplace

(AMT), un espace de *crowdsourcing* rémunéré sur le web où les institutions et les sociétés peuvent proposer des microtâches appelées HIT (Human Intelligence Tasks. De l'autre coté de la plateforme, des internautes viennent rechercher des tâches à réaliser. La plupart du temps, ces tâches ne demandent pas une très grande qualification mais restent impossible à faire faire par des algorithmes ou par des programmes. Il s'agit, par exemples, de transcription d'enregistrement vidéos ou audio, d'indexation de documents ou d'image, de classification, de résumés, de votes, d'identification d'images et notamment d'images pornographiques, de rédaction de commentaires et d'avis sur des sites participatifs, d'ajout de relations ou de « likes » sur les réseaux sociaux. Il peut même parfois s'agir de correction de textes d'OCR brutes.

Amazon se rémunère grâce à une commission de 10 % à 20 % pour la mise en place et la maintenance de ce service aux entreprises et aux internautes et aurait ainsi un revenu situé entre 1 et 30 millions de dollars par an pour ce service (Ipeirotis, 2010). La commission perçue par Amazon aurait récemment augmenté en 2015[27].

Avant d'engager un travailleur, le commanditaire peut exiger certaines qualifications et mettre en place des tests de sélection. Des statistiques de réputation des travailleurs sont également accessibles sur le modèle des sites de e-commerce qui permettent de s'assurer des statistiques de tel ou tel vendeur. Une fois le travail effectué, il peut le valider ou le refuser.

[27] D'après Nicolas Gary (24/06/2015). Le Mechanical Turk d'Amazon augmente ses tarifs d'intermédiaire. Actualitte.com

Voici les statistiques collectées dans la littérature (Ipeirotis, 2010 ; Ross, 2010 ; Fort, 2011 ; Göttl, 2014) :

Date	**Nombre de travailleurs et sociétés inscrits**	**Nombre de HITs en cours à n'importe quel moment**	**Nombre de HITs proposés**	**Valeur marchande échangée**
Entre janvier 2009 et avril 2010 (16 mois)			165 368 HITs soit 10 335 HITs par mois	
En avril 2010 (depuis la création de AMT en 2005)	Plus de 400 000 travailleurs 9 436 sociétés	Entre 50 000 et 100 000	6 701 406 HITs	529 259 $
En 2010			Entre 138 654 HITs et 396 106 HITs par semaine	30 000 $ à 40 000 $ par jour 10 millions à 150 millions $ par an
En 2011	Environ 500 000 travailleurs dans 190 pays			
En novembre 2013			Plus de 700 000 HITs par semaine	
En 2014			200 000 HITs par jour	40 000 $ par jour

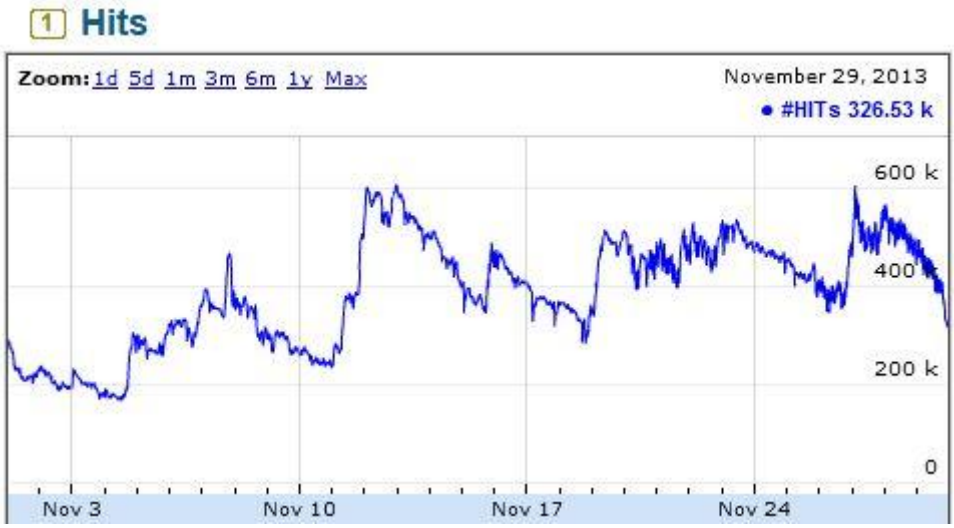


Figure 46. Nombre de HITS en novembre 2013 d'après le Mechanical Turk tracker

D'après (Ipeirotis, 2010), les travailleurs de la marketplace, viendraient principalement des États-Unis, seraient plus jeunes et plus instruits que la population en général et seraient surtout intéressés par un moyen d'obtenir un peu d'argent. Leurs motivations seraient donc à la fois intrinsèques (une manière fructueuse et amusante de passer du temps libre) et extrinsèques (une source de revenus) (Kaufmann, 2011). Mais, avec le développement du chômage dans les pays occidentaux et depuis que la plateforme permet aux participants indiens de pouvoir être payés en roupies, ce type de motivation pourrait laisser place à la satisfaction de besoins beaucoup plus nécessaires.

En février 2010, une enquête a été menée par Panos Ipeirotis sur la plateforme en payant chaque participant 10 cents, puis une autre étude, menée par (Ross, 2010) a également été produite en rétribuant également 0,10 $ sur l'Amazon Turk Mechanical Marketplace chaque travailleur répondant à une enquête de moins de 2 minutes. Ces enquêtes révèlent que la part des travailleurs américains (47 %) diminuerait au profit des travailleurs indiens (34 % pour Ipeirotis, 36 % pour Ross). Cette tendance est d'ailleurs relativement en cohérence avec celle rapportée par (Fort, 2011) et qui signale que les travailleurs indiens représentaient 10 % des travailleurs de la plateforme en 2008, 33 % en 2010 et 50 % en mai 2010.

Les travailleurs américains seraient plutôt féminins, alors que les travailleurs indiens seraient, au contraire, plutôt masculins, mais, d'après (Ross, 2010), la proportion d'hommes aurait tendance à augmenter :

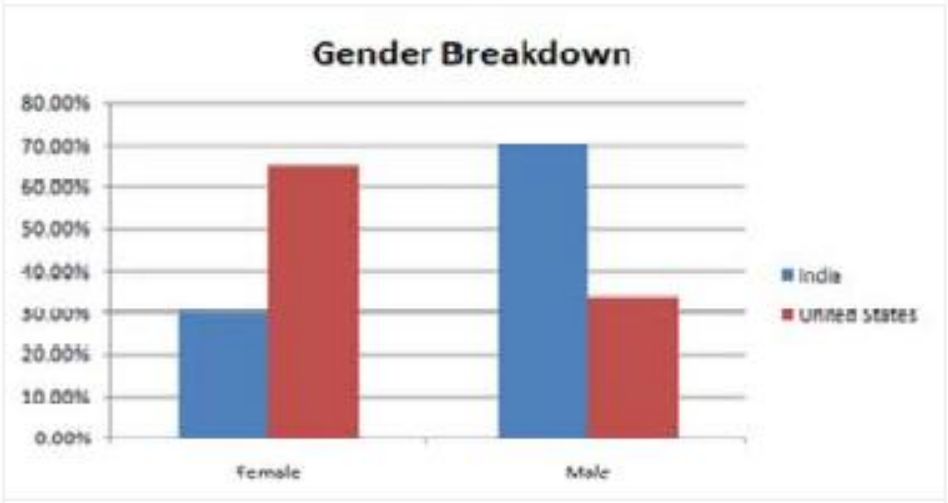


Figure 47. Répartition des travailleurs indiens et américains sur l'AMT le sexe (d'après Ipeirotis, 2010)

Si on considère que les femmes américaines sont d'avantage touchée par le chômage et le temps partiel, ce résultat n'a rien d'étonnant.

Les travailleurs semblent généralement être relativement jeunes et leur moyenne d'âge aurait même tendance à diminuer. Ils seraient sans enfants, et seraient assez diplômés :

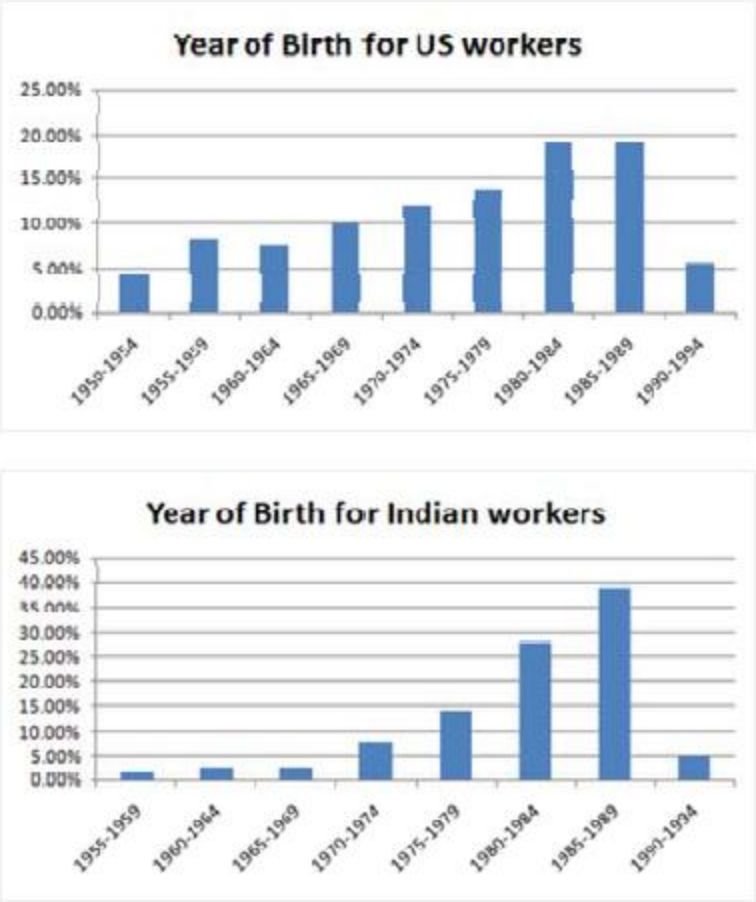


Figure 48. Années de naissance des travailleurs sur l'AMT (d'après Ipeirotis, 2010)

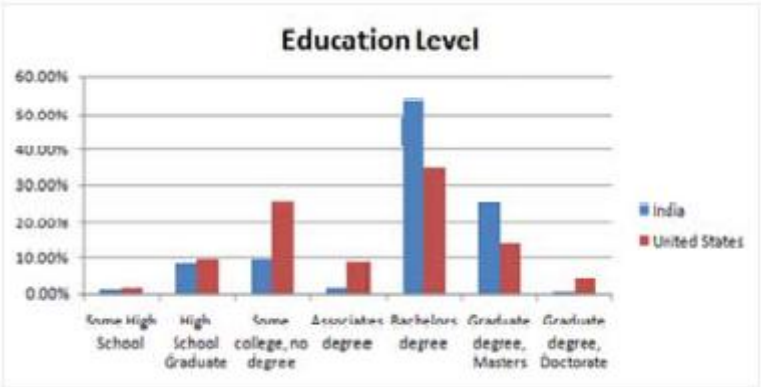


Figure 49. Niveau scolaire des travailleurs sur l'AMT (d'après Ipeirotis, 2010)

La majeure partie d'entre eux passeraient entre 4 et 8 heures par semaine sur la plateforme :

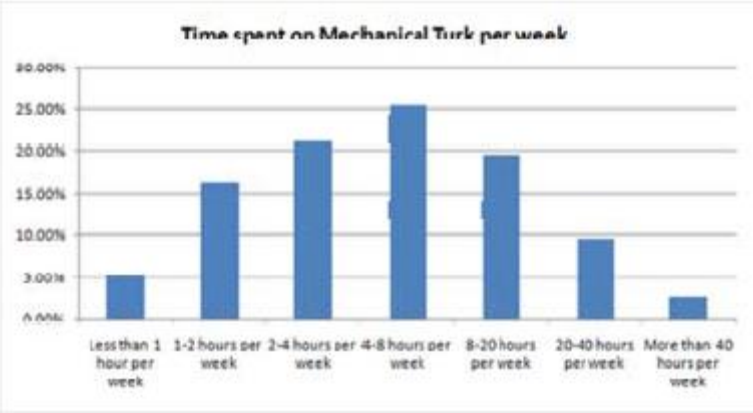


Figure 50. Temps moyen consacré à l'AMT (d'après Ipeirotis, 2010)

La grande majorité des travailleurs s'y consacreraient moins de 5 heures par semaine, mais 18 % d'entre eux travailleraient quand même plus de 15 heures par semaine.

Ils gagneraient entre 1 et 5 $ par semaine :

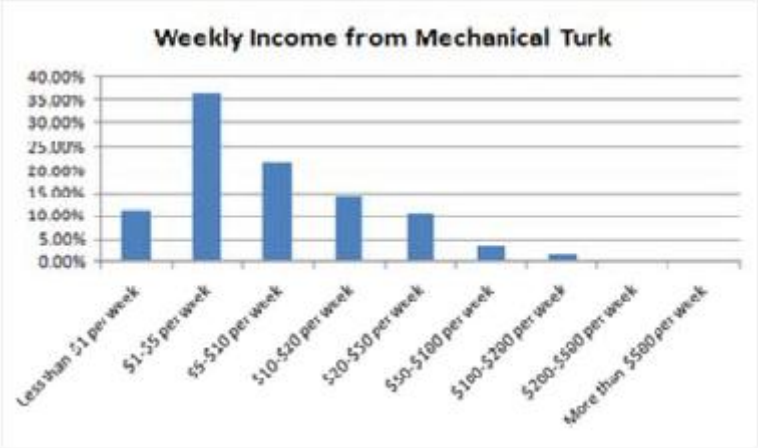


Figure 51. Revenus moyens tirés de l'AMT (d'après Ipeirotis, 2010)

La valeur moyenne du HIT serait de seulement 7,9 cents (Ross, 2010). 10 % des HITs seraient payés seulement 0,02 $ maximum, 50 % autour de 0,10 $ et 15 % d'entre eux seraient payés 1 $ ou plus (Ipeirotis, 2010). Dans ces conditions, et selon un sondage réalisé auprès de 400 000 travailleurs inscrits sur MTurk, l'utilisateur américain moyen aurait gagné 2,30 $ par heure en 2009 et l'utilisateur indien 1,58 $ par heure (Ross, 2010).

Mais peu d'entre ces travailleurs (27 %) considéreraient qu'il s'agit de leur première source de revenus :

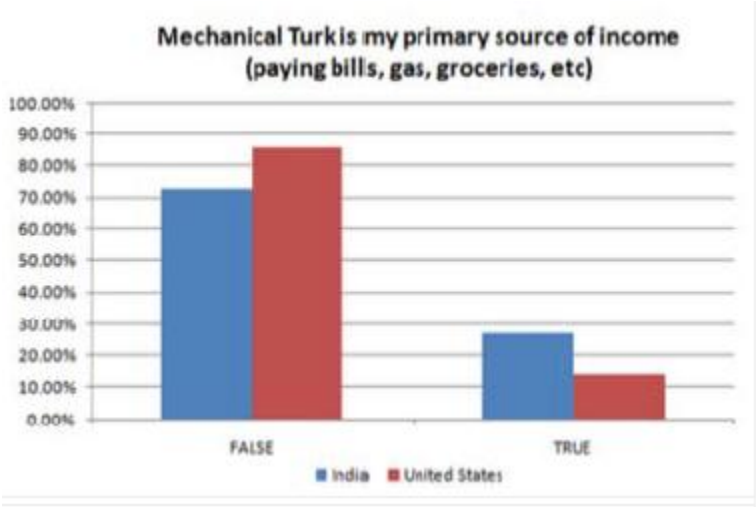


Figure 52. Nombre de travailleurs déclarant que l'AMT est leur première source de revenus (d'après Ipeirotis, 2010)

Les seulement 0,1 % des entreprises ayant proposé le plus de HITs représenteraient une part non négligeable de 30 % de l'activité de la plateforme. De la même manière, d'après (Fort, 2011), 80 % des HITS seraient réalisés par seulement 20 % des travailleurs qui passeraient plus de 15 heures par semaine sur la plateforme. Ces travailleurs pourraient être entre 3011 et 8582 d'après les calculs de l'étude. 20 % d'entre eux considéreraient cette activité comme leur principale source de revenus et 50 % seulement comme une source secondaire de revenus.

Les travailleurs les moins assidus seraient motivés par un moyen de passer le temps, à l'inverse des travailleurs les plus assidus qui seraient d'avantage motivés par la variété des tâches et l'identification à une communauté.

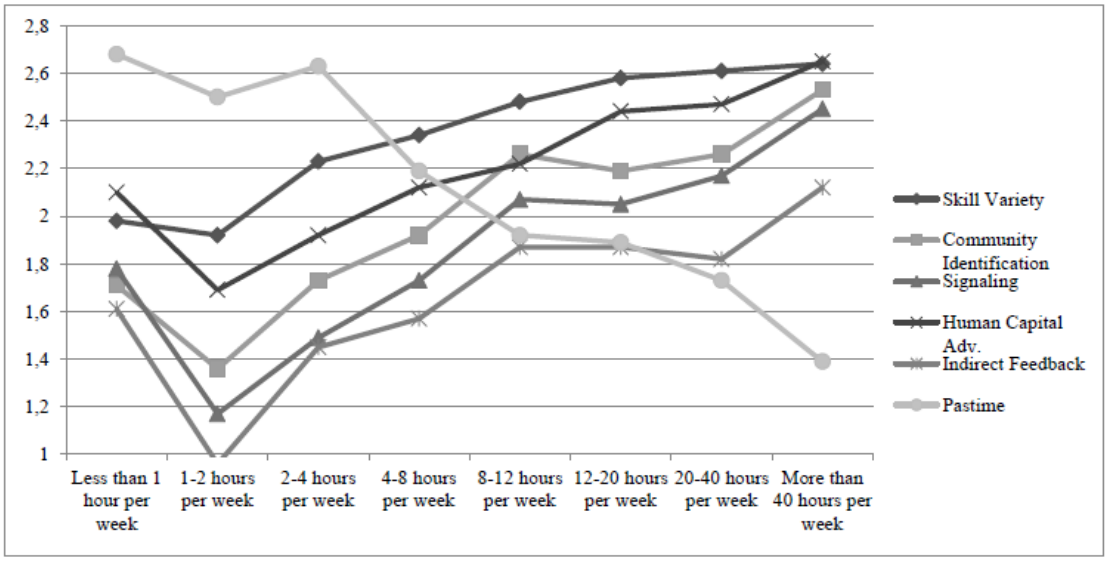


Figure 4: Mean construct score distribution over Weekly Time on MTurk

Figure 53. Types de motivations en fonction de la plus ou moins grande assiduité des travailleurs sur la plateforme AMT d'après (Kaufmann, 2011)

En fonction de l'expérience accumulée (nombre de HIT approuvés supérieur ou égal à 1000) et de la qualité du travail fourni (taux de HITs approuvés supérieur ou égal à 95 %), un travailleur peut obtenir la qualification de "Master" worker". La flexibilité et l'autonomie procurées par ce type de travail pourrait être satisfaisantes y compris du point de vue des salariés. Ainsi, une travailleuse américaine de 28 ans témoigne que sur la base de 5000 HITs par semaine, elle y travaille à plein temps et gagne plus d'argent qu'avec son ancien travail, tout en ayant la liberté de choisir, en fonction de ses besoins, de ses compétences et de ses envies, son lieu de travail, ses horaires de travail, sa durée de travail, son type de travail, et même ses patrons devenus ses clients. (Deng, 2013). Dans un contexte de développement de modèles "à la demande" où le jobcrafting se développe et où les salariés souhaitent de plus en plus modifier leurs fiches de

poste, ce type de plateforme pourrait donc être très attractif et bénéficier autant aux employeurs qu'aux employés.

Néanmoins, l'omniprésence de travailleurs malhonnêtes et de spammeurs non écartés et non sanctionnés sur la plateforme conduirait à une dépréciation de la valeur du travail sur la plateforme et à des prix très bas, les meilleurs travailleurs finissant par partir (Sagot, 2011). Ainsi, selon (Ross, 2009) rapporté par (Sagot, 2011), 70 % des Turkers utiliseraient la plateforme depuis moins de 6 mois.
Les chercheurs universitaires seraient les premiers utilisateurs[28].

D'autres plateformes sur le même modèle que l'Amazon Mechanical Turk existent et fonctionnent sur un modèle assez similaire : 99design, CloudCrowd, Cloud-Flower, CrowdFlower, eLance, Foule Factory, Freelancer, Guru, Innocentive, ManPower, Mob4hire, MobileWorks, oDesk, Postmates, quora.com, Samacource, sparked.com, TaskRabbit, Topcoder, Trada, Turkit, uTest…

Concernant l'utilisation de *crowdsourcing* rémunéré au bénéfice de projets de numérisation du patrimoine des bibliothèques qui nous intéresse plus particulièrement dans le cadre de cette thèse, des expérimentations de transcription de manuscrits et de correction d'OCR sont relatées (Lang, 2011) (Saylor, 2011). Une institution a, par exemple, mis son OCR brute sur des Google Docs ouverts en écriture puis a payé des internautes dans le monde pour effectuer la correction de l'OCR via l'Amazon Mechanical Turk Marketplace. La liste des URL de chaque Google Doc a simplement été versée en tant que fichier CSV sur la plateforme. Après le travail de correction effectué, d'autres internautes ont ensuite été payés pour contrôler l'OCR corrigé dans lequel des erreurs avaient volontairement été introduites afin de vérifier que le travail avait bien été effectué.
Voici les résultats comparatifs entre les coûts induits par l'Amazon Mechanical Turk Marketplace et l'estimation de ceux qui auraient été ceux d'un prestataire traditionnel :

[28] D'après Nicolas Gary (24/06/2015). Le Mechanical Turk d'Amazon augmente ses tarifs d'intermédiaire. Actualitte.com

Travail effectué	Coûts (avec l'Amazon Mechanical Turk Marketplace)	Estimations (avec un prestataire traditionnel)
transcription d'un document de 6 à 8 pages	0,08 $ (soit 0,06 €) en une semaine environ	
relecture de la transcription d'un document de 6 à 8 pages	0,10 $ (soit 0,07 €) en une semaine environ	
transcription avec contrôle qualité d'un document de 6 à 8 page	0,18 $ (soit 0,13 €)	de 2 à 8 $ (soit de 1,45 à 5,8 €)
coût total 72 pages	22,86 $ (soit 16,5 €)	144 à 576 $ (soit 104,3 à 417,3 €)
coût total 200 pages	60 $ (soit 43, 5 €)	400 $ (soit 290 €)

Tableau 8. Coûts comparés entre une correction d'OCR via l'AMT et via un prestataire

Une autre expérimentation autour de *crowdsourcing* rémunéré a également été relatée autour de l'annotation de gravures de fleurs conservées au Rijksmuseum d'Amsterdam avec l'aide de la plateforme CrowdFlower (Oosterman, 2014 et Oosterman, 2014 bis). Mais, la foule n'était peut être pas la bonne cible pour ce projet et le Rijksmuseum a constaté qu'il avait plutôt besoin de d'amateurs, experts, d'autodidactes et de professionnels retraités, c'est à dire de faire appel au *communitysourcing* plutôt qu'au *crowdsourcing*. (De Boer, 2012)

En nous inspirant des calculs de (Ipeirotis, 2011) de (Geiger, 2012), de (Zarndt, 2014) et de nos propres estimations, nous avons estimé quels pourraient être le bénéfice financier de différents projets de correction participative de l'OCR en calculant ce qui aurait été payé par les institutions s'ils avaient eu recours à de la main d'œuvre pour effectuer les corrections d'OCR.

Projet	Coût non dépensé
California Digital Newspaper Collection	53 130 $ cumulés en juin 2014
TROVE	2 580 926 $ cumulés en mai 2014
Digitalkoot	Entre 31 000 et 55 000 € cumulés en octobre 2012
Google Books et reCAPTCHA	146 millions d'euros par an au rythme de 2008

Tableau 11. Estimation du coût non dépensé en prestation de correction d'OCR grâce au recours au *crowdsourcing*

Le *crowdsourcing* appliqué à la correction de l'OCR présente donc un enjeu économique loin d'être négligeable. Il permettrait d'éviter aux bibliothèques de gaspiller un argent public devenu si rare, comme elles le font actuellement en faisant travailler de la main d'œuvre dans des pays en voie de développement et dans des conditions parfois critiquables tout en permettant d'améliorer la qualité des textes produits dans le cadre de leurs projets de numérisation afin d'offrir de meilleures possibilités de recherches en texte intégral, une meilleure visibilité sur Internet et une meilleure indexation des contenus par les moteurs de recherche, de produire des fichiers lisibles sur tablettes et de permettre une réutilisation et une exploitation sémantique des données textuelles.

A la lecture de ce tableau comparatif, on constate également que les résultats obtenus pas Google Books avec reCAPTCHA sont sans commune mesure avec les meilleurs résultats obtenus par les sites de bibliothèques ayant recours au *crowdsourcing* classique de volontaires. Le type de *crowdsourcing* utilisé est également très différent puisque dans certains cas, l'internaute participe volontairement à la correction de l'OCR (TROVE, California Digital Newspaper Collection…) et dans d'autres, il est utilisé involontairement sans même qu'il en ait souvent conscience (reCAPTCHA). Ce type de *crowdsourcing* inconscient et involontaire est généralement qualifié en anglais de "implicit *crowdsourcing*" et pourrait être traduit en français par "*crowdsourcing* implicite" bien que ce terme soit encore quasiment inexistant dans la littérature française en septembre 2014 :

	Nombre d'occurrences sur Google	**Nombre d'occurrences sur Google Scholar**
***crowdsourcing* implicite**	6 occurrences	0 occurrence
implicit *crowdsourcing*	3320 occurrences	50 occurrences
***crowdsourcing* involontaire**	0 occurrence	0 occurrence
involuntary *crowdsourcing*	50 occurrences	1 occurrence
***crowdsourcing* inconscient**	7 occurrences	0 occurrence
unconscious *crowdsourcing*	1 occurrence	0 occurrence

Au delà de la division entre *crowdsourcing* implicite et explicite, l'étude de divers projets nous a permis de discerner des projets avec ou sans *gamification*, de correction de textes imprimés ou de transcriptions de manuscrits, de correction dans le contexte ou de correction hors contexte.

La correction classique de l'OCR dans le contexte du texte répond bien aux motivations des bénévoles qui souhaitent profiter de leur contribution afin de prendre connaissance et lire des textes qui présentent un intérêt pour eux. On a ainsi vu certains volontaires devenir des spécialistes de tel ou tel sujet grâce à leur participation à des projets de *crowdsourcing* explicite.

Mais, ce type de *crowdsourcing* reste moins efficace que la *gamification* ou le *crowdsourcing* implicite qui proposent de la microtâche et de la correction hors contexte qui ne permettent pas de développement personnel des contributeurs.

Comme nous l'avons déjà dit, le *crowdsourcing* implicite prend en compte le fait que seule une petite minorité d'individus participe à des projets de *crowdsourcing* explicite. Il est donc plus efficace de recycler et de réutiliser l'énergie des internautes au cours de leurs activités courantes sur Internet. Le *crowdsourcing*, explicite, plus classique, utilisé depuis plus de 10 ans, pourrait, par ailleurs connaître ses limites, alors même qu'il commence tout juste à s'implanter en France et à faire l'objet d'études comme celle de (Moirez, 2013).

En effet, comme le suggère (Ayres, 2013) et comme semblent l'indiquer les statistiques du projet TROVE, nous pourrions désormais avoir atteint un seuil critique pour le *crowdsourcing*. Le temps de travail bénévole disponible sur le web n'est pas infiniment extensible et il est réparti entre des projets de plus en plus nombreux. *Crowdsourcing* implicite et explicite pourraient ainsi coexister et se compléter. Dans le premier cas, les internautes jouent ou créent un compte sur le web et améliorent involontairement la qualité de textes indéterminés (ou parmi les plus consultés sur le web). Dans le second, ils cherchent à la fois à consommer des textes déterminés et à en améliorer la qualité. Néanmoins, l'amélioration des capacités des logiciels de reconnaissance de caractères pourrait également rendre caduques les projets de correction participative de l'OCR dans les 5 prochaines années, l'œil humain pouvant progressivement être remplacé par des algorithmes. La frontière entre ce qu'il est possible de faire faire automatiquement par la machine et ce qui doit encore être pris en charge par l'homme est une frontière mouvante. A mesure que les technologies OCR s'amélioreront, le recours au *crowdsourcing* deviendra de moins en moins nécessaire et le recours aux bénévoles pour corriger les textes pourrait ainsi bientôt ne plus avoir d'intérêt que pour la transcription de manuscrits.

Dans tous les cas, si on considère l'évolution du nombre de corrections dans le plus grand projet de correction participative de l'OCR, le projet australien TROVE, on observe effectivement un ralentissement ces dernières années :

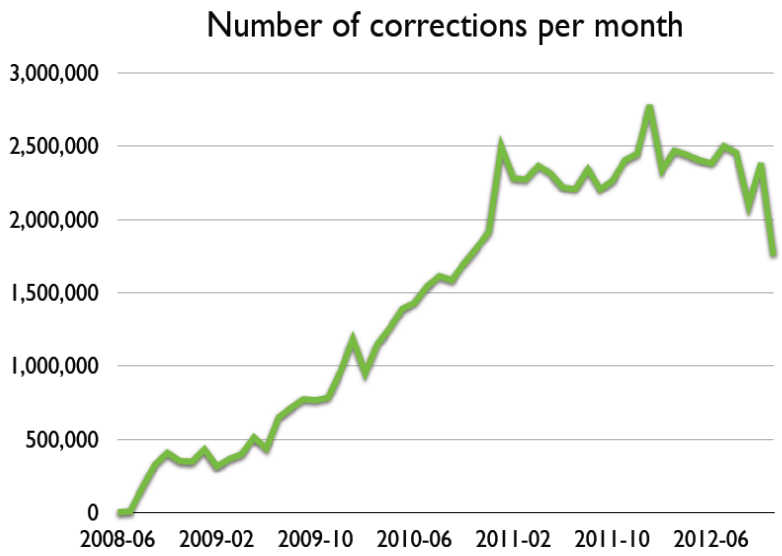


Figure 58. Nombre de corrections sur TROVE entre 2008 et 2012, d'après (Hagon, 2013)

Et cette stagnation ne s'explique pas par une stagnation du nombre de contenus proposés à la correction car il a continué à croître comme l'illustre le diagramme suivant :

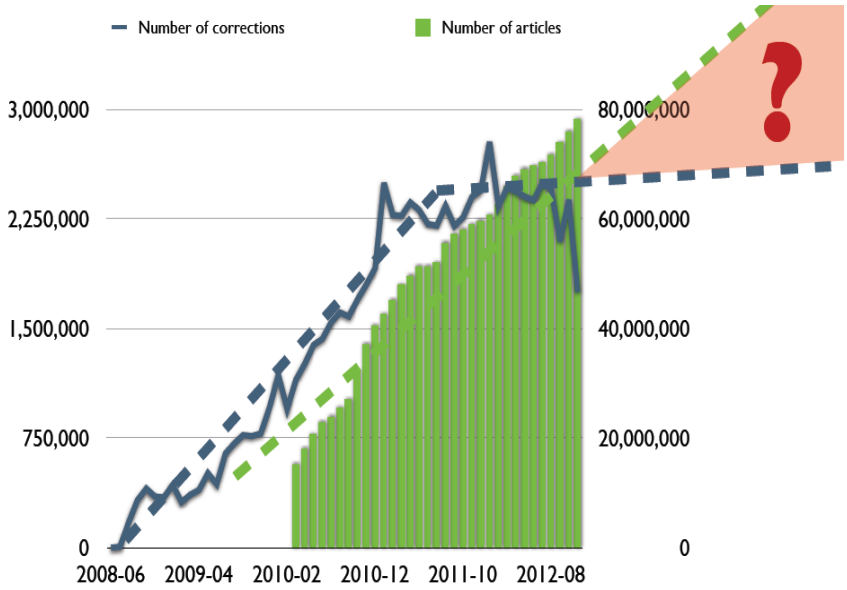


Figure 59. L'évolution du nombre contenus comparée

à celle du nombre de corrections sur TROVE, d'après (Hagon, 2013)

En théorie le *crowdsourcing* explicite s'adresse à des foules de participants. En réalité, la majorité des contributions provient d'une toute petite minorité de volontaires motivés. Les contributeurs des projets TROVE, Cambridge Public Library, ou encore California Digital Newspaper Collection sont des généalogistes :

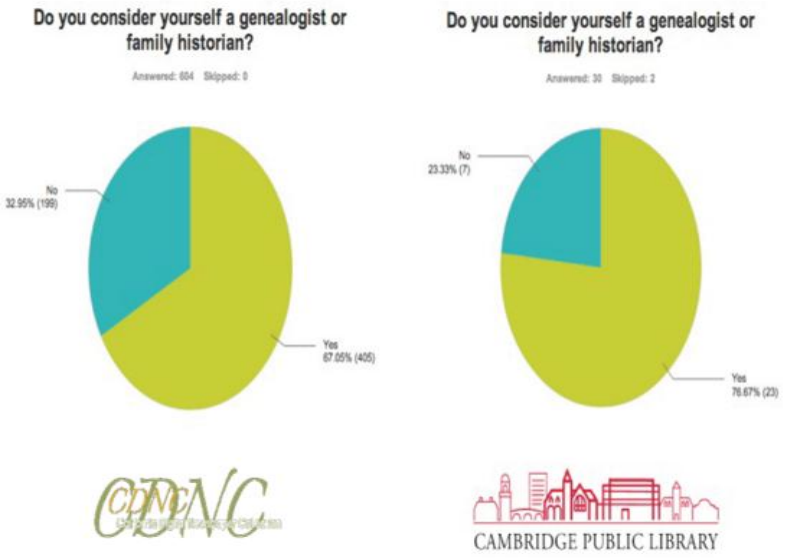


Figure 60. Part des généalogistes parmi les contributeurs d'après une enquête CDNC / Cambridge Public Library

Ce sont également plutôt des retraités (ce qui ne serait pas le cas avec la *gamification*) :

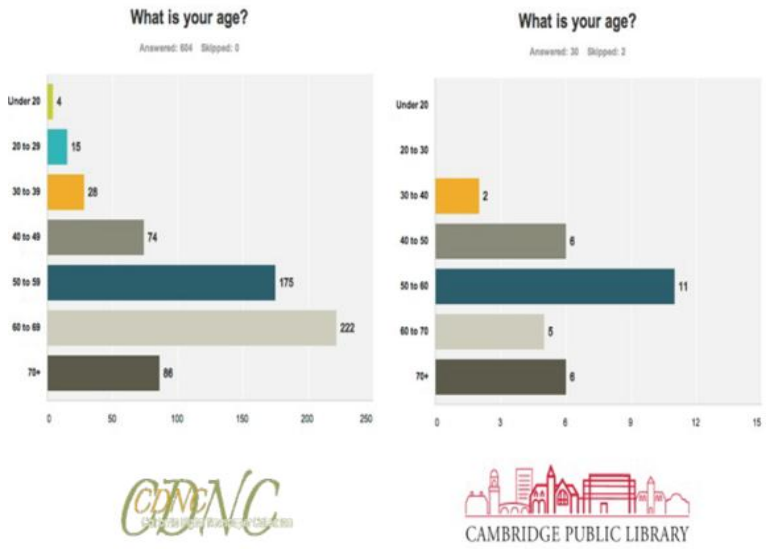


Figure 61. Répartition des bénévoles par classes d'âge d'après une enquête CDNC / Cambridge Public Library

et ils s'intéressent surtout à la généalogie :

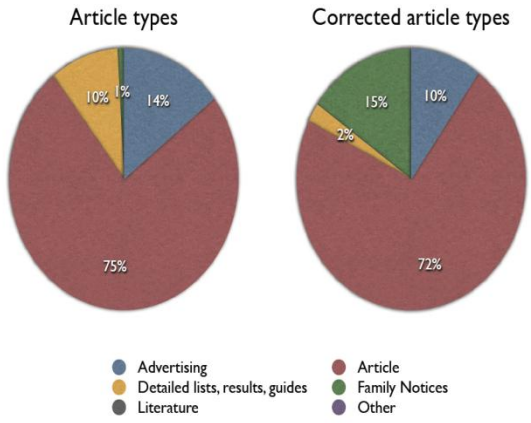


Figure 62. Les types de documents diffusés sur Trove comparés aux types de documents qui y sont corrigés, d'après (Hagon, 2013)

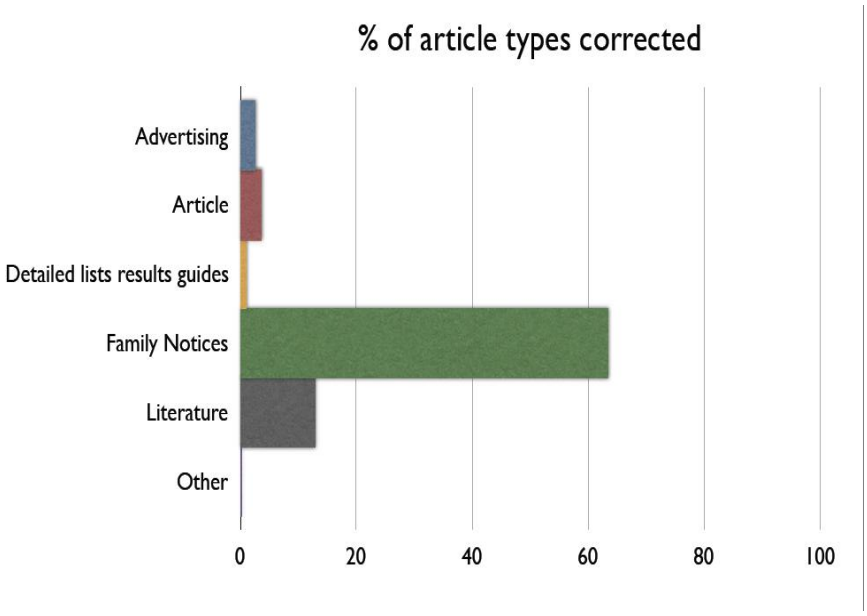


Figure 63. Les types de documents les plus corrigés sur Trove, d'après (Hagon, 2013)

Si la majeure partie des bénévoles se recrute parmi les retraités aisés s'intéressant à la généalogie ou à l'histoire locale, la question se pose de la pérennité de ce type de service. Lorsqu'une autre génération de retraités aura remplacé l'actuelle car rien n'indique qu'elle se passionnera autant pour la généalogie et l'histoire locale. La question de la représentation de l'ensemble de la société que les institutions culturelles sont sensées servir se pose également.

Au delà de la niche du *crowdsourcing* appliqué aux bibliothèques, ce phénomène qui veut que seule une minorité d'internautes soit à l'origine de la plupart des contenus se vérifie également à l'échelle de Wikipédia où 90 % des contributions sont le fait de seulement 10 % des utilisateurs.

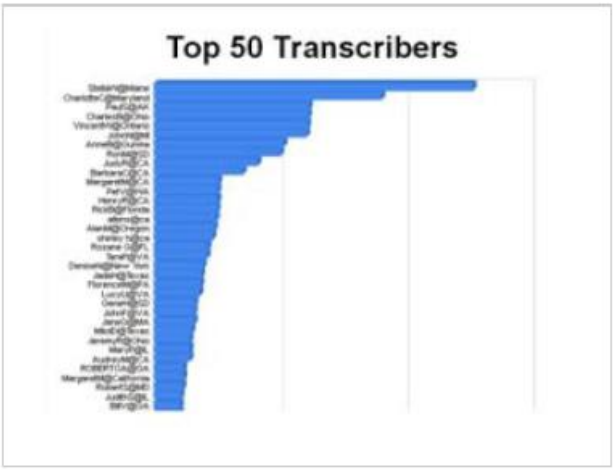


Figure 64. Le top 50 des plus gros contributeurs de Wikipédia

Dans le cas des institutions culturelles, en particulier, il est donc finalement assez difficile de parler véritablement de *crowdsourcing* car il ne s'agit pas véritablement d'une foule d'internautes indifférenciés et anonymes contribuant irrégulièrement ou un nombre limité de fois, mais plutôt de petites communautés de fidèles bénévoles qui s'assistent mutuellement. La plupart des projets de *crowdsourcing* réussis n'ont pas bénéficié de grandes foules d'anonymes mais sont parvenus à provoquer la participation de quelques bénévoles engagés (Owens, 2013). Ainsi, d'après une étude de (Carletti, 2013), sur 36 projets de *crowdsourcing* étudiés, ces contributeurs seraient entre quelques centaines et quelques dizaines de milliers, la moyenne des participants tournant autour de 5000 ou 6000 internautes.

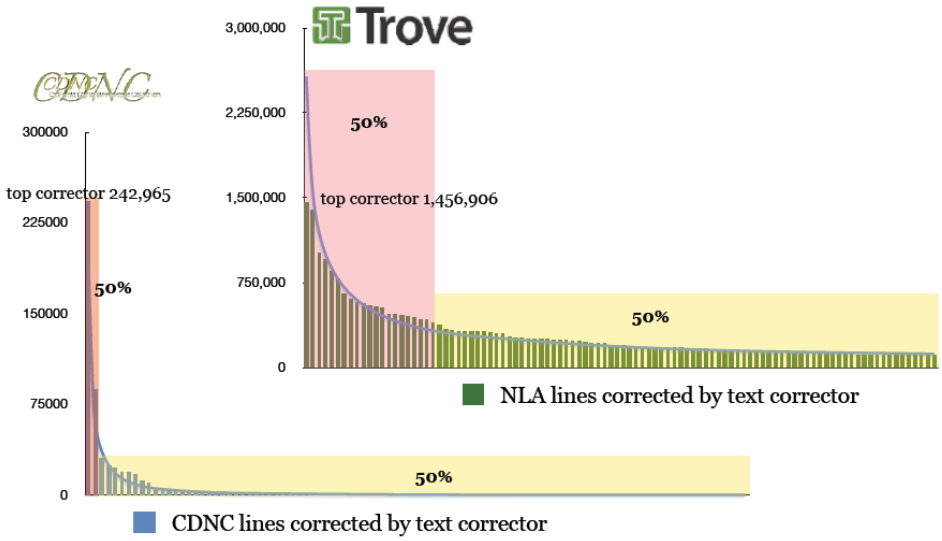


Figure 65. Classement des contributeurs selon le nombre de lignes corrigés pour les projets TROVE et CDNC d'après (Zarndt, 2014)

Voici une manière plus originale d'illustrer ce phénomène :

Figure 66. Part du travail accompli par chaque contributeur du projet Old Weather proposant de transcrire des observations météorologiques (d'après Brumfield, 2013)

Dans le diagramme qui précède, chaque carré correspond à un bénévole et la taille de chacun d'entre eux est proportionnelle à la quantité de transcriptions effectuées.

On constate ainsi que sur 1,6 millions de fiches produites par la British Royal Navy entre 1905 et 1929 et transcrites par environ 16 000 bénévoles, 10 % des transcriptions sont l'œuvre de seulement 20 bénévoles très productifs.

Certains auteurs comme (Causer, 2012) préfèrent ainsi le terme de "*communitysourcing*" à celui de *crowdsourcing* pour exprimer le fait qu'on recherche à mobiliser une communauté plutôt qu'une foule d'internautes indifférenciés.

Enfin, la question de la réintégration des données produites par les internautes se pose. La collaboration entre la BnF et Wikisource illustre les difficultés qui peuvent se poser pour réintégrer les données ainsi produites. Les logiciels de reconnaissance de caractères ont ainsi produit des fichiers ALTO en XML qui se présentent sous la forme d'un index des mots du texte océrisé avec les coordonnées dans l'image de chaque mot. Grâce au fichier ALTO, les recherches en texte intégral sont possibles ainsi que les fonctionnalités de surlignage des mots recherchés. Or, ces fichiers ALTO n'ont pas pu être corrigés et mis à jour par les internautes de Wikisource. Seul le texte a pu être corrigé et la reconstruction des fichiers ALTO, c'est à dire du lien entre l'image et le texte a probablement été difficile pour la BnF. C'est la raison pour laquelle, il vaut mieux que la question de la réintégration des données produites par les bénévoles doit être posée dès la conception du projet si on ne souhaite pas demeurer dans une simple logique de publicité institutionnelle autour d'un sujet à la mode. Dans ce cas, l'interface de correction de l'OCR pourra, par exemple, être directement intégrée à celle de la bibliothèque numérique.

En France, la correction participative de l'OCR suscite un intérêt assez récent. Il a ainsi fallu attendre 2013, bien après le commencement de cette thèse, pour que la Bibliothèque nationale de France publie, dans le cadre du projet Ozalid, une première publication en langue française sur le sujet. Par contre, dans le domaine des archives en particulier, le *crowdsourcing* commence à être

couramment utilisé. Ainsi, aux Archives départementales de l'Ain, environ 500 000 pages ont été indexées de manière participative en 2 ans (Moirez, 2012). Dans le Cantal, ce sont près de 1000 micro-tâches d'indexation qui seraient réalisées chaque jour. Dans le domaine de l'imprimé, le code source du logiciel développé par le projet TROVE (Bibliothèque nationale d'Australie) a été réutilisé, par le projet PlaIR[29], porté par l'Université de Rouen et par les Archives départementales de Seine Maritime afin de proposer des fonctionnalités de correction participative de l'OCR.

Ces éléments sur la correction participative ont fait l'objet d'un article (Andro, 2015, 2)

29 http://plair.univ-rouen.fr (consulté le 23 juin 2016)

2.5- Folksonomie, catalogage et indexation participatives

La folksonomie est un terme inventé en 2004 par Thomas Vander Wal à partir des mots folk and taxonomy. Il est synonyme de collaborative tagging, social tagging, social classification, et social indexing. D'autres auteurs parlent également de "potonomie" ou de "peuplonomie" (Elie, 2007).

Ce sujet a déjà été beaucoup étudié y compris dans la littérature française et il n'aurait guère été opportun de concentrer trop d'effort sur le sujet. Nous limiterons donc le sujet d'étude à quelques projets représentatifs et plus originaux.

2.5.1- Le *crowdsourcing* explicite par tagging volontaire : Flickr: The Commons

Le site Flickr a été créé en 2004. Il rassemble une communauté de 51 millions de photographes enregistrés. D'après Yahoo, 4,5 millions de photographies y seraient mises en ligne chaque jour. (Colquhoun, 2013). Le 16 janvier 2008, 3000 photos ont été placées par la Bibliothèque du Congrès sur Flickr afin d'en accroître la visibilité et d'en permettre l'indexation par les internautes, mais aussi les commentaires, le partage, l'enregistrement dans les favoris et la réutilisation. Le choix de Flickr plutôt que Picassa, Wikimédia ou d'autres s'est principalement justifié du fait de l'existence d'une communauté préexistante importante. En 2013, 56 institutions dans 14 pays, au delà de la Bibliothèque du Congrès participaient également au projet.

Les statistiques suivantes ont été récoltées dans la littérature (Holley, 2010 ; Paraschakis, 2013 ; Earle, 2014) :

Date	**Nombre d'images à indexer**	**Nombre de visiteurs**	**Nombre de tags**	**Nombre de commentaires**	**Nombre d'internautes participants**
24 heures après la mise	3000 photos	1,1 million de vues			

en ligne du 16 janvier 2008					
1 semaine après la mise en ligne	3000 photos	3,6 millions de vues			
Au 23 octobre 2008 (10 mois après la mise en ligne)		10,4 millions de vues de photos et environ 6 millions de visites (soit 500 000 visites par mois)	67 176 tags sur 4615 photographies	7 166 commentaires	2 518 internautes pour les tags 2 562 internautes pour les commentaires
En janvier 2013 (en 5 ans)	250 000 images		2 millions de tags	165 000 commentaires	165 000 contributeurs

Tableau 9. Statistiques collectées dans la littérature à propos du projet Flickr The Commons

D'autres institutions comme la Smithsonian déclaraient qu'en seulement 3 mois, leurs photographies avaient reçu, en moyenne 2348 visites par jour, soit autant de visites qu'en 5 ans, lorsqu'elles étaient auparavant sur le site de l'institution. Pour la Smithsonian, entre juin et octobre 2008, 513 commentaires ont ainsi été ajoutés sur 254 photographies (22 % du corpus) avec une moyenne de 2 commentaires par image et 1 commentaire pour 2089 visites.

Flickr a également été utilisé pour collecter des photographies des internautes par les Bibliothèques et Archives du Canada. "*En plus de cent ans d'existence, les diverses branches de Bibliothèque et Archives Canada (BAC) ont*

collecté plus de 25 millions de photographies. Il aura fallu seulement six ans pour voir le site web Flickr rassembler 5 milliards d'images." (Casemajor Loustau, 2011).

En France, comme le relate (Peccatte, 2009), une initiative sous Flickr a également été conduite sur des archives de photographies de la Normandie pendant la 2ème guerre mondiale. Ainsi, entre 2763 photos ont été versées en janvier 2007 afin d'en permettre la redocumentarisation par des spécialistes sur le web, tout en générant en moyenne 1300 visites par jour en 2007-2008.

Comme pour l'ensemble des projets de *crowdsourcing* appliqué au patrimoine culturel, l'essentiel du contenu est produit par une minorité active d'internautes. Concernant Flickr: The commons, 40 % des tags sont ainsi le fait de seulement 10 "super taggers" ayant ajouté plus de 3 000 tags chacun (Holley, 2010).

Les internautes ont ainsi permis aux institutions qui participent au projet d'identifier des personnes, des lieux, des événements…

Concernant la qualité une étude de (Guy & Tonkin, 2006) et rapportée par Earle estimait, sur un échantillon que seuls 40 % des tags avaient une occurrence dans le dictionnaire Open Source Aspell.

2.5.2- Le recours à la *gamification* : Art Collector

Art Collector est un jeu plus spécifiquement destiné au patrimoine numérisé et inspiré des expérimentations de *gamification* précédemment étudiées. C'est un jeu développé à titre expérimental sur Facebook autour du Swedish Open Cultural Heritage (SOCH) qui agglomère près de 100 000 images collectées sur divers sites web valorisant le patrimoine culturel suédois.

Comme le relève (Paraschakis, 2013), le choix du réseau social Facebook vient du trafic très important généré par ses jeux comme Mafia Wars ou Farmville. L'objectif du jeu est de constituer une collection d'images et de tableaux et de devenir le plus grand collectionneur d'art en additionnant la valeur des œuvres cumulées, la valeur d'une œuvre étant proportionnelle au nombre de tags qui la décrivent.

Il existe 2 types de collections : les collections privées constituées par les joueurs et les collections publiques dont les pièces n'appartiennent encore à personne. Afin de s'approprier une image, le joueur doit avoir proposé plus de la moitié de ses tags. A la fin de ce premier round ("Tag It !") pour lequel 4 images sont proposées à chacun (il reste possible de passer une image), les joueurs sont rémunérés par 4 jetons pour chaque tag original et 2 jetons pour chaque tag déjà existant, c'est à dire commun avec un ou plusieurs joueurs. L'objectif de ce premier round est de proposer un maximum de mots clés. Il est également possible de gagner 40 jetons si l'un des amis de son réseau social accepte de participer. Les 3 meilleurs joueurs reçoivent une médaille (or, argent ou bronze). Un joueur ayant saisi plus de 50 tags obtient la médaille de "Power Tagger" et celui en ayant saisi plus de 100, celui de "Super Tagger".

Figure 54. Capture d'écran du jeu Art Collector 1er round, d'après (Paraschakis, 2013)

Les œuvres qui obtiennent au moins 4 tags sont ensuite achetées par la collection publique pour le second round ("Challenge").

Dans cette partie, les joueurs cherchent à gagner des œuvres appartenant à la collection publique ou à des collections privées, en particulier, de joueurs faisant partie du même réseau social, pour enrichir leurs collections privées en devinant les mots clés qui leurs correspondent. Chaque tentative coûte 20 jetons. L'œuvre est gagnée si plus de la moitié de ses tags ont été devinés.

Figure 55. Capture d'écran du jeu Art Collector Round 2, choix d'une pièce, d'après (Paraschakis, 2013)

Figure 56. Capture d'écran du jeu Art Collector. Round 2, essayer de gagner une œuvre, d'après (Paraschakis, 2013)

Avec la première partie du jeu (premier round), des indexations de documents sont produites. Avec le second round, ces indexations sont validées.

D'après les statistiques qui ont été publiées, 103 utilisateurs se seraient connectés au jeu en 2 semaines. Parmi ces joueurs, 56,3 % d'entre eux ont fait plusieurs parties. Néanmoins, aucun tag n'a été ajouté par 35 % d'entre eux. Sociologiquement, parmi les classes d'âges des joueurs, la plus nombreuses est la classe des 25-34 ans. Les hommes sont 10 % plus nombreux que les femmes.

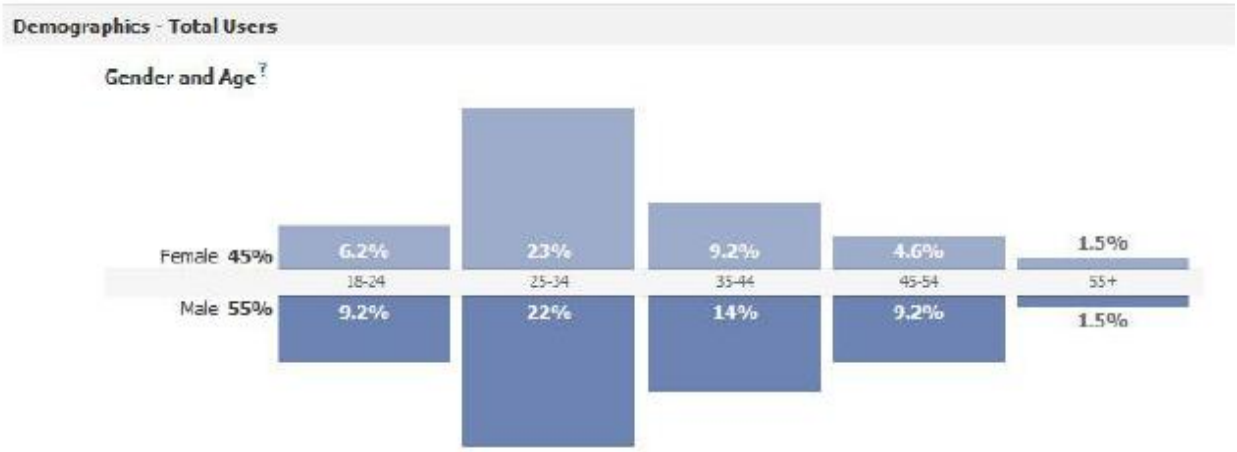


Figure 57. Genre et âge des joueurs de Art Collector d'après (Paraschakis, 2013)

Comme le souligne (Paraschakis, 2014), le jeu Art Collector fait appel à la fois à l'esprit de compétition (tableau des meilleurs joueurs, résultats, challenges), de communication (recherche d'amis, notifications) et de collaboration (partager un trophée).

Pour un panorama plus complet, nous aurions également pu évoquer le projet 1001 Stories Denmark, le wiki du Netherlands Institute for Sound and Vision, le Historical Maps Pilot de la British Library, le projet Mtagger de l'Université du Michigan, le projet PennTags de l'Université de Californie, le projet Social OAC de la Daniel Library, le projet australien "Describe me" du Museum Victoria, les projets "Tag! You're It!" et "Freeze tag!" du Brooklyn Museum, le projet britannique "Your Paintings Tagger" et le projet britannique "Operation war diary" de l'Imperial War Museums et le jeu Tagging Wasida avec lequel les internautes marquent autant de points que leurs tags sont validés par d'autres (Yoshimura & Shein 2011).

Le domaine de la folksonomie est trop important et utilisé depuis bien trop longtemps pour qu'il soit utile et judicieux, dans le cadre de cette thèse, de chercher à étudier l'exhaustivité des publications à son sujet. C'est pourquoi, nous nous sommes limités aux initiatives et aux publications les plus représentatives, significatives et novatrices en insistant particulièrement sur la *gamification*.

En France, selon un rapport d'étape sur la situation des archives concernant l'indexation collaborative (Bouyé, 2012), 20 % des départements seraient engagés dans le web 2.0 et la plupart devraient proposer ce type de fonctionnalités en 2015.

Le tagging est un dialogue entre le visiteur et l'œuvre, entre le visiteur et le musée (Trant, 2006). Si la folksonomie offre une liberté totale à l'usager sur lequel elle est centrée au delà de tout langage contrôlé, contraignant et coûteux, elle peut aussi être détournée par ceux qui cherchent à améliorer le référencement d'une page web et devenir source d'info-pollution et contribuer à générer une "Babel sémantique" (Le Deuff, 2006). Ainsi, selon une étude de (Guy & Tonkin, 2006) rapportée par Earle, seulement 40 % des tags auraient une occurrence dans le dictionnaire Open Source Aspell.

2.6- La traduction participative

La traduction participative des sites web des bibliothèques numériques est évoquée dans la littérature (Budzise-Weaver, 2012). Mais la traduction des textes patrimoniaux numérisés eux-mêmes n'a, à notre connaissance, pas encore fait l'objet de nombreux projets, en dehors du Suda On-Line project qui propose de traduire une encyclopédie byzantine du 10è siècle, du projet européen Organic Lingua (http://www.organic-edunet.eu/en consulté le 23 juin 2016) mais qui concernait des textes nativement électroniques, non des textes patrimoniaux numérisés, et, dans un domaine assez proche, de Wikiaudia, un projet de traduction orale de textes numérisés afin de produire des livres audio (Venkatesh, 2015).

Bien que ce site ne soit pas, non plus, un site en rapport avec la numérisation du patrimoine des bibliothèques, signalons quand même que Luis Von Ahn, l'inventeur du reCAPTCHA et du Google Image Labeller, a ensuite développé un nouveau projet de *crowdsourcing* implicite, Duolingo.com, dont le modèle économique illustre les possibilités du *crowdsourcing* implicite. Les internautes peuvent se former gratuitement aux langues étrangères. En

contrepartie les traductions qu'ils effectuent en se formant sont revendues afin de financer le site.

Pour les bibliothèques numériques, la traduction participative représente un chantier important qui reste à mettre en œuvre dans un cadre mutualisé comme Google Books, Internet Archive, Hathi Trust ou Gallica.

Afin de ne pas alourdir ce deuxième chapitre relatif au panorama des projets, nous n'avons sélectionnés que les projets les plus significatifs dans ce chapitre et placé, en annexe, les données recueillies concernant d'autres projets moins représentatifs.

Chapitre 3- Analyses, du point de vue des sciences de l'information et de la communication, sur le *crowdsourcing* pour les bibliothèques numériques

A partir du panorama des projets et d'une lecture de la littérature sur le sujet, nous avons réalisé une synthèse sur le sujet sous la forme d'un état de l'art. Cette synthèse comporte également des analyses originales non issues de la littérature.

3.1- Typologies et taxonomies des projets

Bien que nous ayons déjà introduit cette thèse par une nécessaire définition introductive du *crowdsourcing*, nous sommes amenés, dans cette nouvelle partie, destinée à des analyses dans le domaine des sciences de l'information et de la communication, à revenir sur cette définition. Cette fois-ci, nous le ferons de manière moins générale et de manière plus appliquée au domaine des bibliothèques numériques et en produisant une taxonomie originale.

La taxonomie, en sciences naturelles, consiste à classer les espèces en fonction de leurs traits et de leurs caractères en classes, ordres, familles et genres. Cette science particulière a inspiré d'autres disciplines et notamment les sciences de l'information et de la communication. Concernant le domaine du *crowdsourcing*, il existe de nombreuses taxonomies proposées dans la littérature que nous allons résumer ici avant de proposer notre propre classification originale de projets de *crowdsourcing* dans le domaine des bibliothèques numériques.

Initialement, John Howe, l'inventeur du terme de "*crowdsourcing*", avait distingué les quatre grands types suivants :

- Intelligence collective : résoudre des problèmes grâce à la sagesse des foules.
- Crowdcreation : utiliser la créativité collective.
- Crowdvoting : demander l'opinion et l'avis des internautes
- *Crowdfunding* : faire appel au financement participatif

Dans la continuité de Howe, Harris reprend sa taxonomie mais distingue plus particulièrement micro-tâches et macro-tâches, ces dernières ayant d'avantage recours à de l'innovation via des internautes ou via la sagesse de foules d'internautes :

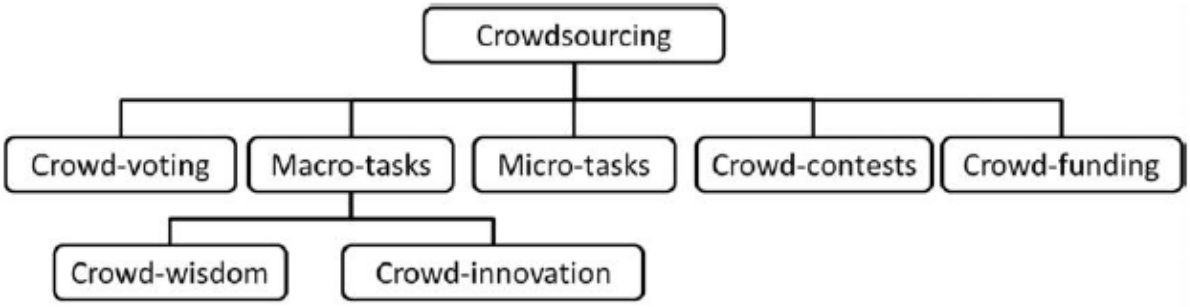


Figure 67. **Taxonomie du *crowdsourcing*, d'après (Harris, 2013)**

La plupart des auteurs ont également cherché à classer les projets en fonction du niveau d'engagement et d'initiative, distinguant ainsi l'engagement, la participation, la contribution, et le volontariat ou, plus simplement, distinguant le *crowdsourcing* participatif ou contributif et le *crowdsourcing* collaboratif (Bonney, 2009 ; Oomen, 2011 ; Dunn, 2012 ; Tweddle 2012 ; Boeuf 2012 ; Radice, 2014) :

- *crowdsourcing* participatif ou *crowdsourcing* contributif : le public contribue en produisant simplement des données dans le cadre de projets conçus et pilotés par des institutionnels. Le travail du public est déterminé, limité et fait appel à un investissement individuel relativement faible (microtâches).
- *crowdsourcing* collaboratif ou de co-création : le public prend part plus activement aux décisions de la politique documentaire du projet et faisant appel à un investissement individuel plus fort (macrotâches). Certains auteurs (Bonney, 2009 et Radice, 2014) distinguent parfois *crowdsourcing* collaboratif et co-création :
  - *crowdsourcing* collaboratif : partenaires actifs et qui interagissent entre eux, mais dans un cadre contrôlé par l'institution.

- Co-création : partenaires qui participent également à la politique et à la définition des objectifs du projet ou qui peuvent même être à l'initiative des projets.

De manière plus précise, (Bonney, 2009) propose le tableau suivant que nous avons adapté à notre domaine et traduit en français :

Étapes scientifiques	**Projets contributifs**	**Projets collaboratifs**	**Projets de co-création**
Choisir et définir une question	Non	Non	Oui
Rassembler des informations et des ressources	Non	Non	Oui
Faire des analyses et des hypothèses	Non	Non	Oui
Concevoir des méthodologies de collectes de données	Non	Éventuellement	Oui
Produire des données	Oui	Oui	Oui
Analyser des données	Éventuellement	Oui	Oui
Interpréter des données et en tirer des conclusions	Non	Éventuellement	Oui
Diffuser les conclusions	Éventuellement	Éventuellement	Oui
Discuter les	Non	Non	Oui

résultats et définir de nouvelles questions			

Tableau 12. Modèle de participations du public inspiré de (Bonney, 2009)

Parmi les tâches identifiées dans notre panorama (mise en ligne, numérisation et impression à la demande, correction de l'OCR, transcription, indexation), il n'existe manifestement à ce jour que des projets contributifs en bibliothèques. Seule la numérisation à la demande et la mise en ligne pourraient être qualifiées de collaboratives dans la mesure où le public participe ainsi à la constitution de la collection, donc à la politique d'acquisition et à la politique documentaire de la bibliothèque numérique.

Selon (Simon, 2010) et (Stiller, 2014), il existerait 5 étapes successives d'engagement :

1- des individus qui consomment du contenu

2- des individus qui interagissent avec du contenu

3- les interactions individuelles sont mises en réseau ensemble

4- les interactions individuelles sont mises en réseaux sociaux

5- les individus s'engagent socialement les uns pour les autres

Il est également possible d'affiner cette distinction entre participatif et collaboratif en fonction des types de participation dans la mesure où, sur le web, on trouve déjà les catégories suivantes selon l'étude Forrester's NACTAS Q4 2006 Devices & Access Online Survey rapportée par (Radice, 2014) :

- Les "créateurs" (13 %) qui publient des sites web ou des blogs, mettent en ligne des vidéos
- Les "critiques" (19 %) qui commentent et évaluent
- Les "collectionneurs" (15 %) qui font du partage sur les réseaux sociaux
- Les "sociables" (19 %) qui utilisent les réseaux sociaux
- Les "spectateurs" (33 %) qui lisent des contenus sur Internet

- Les “inactifs” (52 %) qui ne rentrent dans aucune des catégories précédentes

En s’inspirant de cette classification générale et en l’appliquant au domaine culturel, on obtiendrait les catégories suivantes des conservateurs (minoritaires), des producteurs, des commentateurs, des partageurs de contenus, et des consommateurs (majorité), d’après (Radice, 2014).

On pourrait également classer les projets en fonction des critères suivants :

- Qui contribue ? Une foule indéterminée et ouverte d'internautes (*crowdsourcing*) ou un groupe plus spécifique et déterminé, une communauté (*communitysourcing*), des populations locales ?
- Pourquoi la foule contribue ? Pour quel type de motivations ? Des motivations intrinsèques ou plutôt extrinsèques ?
- Comment la foule contribue ? Par compétition ou au contraire, par collaboration ?
- Pour qui la foule contribue-t-elle ? Pour des intérêts privés ou pour des intérêt publics ?
- Quel est l’objectif principal du projet ? Obtenir des données ou mobiliser la foule autour des collections afin de mieux la sensibiliser.
- Qu’apportent les contributeurs ? De l’argent ? (*crowdfunding*) Du travail ? Des connaissances ? Des idées ?

Il est également possible de classer les projets selon le niveau d’interaction et de compétition de la foule (Renault, 2014 bis) en distinguant :

- Le *crowdsourcing* cumulatif : la juxtaposition et l’agrégation de participations individuelles susceptibles de découvertes inattendues (“les petits ruisseaux font les grandes rivières”[30])
- Le *crowdsourcing* collaboratif : la collaboration des individus à orchestrer (“l’union fait la force”)
- Le *crowdsourcing* compétitif : la compétition à arbitrer (“que le meilleur gagne”)

[30] "Vous voyez peu de fleuves larges dès leur source ; la plupart se grossissent des ruisseaux qui se jettent dans leur sein" (Ovide, Œuvres complètes)

- Le *crowdsourcing* coopétitif : la coopération dans un esprit compétitif (“tous pour un, un contre tous”)

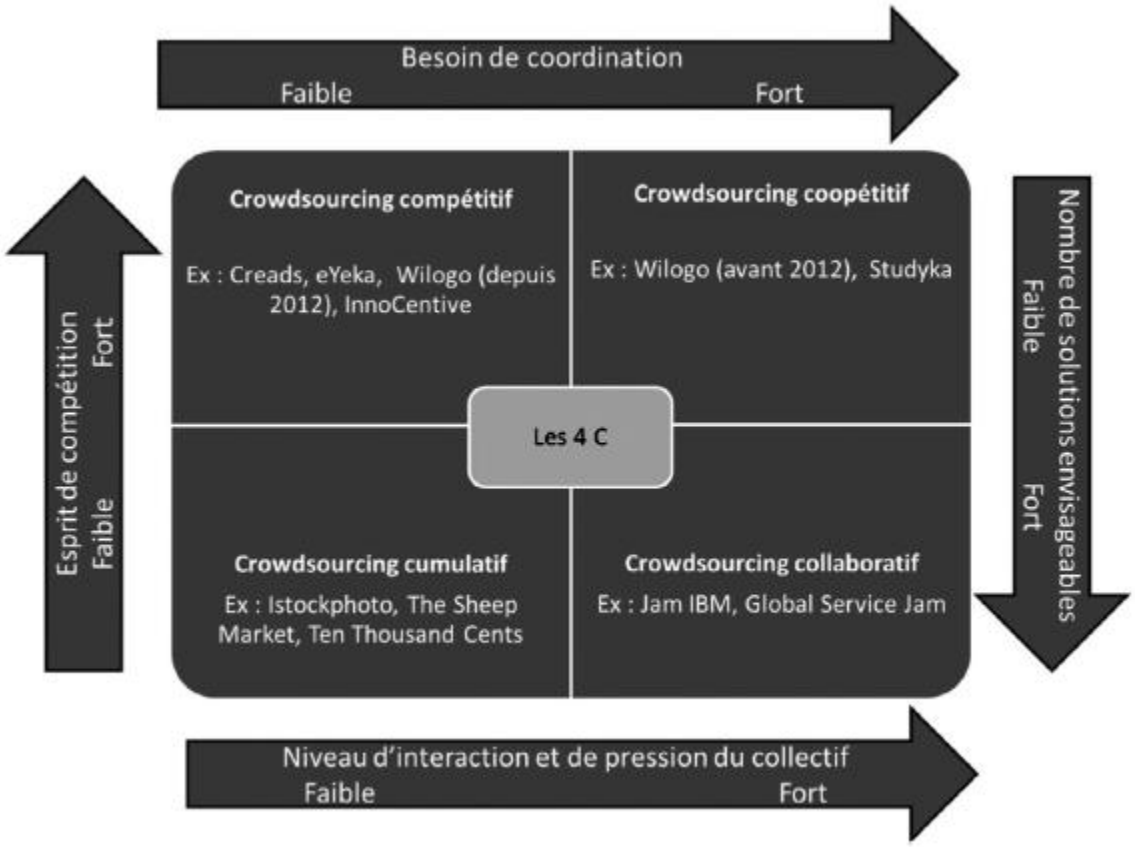


Figure 68. La taxonomie des 4C du *crowdsourcing* d’après (Renault, 2014 bis)

Enfin, il est évidemment également possible de classer les projets en fonction du type d’activité proposé, comme nous l’avons d’ailleurs fait, en partie, dans notre panorama des projets, distinguant mise en ligne participative, numérisation à la demande par *crowdfunding*, correction participative de l’OCR et folksonomie.

Dans le cas de la numérisation des bibliothèques, on peut ainsi identifier les activités suivantes :

- Sélection des documents susceptibles d’être numérisés selon des critères juridiques (auteur mort depuis plus de 70 ans), de pertinence (document pas déjà numérisé ailleurs), ou des critères scientifiques et thématiques.

- Description matérielles des documents à numériser (format, nombre de feuillets, angle d'ouverture, état du document)
- Numérisation (organisation des trains de numérisation, numérisation, expédition des trains)
- Production d'un OCR brut (avec des logiciels de reconnaissance optique de caractères)
- Contrôle qualité de la numérisation (travail ingrat généralement pris en charge en interne par les bibliothèques qui contrôlent certains points sur l'ensemble des livraisons ou sur des échantillons. Ce travail peut parfois être sous-traité par un prestataire qui contrôle le travail d'un premier prestataire)
- Mise en ligne des documents
- Catalogage ou reprise de catalogage des documents mis en ligne
- Indexation ou reprise d'indexation des documents
- Correction de l'OCR (généralement réalisée pour des projets éditoriaux, la production de fichiers EPUB ou MOBI ou des projets de *text mining* avec l'aide de prestataire faisant appel à de la main d'œuvre à bas coût à Madagascar, en Inde par exemples.)
- Archivage pérenne (transfert sur des serveurs d'archivage pérenne de fichiers haute résolution dans des formats de conservation accompagnés de métadonnées techniques et bibliographiques)
- Valorisation éditoriale et contextualisation en ajoutant un appareil critique pour chaque texte (Informations bibliographiques, résumés, tables des matières, actualités, analyses, documents en relation...)
- Création de livres électroniques lisibles sur tablettes aux formats EPUB et/ou MOBI

Nous aurions également pu ajouter les activités suivantes : créer des liens, commenter, catégoriser, cataloguer, contextualiser, géoréférencer et traduire en reprenant la classification proposée par (Dunn, 2012). D'autres classifications, selon le type de documents (images, textes, manuscrits, vidéos, sons, cartes) ou encore selon le type de données produites par les contributeurs (textes traduits ou

transcrits ou corrigés, métadonnées, résumés, connaissances, argent...) seraient également envisageables.

Dans le cadre de notre diaporama de projets, nous avons également été amené à distinguer de manière plus originale, des types de *crowdsourcing* eux mêmes

- le *crowdsourcing* explicite
  - le *crowdsourcing* explicite gratuit (recours aux internautes bénévoles)
  - le *crowdsourcing* explicite rémunéré (recours aux internautes rémunérés)
- le *crowdsourcing* implicite (recours au travail involontaire des internautes)
- la *gamification* (recours au travail des internautes sous la forme de jeux)
- le *crowdfunding* (recours aux contributions financières des internautes

Cette taxonomie est originale. Au 27 mars 2015, la notion de "*crowdsourcing* implicite" pourtant présente dans la littérature internationale sous la forme de "*implicit crowdsourcing*" ne donne que 6 résultats sous la forme française de "crowdsourcing implicite" dans Google. Cette distinction s'inspire de (Harris, 2013) qui, concernant la participation volontaire, parle de *crowdsourcing* explicite et, concernant la participation involontaire, de *crowdsourcing* implicite.

A partir de cette taxonomie issue d'une analyse de la littérature et des taxonomies produites précédemment, nous avons recherché à croiser ces diverses formes de modèles avec les différentes activités d'un projet de développement d'une bibliothèque numérique. En les croisant, nous avons cherché à identifier d'éventuelles formes de *crowdsourcing* qui resteraient à inventer et qui n'auraient, à notre connaissance, encore jamais fait l'objet d'expérimentations.

Sur l'axe vertical du tableau, on trouve, les différentes tâches d'un projet de numérisation :

- la sélection des documents qui méritent d'être numérisés sur des critères scientifiques, historiques et après vérifications qu'ils ne l'ont pas déjà été et qu'ils peuvent l'être d'un point de vue juridique ;
- la numérisations ;

- le financement ;
- le contrôle qualité de la numérisation, de l'OCR, des métadonnées ;
- le catalogage ;
- l'indexation ;
- la correction de l'OCR d'imprimés ;
- la transcription de manuscrits.

Sur l'axe horizontal, on trouve notre taxonomie :

- *crowdsourcing* explicite ;
- *gamification* ;
- *crowdsourcing* implicite.

Et pour chaque catégorie, on a créé des sous catégories pour distinguer le travail bénévole et gratuit du travail rémunéré bien qu'il existe une multitude de formes intermédiaires. Nous avons également choisi de distinguer les degrés quantitatifs d'engagement :

- participatif : les internautes produisent des données sous la forme de microtâches et avec un engagement relativement faible pour les institutions dans un cadre limité
- collaboratif : les internautes participent à la politique et à la définition des objectifs du projet et s'engagent plus fortement

Ainsi, si on croise les types d’activités liées aux projets de numérisation avec toutes ces variables, on obtient la matrice taxonomique originale ci-après. Cette matrice a l’intérêt de permettre d’identifier des formes nouvelles de crowdsourcing appliqué aux projets de numérisation qui resteraient à inventer. Sa limite réside dans le caractère artificiel et parfois manquant de sens de certains croisements. Ainsi, toutes les formes en rouge ne nous semblent pas pouvoir trouver une application et elles demeurent majoritaires dans le tableau.

	crowdsourcing explicite				*Gamification*				*crowdsourcing* implicite			
	gratuit		rémunéré		gratuite		rémunérée		gratuit		rémunéré	
	participatif	collaboratif	participatif	collaboratif	participatif	collaboratif	participatif	collaboratif	participatif	collaboratif	participatif	collaboratif
Sélection	111a	112a	121a	122a	211a	212a	221a	222a	311a	312a	321a	322a
Numérisation	111b	112b	121b	122b	211b	212b	221b	222b	311b	312b	321b	322b
Financement	111c	112c	121c	122c	211c	212c	221c	222c	311c	312c	321c	322c
Contrôle qualité	111d	112d	121d	122d	211d	212d	221d	222d	311d	312d	321d	322d
Catalogage	111e	112e	121e	122e	211e	212e	221e	222e	311e	312e	321e	322e
Indexation	111f	112f	121f	122f	211f	212f	221f	222f	311f	312f	321f	322f
Correction de l'OCR	112g	112g	121g	122g	211g	212g	221g	222g	311g	312g	321g	322g
Transcription	111h	112h	121h	122h	211h	212h	221h	222h	311h	312h	321h	322h

Tableau 13. Activités d'un projet de numérisation croisées avec les types de *crowdsourcing*

Parmi toute cette systématique des formes de *crowdsourcing* susceptibles d'exister dans le domaine de la numérisation en bibliothèque, certaines, en rouge, n'ont, de notre point de vue, malheureusement pas beaucoup de sens et ne permettraient pas de trouver un jour une application.
D'autres formes de *crowdsourcing*, en vert, existent déjà comme :

Code	Type	Forme
111a	le *crowdsourcing* explicite gratuit et participatif appliqué à la sélection des documents	suggérer la numérisation d'un document
311a	le *crowdsourcing* implicite gratuit et participatif appliqué à la sélection des documents	utiliser les statistiques de consultation et d'emprunts des catalogues de bibliothèques pour identifier les documents à numériser
111b	le *crowdsourcing* explicite gratuit et participatif appliqué à la numérisation	avoir recours aux numérisations amateurs de livres ou d'archives par les internautes pour enrichir des bibliothèques numériques (Internet Archive, par exemple)
111c	le *crowdfunding* appliqué à la numérisation	la numérisation à la demande de documents par financements participatifs (Numalire, par exemple)
111f	le *crowdsourcing* explicite gratuit et participatif appliqué à l'indexation	le tagging (folksonomie, Steve Museum, par exemple)
211f	la *gamification* gratuite et participative appliquée à l'indexation	les jeux autour de l'indexation des documents numérisés (Google Image Labeler par exemple)
111g	le *crowdsourcing* explicite gratuit et participatif appliqué à la correction de l'OCR	la correction participative de l'OCR (Wikisource, par exemple)

121g	le *crowdsourcing* explicite rémunéré et participatif appliqué à la correction de l'OCR	la correction rémunérée de l'OCR par des internautes (sur Amazon Mechanical Turk Marketplace, par exemple)
211g	la *gamification* gratuite et participative appliquée à la correction de l'OCR	la *gamification* autour de la correction de l'OCR (Digitalkoot, par exemple)
311g	le *crowdsourcing* implicite gratuit et participatif appliqué à la correction de l'OCR	l'utilisation des saisies des internautes pour des raisons de sécurité (reCAPTCHA par exemple)
111h	le *crowdsourcing* explicite gratuit et participatif appliqué à la transcription de manuscrits	la transcription participative de manuscrits (Transcribe Bentham, par exemple)
121h	le *crowdsourcing* explicite rémunéré et participatif appliqué à la transcription de manuscrits	la transcription rémunérée de manuscrits (sur Amazon Mechanical Turk Marketplace, par exemple)

Tableau 14. Types existants de *crowdsourcing* appliqués à la numérisation

D’autres enfin, en orange, n’ont pas été identifiés ou restent encore à inventer :

Code	Type	Forme
112a	le *crowdsourcing* explicite gratuit et collaboratif appliqué à la sélection des documents	permettre aux internautes d’influer directement sur la politique de numérisation
121a	le *crowdsourcing* explicite rémunéré et participatif appliqué à la sélection des documents	payer des internautes pour qu’ils identifient les documents qui méritent d’être numérisés en retrouvant les dates de décès des auteurs et en vérifiant s’ils n’ont pas déjà été numérisés
211a	la *gamification* gratuite et participative appliquée à la sélection des documents	faire un jeu dans lequel les internautes ont à trouver si le document peut être numérisé en attribuant une note sur son intérêt, en retrouvant les dates de décès des auteurs, en vérifiant s’il n’a pas déjà été numérisé
221a	la *gamification* rémunérée et participative appliquée à la sélection des documents	rémunérer les meilleurs joueurs du jeu précédemment évoqué
121b	le *crowdsourcing* explicite rémunéré et participatif appliqué à la numérisation	rémunérer les internautes ou les lecteurs pour les documents inédits qu’ils numériseraient et mettraient en ligne sur une bibliothèque numérique
311b	le *crowdsourcing* implicite gratuit et participatif appliqué à la numérisation	conserver automatiquement les images des photocopies réalisées par les internautes systématiquement associées aux références du document identifié via RFID afin de

		pouvoir les verser ensuite dans les bibliothèques numériques
111d	le *crowdsourcing* explicite gratuit et participatif appliqué au contrôle qualité de la numérisation	demander aux internautes de valider la qualité de telle ou telle page numérisée sur tel ou tel point de contrôle et confronter leurs validations
121d	le *crowdsourcing* explicite rémunéré et participatif appliqué au contrôle qualité de la numérisation	rémunérer les internautes pour ce travail de validation
211d	la *gamification* gratuite et participative appliquée au contrôle qualité de la numérisation	faire un jeu dans lequel les internautes valident la qualité des pages numérisées sur tel ou tel critère
221d	la *gamification* rémunérée et participative appliquée au contrôle qualité de la numérisation	rémunérer les meilleurs joueurs du jeu précédemment évoqué
111e	le *crowdsourcing* explicite gratuit et participatif appliqué au catalogage des documents numérisés	demander aux internautes de cataloguer les documents numérisés
121e	le *crowdsourcing* explicite rémunéré et participatif appliqué au catalogage des documents numérisés	rémunérer les internautes pour ce travail de catalogage
211e	la *gamification* gratuite et participative appliquée au catalogage des documents numérisés	faire un jeu de catalogage des documents numérisés
221e	la *gamification* rémunérée et participative appliquée au catalogage des documents numérisés	rémunérer les meilleurs joueurs du jeu précédemment évoqué
121f	le *crowdsourcing* explicite rémunéré et participatif appliqué à l'indexation	rémunérer les internautes pour leurs mots clés et leurs tags
221f	la *gamification* rémunérée et	rémunérer les meilleurs joueurs des

	participative appliquée à l'indexation	jeux de tagging
221g	la *gamification* rémunérée et participative appliquée à la correction de l'OCR	rémunérer les meilleurs joueurs des jeux de correction de l'OCR
211h	la *gamification* gratuite et participative appliquée à la transcription de manuscrits	faire un jeu de transcription de manuscrits sur le modèle de ceux qui existent déjà pour la correction de l'OCR
221h	la *gamification* rémunérée et participative appliquée à la transcription de manuscrits	rémunérer les meilleurs joueurs des jeux de transcriptions de manuscrits
311h	le *crowdsourcing* implicite gratuit et participatif appliqué à la transcription de manuscrits	utiliser le système de reCAPTCHA pour les manuscrits

Tableau 15 des types restant à inventer de *crowdsourcing* appliqués à la numérisation

En résumé de ce qui précède, voici la taxonomie originale que nous proposons :

Types de *crowdsourcing*		**Définition**	**Exemples**
crowdsourcing explicite	*crowdsourcing* explicite gratuit	*recours au travail volontaire des internautes bénévoles*	TROVE
	crowdsourcing explicite rémunéré	*recours au travail volontaire des internautes rémunérés*	Amazon Mechanical Turk Marketplace
crowdsourcing implicite		*recours au travail involontaire des internautes*	reCAPTCHA
Gamification "*human computation*" "*games with a purpose*"		*recours au travail des internautes sous la forme de jeux*	Digitalkoot
Crowdfunding		*recours aux contributions financières des internautes*	Numalire

Tableau 16. Taxonomie du *crowdsourcing* appliqué à la numérisation

3.1.1- Le *crowdsourcing* explicite

3.1.1.1- Le *crowdsourcing* explicite gratuit

Cette forme de *crowdsourcing* est la plus ancienne et la plus répandue. Elle consiste à faire appel au travail bénévole et gratuit des internautes pour ajouter des documents numérisés dans une bibliothèque numérique, en corriger l'OCR ou en transcrire l'écriture, y ajouter des métadonnées et des mots clés.

3.1.1.2- Le *crowdsourcing* explicite rémunéré

Cette forme de *crowdsourcing* encore peu répandue en bibliothèque consiste à demander aux internautes de réaliser le même type de travail, mais en étant

rémunéré. Les rares expérimentations relatée dans la littérature, et que nous avons rapportées dans notre panorama, ont été effectué sur l'Amazon Mechanical Turk Marketplace ou CrowdFlower.

3.1.2- Le *crowdsourcing* implicite

Cette forme de *crowdsourcing* encore moins répandue en bibliothèque n'est, à notre connaissance utilisé que par le projet reCAPTCHA qui permet de faire corriger involontairement par les internautes l'OCR des 30 millions de livres numérisés par Google Books lorsqu'ils saisissent des mots déformés pour prouver qu'ils ne sont pas des robots au moment de la création de comptes.

3.1.3- La *gamification*

Cette forme de *crowdsourcing* consiste à demander aux internautes de produire un travail en jouant. Comme nous l'avons vu dans le diaporama des projets, il existe de multiples expérimentations de *gamification* appliquée à la numérisation des bibliothèques.

Si on considère que de nombreuses tâches demeurent encore impossibles à effectuer pour les ordinateurs alors qu'elles le sont pour les humains, et que ces derniers consacrent une partie croissante de leur temps à jouer devant un ordinateur, il apparaît bien qu'il y a une opportunité à utiliser l'intelligence humaine pour toutes sortes de tâches coûteuses et à la mobiliser sous la forme de jeux.

Le jeu en ligne, se situerait, d'après (Paraschakis, 2013) juste derrière les réseaux sociaux parmi les activités les plus répandues sur le web. Des jeux comme Yahoo! Games, MSN's The Zone ou Pogo.com rassemblent fréquemment plus de 100 000 visiteurs. D'après (Paraschakis, 2013), les jeux sur les réseaux sociaux attireraient 120 millions de personnes parmi lesquelles 81 millions qui joueraient tous les jours et 49 millions plusieurs fois par jour. Le jeu Facebook Farmville, en particulier, attirerait 83 millions de joueurs par mois et celui appelé Mafia Wars en attirerait, quant à lui, 25 millions par mois. Facebook resterait leader avec 91 % de joueurs, devant Google+ (17 %), MySpace (15 %) et Bebo (7 %). D'après (Von Ahn, 2008) reprenant un rapport de l'Entertainment Software

Association, 200 millions d'heures cumulées sont consacrées chaque jour aux jeux vidéos aux Etats-Unis, 65 % des ménages américains jouent aux jeux vidéos et un citoyen américain aurait déjà joué, en moyenne, lorsqu'il a atteint l'âge de 21 ans, pas moins de 10 000 heures à des jeux vidéos. Ces 10 000 heures représentent quand même l'équivalent de 5 années de travail à temps plein, soit 40 heures par semaine. Au delà d'un demi-milliard de personnes jouent, dans le monde, à des jeux sur le web et ce, pendant au minimum une heure tous les jours. Aux USA, ils seraient 183 millions (Eickhoff, 2012). Et concernant les *casual games* ou jeux occasionnels de type puzzle, solitaires, réussites, ou encore démineurs, ce sont 200 millions de personnes dans le monde qui y joueraient (Ridge, 2011). Une enquête de 2006 de la société PopCap, rapportée par cet auteur, révélerait que 76 % des joueurs seraient des femmes dont la moyenne d'âge serait de 48 ans.

Dans un contexte de ludification de la culture, où le plaisir semble prendre une importance croissante dans la société, les études, comme le travail pourraient être considérés comme des successions de challenges, avec des épreuves, des quêtes, des changements de niveaux, des points, des bonus. Les organisations pourraient donc ainsi s'inspirer des jeux vidéos pour stimuler la motivation de leurs étudiants ou de leurs collaborateurs. Il serait ainsi possible de réutiliser les principaux ressorts et mécaniques du jeu vidéo pour le réutiliser dans d'autres contextes. Le jeu est une activité volontaire, autonome, permettant d'avoir de nouvelles expériences, de tester et de monter en compétences dans un environnement sûr et sans possibles conséquences mauvaises. D'après (Chrons, 2011), la *gamification* permettrait d'obtenir de meilleurs résultats que le *crowdsourcing* traditionnel en termes de participation. En effet, les individus ont généralement des réticences à consacrer une part importante de leur temps à accomplir des travaux difficiles ou à remplir des tâches ingrates. Mais ils ont aussi parfois des difficultés à s'arrêter de jouer sur les jeux vidéos. Il peut donc être opportun de transformer des tâches ingrates en jeux vidéos. Le processus qui consiste à convertir des activités productives en jeux est appelé *gamification* et pourrait être traduit en français par ludification ou encore par ludicisation. La *gamification* pourrait également être définie comme le fait d'appliquer des

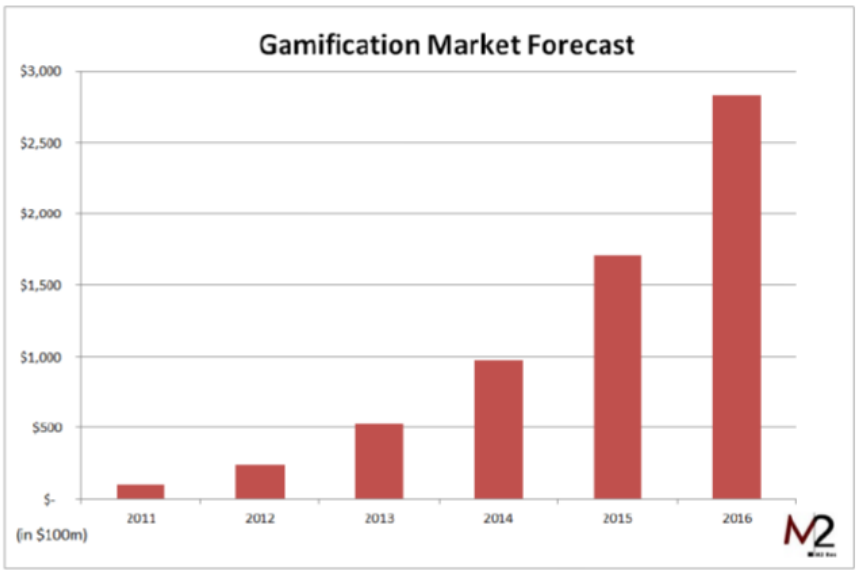
éléments de design, de psychologie et de mécanismes du jeu vidéo dans d'autres contextes (Deterding, 2011).

Le terme de "*gamification*" a été proposé par Nick Pelling en 2002, celui de "*Human Computation*" par Luis Von Ahn, dans sa thèse, en 2005. Il pourrait être traduit par "calcul humanoïde" (Néroulidis, 2015). Le terme de "*Games with a purpose* (GWAP)" a, quant à lui, été proposé en 2008 par Von Ahn & Dabbish. Il pourrait être traduit en français par "jeu avec une finalité". Comme le suggère (Von Ahn, 2006), il suffirait de considérer les cerveaux humains comme autant de processeurs en réseau au sein d'un système distribué. Grâce à ce système, chacun individu pourrait participer à produire un calcul massif. (Quinn, 2011) a produit une contribution spécifique afin de définir la notion de *Human Computation* en reprenant ces différentes définitions. En s'inspirant de ses travaux, nous pourrions définir l'*Human Computation* comme l'utilisation de l'intelligence humaine collective mobilisée, par des jeux, afin de résoudre des problèmes que les ordinateurs ne pas encore en capacité de prendre en charge ou des problèmes qui ne peuvent être résolus par des groupes trop restreints d'humains. De la même manière que le *crowdsourcing* remplace des salariés par des internautes, l'*Human Computation* remplace des ordinateurs par des humains.

Le potentiel de la *gamification* est très important. (Von Ahn, 2004) prétendait que l'intégralité des images de Google Images pourrait être indexée en 31 jours par 5000 internautes qui joueraient 24 heures sur 24 au jeu ESP Game. Il signalait également que 1000 joueurs pourraient indexer 12 000 images par jour si chacun y consacrait une heure de sa journée tandis qu'il faudrait qu'un employé classique tague 900 images par jour pendant plus de 125 jours ou près de 4 mois de travail à plein temps pour obtenir le même résultat au risque de faire un "burn out". Dans le domaine des sciences participatives, si les centaines de millions de joueurs de jeux vidéos dans le monde qui passent 3 billions d'heures par semaine à jouer consacraient seulement 1 % de ce temps au jeu fold.it, les importants résultats obtenus en 3 ans de projet pourraient l'être chaque semaine (Good, 2011).

Plus récemment et d'un point de vue plus large, d'après un communiqué de presse de la société Gartner publié en 2011, plus de 50 % des organisations en

gestion des processus d'innovation pourraient incorporer des mécanismes autour de la *gamification* dans leurs entreprises d'ici à 2015. Nous constatons toutefois que de la même manière que la *gamification* est régulièrement confondue avec les *serious games* qui visent uniquement à se former individuellement par le jeu et non à produire des données, elle est aussi régulièrement confondue avec la « pointification ». Or, attribuer des points pour toutes sortes d'actions n'a rien à voir avec la *gamification*. Les jeux nécessitent une durée et un espace, ils sont présidés par des règles, sont soumis à des finalités, et ont recours à des personnes volontaires.



(in millions)	2011	2012	2013	2014	2015	2016
Total	$100	$242	$522	$980	$1,707	$2,830

Figure 69. Evolution temporelle depuis 2011 et prévision du marché futur de la *gamification* (d'après Ollikainen, 2013)

La *gamification* a déjà trouvé de nombreuses applications dans de multiples domaines comme l'indexation de vidéos ou d'images, la traduction, la transcription, le résumé de documents, l'enseignement, et même la vidéosurveillance.

Son potentiel dans le domaine de la numérisation des bibliothèques pourrait être d'autant plus importants que le score des participants serait affiché. Un projet français de plateforme de *gamification* est en train de voir le jour sous le nom de

fungears.com. Dans le domaine des institutions culturelles, on peut citer DigiTalkoot (National Library of Finland) pour la correction de l'OCR, Alum Tag (Rauner Special Collections Library, Dartmouth College) pour l'indexation de photographies, Tag! You're it! (Brooklyn Museum) pour l'indexation des objets ou encore Waisda? (Netherlands Institute for Sound and Vision) pour l'annotation d'émissions télévisuelles.

A la différence du *crowdsourcing* explicite classique qui s'adresse à des sentiments altruistes, la *gamification* fait plutôt appel à des motivations ludiques des internautes. (Harris, C. G., 2013) considère la *gamification* comme étant à l'intersection entre *serious games* et *crowdsourcing*. A l'instar du *serious game*, la *gamification* peut aussi être très « sérieuse », les données produites en s'amusant peuvent avoir un usage très sérieux et être utilisées par des organisations très sérieuses. Toutefois, la *gamification* se distingue du *serious game* car sa finalité est utilitaire pour l'usager qui en attend un bénéfice individuel en terme de développement personnel, de connaissances et de formation alors qu'avec la *gamification*, il cherche principalement à s'amuser tout en réalisant un objectif extérieur à lui. Ainsi, l'objectif de l'internaute n'est pas, contrairement au *serious game*, de se former grâce au jeu, il est plutôt de produire des données utiles en s'amusant. Enfin, la *gamification*, à la différence du *serious game*, fonctionne à partir de microtâches autonomes les unes par rapport aux autres et n'offre pas un scénario construit de manière linéaire comme c'est généralement le cas avec les *serious games*.

Figure 70. *Serious games* et *gamification* d'après (Deterding, 2011)

D'après (Von Ahn, 2008), il existerait 3 grands types de *gamification* :

- Output-agreement games (ou jeux basés sur l'accord en sortie) : chaque joueur dispose des mêmes informations en entrée (par exemple, une même image pour ESP Game) et les résultats produits peuvent être différents.
- Input-agreement games (ou jeux basés sur l'accord en entrée) : on confronte les saisies des joueurs comme pour TagATune (jeu qui envoie des musiques à 2 internautes qui doivent échanger par écrit pour savoir s'il s'agit ou non du même morceau).
- Inversion-problem games (ou jeux basés sur l'inversion de problèmes) : le premier joueur a accès à l'ensemble du problème et essaie de faire deviner la solution au deuxième joueur comme pour Peekaboom.

A sa suite, (Quinn, 2011) a cherché à proposer une classification des projets de *human computation* en fonction des caractéristiques élargies suivantes :

- le type de motivations des internautes :
  - Argent ou gratification
  - Altruisme

- Distraction (*gamification*)
- Réputation
- Travail implicite (avec ReCaptcha, par exemple, l'internaute ignore qu'il travaille pour Google Books)

- le type de contrôle qualité :
  - Accord en sortie entre les contributions de plusieurs internautes qui travaillent indépendamment et simultanément
  - Accord en entrée
  - Incitations financières
  - Conception des tâches de sorte qu'il n'est pas plus facile de tricher que de faire réellement la tâche
  - Réputation avec évaluation des contributions des internautes à la manière des vendeurs ebay
  - Redondance (identifier les mauvaises contributions et les mauvais contributeurs via un système de vote)
  - Pièges (insérer des erreurs volontaires pour vérifier que le travail est bien effectué)
  - Filtrage statistique
  - Examen à plusieurs niveaux (une groupe d'internautes vérifie le travail d'un premier groupe. Dans les marchés de numérisation, il arrive par exemple qu'un prestataire fasse le contrôle qualité d'une premier prestataire)
  - Contrôle par un expert
  - Contrôle automatique par des programmes et des algorithmes
- le type de compétences humaines mobilisées :
  - Reconnaissance visuelle
  - Compréhension de la langue
  - Communication humaine de base
- L'ordre du processus :
  - Ordinateur puis travailleur puis commanditaire
  - Travailleur puis commanditaire puis ordinateur
  - Ordinateur puis travailleur puis commanditaire puis ordinateur

- Commanditaire puis travailleur

- L’architecture des demandes de tâches :
  - Un à un
  - Plusieurs à plusieurs
  - Plusieurs à un
  - Peu à un

Voici les fonctionnalités récurrentes dans les projets de *gamification* d’après (Hamari, 2014 et Göttl, 2014) :

- Fonctionnalités sociales :
  - possibilité de partager sur son réseau social, d’ajouter un “like” et ainsi, de faire connaître l'existence du jeu à ses relations et, en particulier, l’affichage des internautes figurant dans le réseau social du joueur et permettant de leur proposer de jouer avec lui tout en faisant connaître le jeu de manière virale.
  - possibilité de tchater et d’envoyer des messages
  - micro paiements permettant aux internautes d’encourager financièrement le développement des jeux
- Fonctionnalités de jeux :
  - statistiques, nombre de points, médailles, grades, classements, récompenses, challenges, compétitions, défis, objectifs
  - possibilité de jouer avec d’autres joueurs sans simultanéité, en simulant le temps réel.
  - limite de temps, chronomètre

En termes de résultats, le *crowdsourcing* traditionnel et la *gamification* ont été comparés dans plusieurs études. (McCarthy, 2012) a cherché à comparer de manière empirique les résultats entre la correction participative d’OCR de manière traditionnelle avec les résultats obtenus avec *gamification* sur le modèle de Digitalkoot afin de vérifier si la *gamification* pouvait avoir pour effet d’accroître la motivation des participants. Au cours de l’expérience, 2 groupes ont donc été constitués. Avec la *gamification*, on obtiendrait 20 % de participation en plus selon

ses conclusions. De la même manière, d'après (Flanagan, 2012), les jeux permettraient de récolter plus de mots clés par personnes. Ainsi, on obtiendrait en moyenne de 6 tags par visiteurs pour le projet Flickr de la Bibliothèque du Congrès contre 84 tags par visiteurs pour le jeu Tiltfactor Metadata Game.

D'après l'étude menée par (Sabou, 2013) à partir d'une analyse et d'une synthèse de la littérature sur le sujet, le *crowdsourcing* classique serait beaucoup moins coûteux et demanderait moins de temps pour être mis en place par rapport aux jeux. D'après cet auteur, la motivation des bénévoles serait plus facile à maintenir. Enfin, le *crowdsourcing* classique serait mieux perçu, d'un point de vue éthique, par le publique et bénéficierait d'une meilleure image. En comparant les résultats obtenus via un jeu et via l'Amazon Mechanical Turk Marketplace, les auteurs de l'étude estiment que le jeu permettrait de mobiliser une variété moins diversifiée de profils de contributeurs qu'avec le *crowdsourcing* rémunéré via l'Amazon Mechanical Turk Marketplace. Néanmoins, le jeu permettrait de sous-traiter des tâches bien plus complexes et d'obtenir une meilleure qualité de production, il aurait aussi un coût à la tâche très légèrement inférieur et il favoriserait moins la fraude.

Avec le jeu, les joueurs sont d'avantage motivés par des raisons intrinsèques (amusement) tandis que sur l'Amazon Marketplace les motivations extrinsèques (récompense financière) dominent. Il pourrait donc être opportun d'expérimenter un jeu faisant appel aux deux types de motivations avec un jeu qui offre une récompense en argent aux meilleurs joueurs. Ils seraient ainsi à la fois motivés par des raisons aussi bien intrinsèques qu'extrinsèques.

Selon (Harris, C. G., 2013), avoir recours à la *gamification* plutôt qu'au *crowdsourcing* classique permettrait d'améliorer la rapidité et la qualité des contributions mais serait plus coûteux et plus long à mettre en place. (Göttl, 2014) estime également que les Games with a Purpose (GWAP) sont particulièrement coûteux à développer. Dans le cadre de la thèse (Harris, C. G., 2013), l'auteur a cherché à comparer les résultats obtenus pour l'identification d'acronymes selon ces 2 modes de contribution et entre des étudiants et des travailleurs de l'Amazon Mechanical Turk Marketplace. Pour la *gamification*, les joueurs étaient

chronométrés, évalués en temps réel et classés à la fin de la partie. Il a constaté une plus grande précision dans l'identification des acronymes. Mais les étudiants se distingueraient par de plus fortes capacités à résoudre les identifications les plus difficiles. Selon cette étude, la *gamification* devrait donc être privilégiée pour les tâches les plus simples, les plus fastidieuses et qui ne nécessitent pas une trop grande concentration.

Ces éléments à propos de la *gamification* ont fait l'objet d'un article (Andro, 2015, 1)

3.2- Communication et marketing pour recruter les bénévoles

Si le recours aux internautes permet de bénéficier d'une forme de travail bénévole et gratuit, les institutions ne doivent pas négliger que des dépenses importantes devront être consenties afin de développer les plateformes et afin de recruter les bénévoles. Des investissements non négligeables seront donc nécessaires. Néanmoins, les institutions culturelles jouissent déjà d'un public et d'une bonne image auprès de public. En tant que services publics, elles apparaissent comme dignes de confiance, sans buts lucratifs et au service de l'intérêt général. Elles disposent déjà souvent d'une longue expérience de mobilisation de bénévoles, de création d'événements. Parmi les moyens de communication utilisés par les institutions culturelles, nous avons relevé :

- le collage d'autocollants, d'affiches, la production de posters,
- d'articles académiques,
- de dépliants distribués à l'occasion de salons
- de conférences et de congrès, l'organisation de réunions publiques ou d'événements, en identifiant et en contactant des personnes susceptibles de contribuer (Bauer, 2010),
- le recours aux mairies, aux écoles, aux associations, la mobilisation de communautés déjà constituées,
- la production de vidéos, de widgets, de teasers (aguiches) afin d'augmenter le trafic web du site,

- l’utilisation du mailing, la présence active sur les réseaux sociaux, (Twitter, Facebook, Vimeo, LinkedIn)
- et l’utilisation du trafic web déjà généré par le site institutionnel et son catalogue en ligne.
- Transcribe Bentham a même expérimenté, mais malheureusement sans grand succès, l’achat de mots dans le cadre d’une campagne Google Adwords.
- Plus classiquement, les médias traditionnels ont également été utilisés avec succès (campagnes de presse avec communiqués dans la presse spécialisée, locale, nationale, les bulletins des collectivités, émissions de radio et de télévision).

En France, il n’existe encore guère de projets de *crowdsourcing* appliqué aux bibliothèques. Dans ces conditions, il est probable que l’émergence d’un premier projet d’envergure puisse bénéficier d’une certaine couverture médiatique liée à la nouveauté de ce type de démarche.

Donelle McKinley est une doctorante de la Victoria University qui travaille spécifiquement sur le design de l’interface des sites de *crowdsourcing*. D’après ses recommandations (McKinley, 2013), un site de *crowdsourcing* doit avoir une page d’accueil qui décrit le projet et invite à la participation des bénévoles, et d’autres pages Web pour instruire les bénévoles sur l'exécution des tâches. Les internautes doivent avoir une idée claire de pourquoi ils sont là et qu’est-ce qu’ils ont à faire. Pour convaincre un individu de collaborer, il faut qu’il soit intéressé par le sujet, qu’il ait l’impression que sa participation sera utile, que le projet est faisable, qu’il sera accompagné, pourra obtenir des réponses à ses questions, qu’il disposera d’assistance, de forums d’entraide, de listes de discussions, et qu’il sera reconnu pour son travail. D’autres pages encore sont dédiées à l'enregistrement des volontaires. Elles doivent leur présenter des informations détaillées sur le projet, son équipe, son état d’avancement, donner accès aux profils des autres bénévoles. Donelle McKinley recommande enfin de minimiser l’effort de l’utilisateur, de permettre une intégration rapide de nouveaux contributeurs sans formation préalable grâce à un système intuitif et ergonomique. Ainsi, certains internautes n’ont peut être que quelques minutes à consacrer au projet, mais il est nécessaire

de pouvoir capter ces précieuses minutes, d'autant que ces internautes peuvent être légion. Pour pouvoir contribuer, il ne devrait pas être nécessaire d'avoir à effectuer plus de trois clics, comme le signale (McKinley, 2012)

Le contenu de la communication doit être simple, claire, courte et volontariste. Ainsi, comme le signale (McKinley, 2012), dans le cadre du projet What's on Menu ? la phrase "*Aidez la Bibliothèque Publique de New York à enrichir une collection unique*" est à la fois courte et simple, mais elle permet à la fois de dire ce qu'est le projet, qui est le commanditaire, à qui il s'adresse, comment participer, pour quel objectif et quelle est la raison de participer. Des expressions comme celles-ci pourraient également être utilisées : "Aidez-nous à créer un accès libre et gratuit au patrimoine imprimé de la Bibliothèque", "grâce aux efforts de personnes comme vous", "les volontaires du monde entier", "XX % du fonds a été corrigé grâce à vous. Il ne reste plus que XX % à corriger" etc. Le Cleveland Museum of Arts invite les internautes à ajouter des mots clés aux œuvres qu'il diffuse sur le web en affichant le message "aidez les autres à trouver cette œuvre" (Trant, 2006).

Capture d'écran de la communication de What's on the menu? "Help the New York Public Library improve a unique collection. We need you! Help transcribe. It's easy! No registration required!", d'après (Vershbow, 2012)

On trouvera en annexe d'autres exemples illustrant la manière dont les sites des divers projets ont communiqué.

Dans le cadre du projet Steve Museum, le Cleveland Museum of Art invite ses visiteurs ajouter des mots clés en ayant recours à ce slogan : “Help others find this object” (Aidez les autres à trouver cet objet). (Chun, 2006)

Pour mesurer l’importance des mots choisis pour recruter des bénévoles, Jeff Howe relate un projet au cours duquel on invitait les internautes à devenir des journalistes citoyens. Mais, hélas, personne n’avait cliqué sur la phrase “Be a Citizen Journalist”. Lorsque la phrase avait été changée par “Tell Us Your Story” afin d’inviter les internautes à raconter une histoire, le résultat n’avait pas été meilleur. Par contre, lorsque la phrase fut remplacée par “get published” (“soyez publiés”), les internautes avaient enfin afflué en masse (Organisciak, 2010).

Afin de recruter des contributeurs, certaines institutions peuvent aussi, grâce à un dispositif de veille, surveiller les réseaux sociaux, en particulier Twitter, Facebook, les listes de diffusion et les forums, lorsqu’on parle d’elles ou de leurs collections et identifier ainsi de potentiels contributeurs à recruter.

D’un point de vue marketing, plusieurs techniques de psychologie sociale peuvent être mobilisées pour augmenter le nombre de participations :

- La technique de l’étiquetage qui consiste à affirmer qu’on considère déjà que les internautes à qui l’on s’adresse comme des bienfaiteurs qui œuvrent déjà positivement à la valorisation du patrimoine.
- La technique du “pied dans la porte” qui consiste à proposer des tâches très faciles à réaliser pour provoquer un premier acte d’engagement symbolique qui sera généralement suivi d’un engagement plus important. Le simple fait de s’inscrire sur un site est d’ailleurs déjà un acte d’engagement.
- La technique du “pied dans la bouche” qui consiste à se préoccuper des internautes en leur demandant poliment des nouvelles individuelles en début d’interaction.
- La technique du “mais vous êtes libre de” qui consiste à rappeler que l’internaute est libre d’accepter ou de refuser de participer.

- La technique du “un peu, c’est mieux que rien”. En affirmant, par exemple, que même 10 minutes du temps de l’internaute aiderait déjà considérablement la bibliothèque.

3.3- La question des motivations

Si, à la différence des salariés, les bénévoles n’attendent pas nécessairement un apport financier en retour de leurs contributions, ils doivent néanmoins bénéficier d’un retour de la part des institutions qui bénéficient de leur travail. Les projets de *crowdsourcing* se font toujours nécessairement au bénéfice mutuel de l’institution et de l’internaute.

La question des motivations des contributeurs aux projets de *crowdsourcing* est récurrente dans la littérature. On distingue généralement les motivations intrinsèques et les motivations extrinsèques qui peuvent d’ailleurs prédominer de manière très différente d’un individu à l’autre.

Les motivations intrinsèques, internes à l’individu, sont celles qui le poussent à agir pour le seul intérêt du travail et du plaisir qu’il lui procure. L’activité est ainsi une fin en soi, un art pour l'art pratiqué pour lui-même, pour son seul contenu et pour les seules satisfactions que l’on en tire, pour la beauté du geste, pour l’accomplissement de soi ou les responsabilités, de manière passionnée et désintéressée, sans rechercher une reconnaissance ou une récompense qui risqueraient, au contraire, de diminuer la motivation. L’activité est pratiquée par plaisir, par curiosité, par sentiment de compétence, de recherche des finalités, par sentiment de liberté et d’autodétermination.

Au contraire, les motivations extrinsèques, extérieures à l’individu, sont celles qui le poussent à exercer une activité en vue d’obtenir un résultat extérieur à cette activité, à rechercher les effets et les conséquences de l’activité en dehors de l’activité elle-même comme la reconnaissance, la récompense ou la rémunération. Elle est donc plus contraignante et moins libre. L’activité est ainsi un instrument, un simple moyen d’atteindre une fin et d’obtenir le résultat recherché (dans le monde du travail, l’argent ou le fait d’éviter des sanctions).

Les ressorts susceptibles de stimuler un internaute varient en fonction de la diversité des psychologies, des cultures et des classes sociales des individus. Il convient donc de prendre en compte la diversité des profils et la diversité des motivations susceptibles de les animer (Smith, 2013).

En s'inspirant de différentes taxonomies et études sur les motivations des contributeurs trouvées dans la littérature (Rouse, 2010 ; Kaufmann, 2011 ; Alam, 2012 ; Dworak, 2012 ; Alam, 2013 ; Dunn, 2013 ; Owens, 2013 ; Smith, 2013), voici une taxonomie originale que nous proposons :

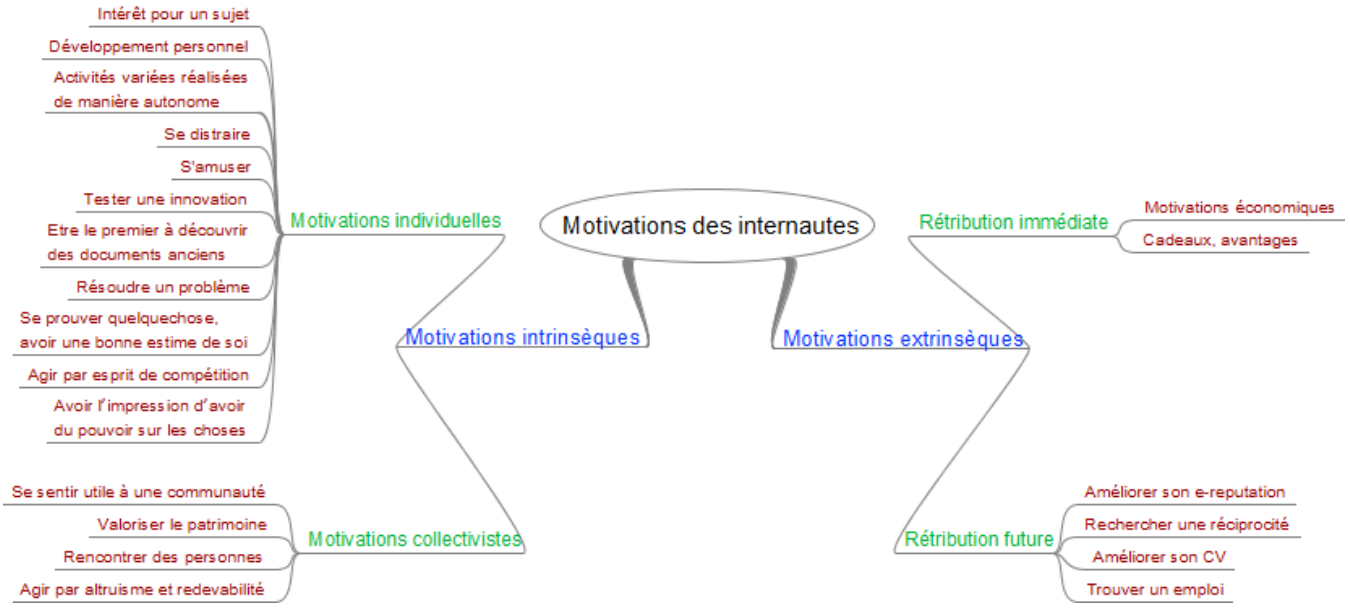


Figure 72. Taxonomie des motivations des bénévoles dans un projet de *crowdsourcing*

Ces grands types de motivations se déclinent plus précisément de la manière suivante :

3.3.1- Les motivations intrinsèques

Parmi les motivations intrinsèques, nous avons relevé, dans la littérature, des motivations liées au plaisir individuel et des motivations collectivistes :

Motivations individuelles :

- Pour l'intérêt pour une science particulière, pour les écrits d'un scientifique, pour une discipline.
- Pour le développement personnel, se cultiver et apprendre, développer des compétences, satisfaire sa soif de connaissances y compris sa connaissance de soi-même (histoire locale, généalogie).
- Pour se distraire, passer le temps, ne pas s'ennuyer et rester actif
- Activités variées réalisées de manière libre, flexible, autonome et responsable. Avec le *crowdsourcing* rémunéré, en particulier, on peut travailler librement quand on veut, où on veut, autant que l'on veut, pour qui on veut et choisir les tâches que l'on veut accomplir.
- Pour le plaisir, pour s'amuser et jouer (*gamification*). Certains projets ont même provoqué une véritable addiction, générant un temps de travail de près de 60 heures par semaine.
- Pour satisfaire sa curiosité de tester une démarche innovante dans l'histoire des nouvelles technologies
- Pour être le premier à lire des manuscrits historiques et avoir l'opportunité d'éditer des documents historiques. A l'instar de l'archéologue qui découvre le premier une relique depuis longtemps ensevelie, être le premier à faire quelque chose d'important pour un document patrimonial.
- Pour résoudre un problème intellectuel ou technologique. Réaliser des tâches qui ne peuvent pas encore être automatisées et prises en charge par des algorithmes.
- Par esprit de compétition, de challenge, pour prouver ce qu'on sait faire, par défi y compris collectif. Ainsi, de nombreux projets affichent, en temps réel, le pourcentage de ce qui reste à accomplir. Ils peuvent également afficher les classements des meilleurs contributeurs de la semaine, du mois, de l'année, depuis toujours pour telle ou telle zone géographique afin que chacun puisse espérer être sur le tableau et provoquer une émulation. Les projets ont aussi intérêt à donner à chacun des tableaux de bord de statistiques personnalisées, à attribuer des médailles et des grades aux contributeurs et, pour les projets de

crowdfunding, à classer les mécènes par ordre d'importance. Sur l'Amazon Mechanical Turk Marketplace, le grade de Master est obtenu par les travailleurs ayant le plus contribué en quantité et en qualité d'après les évaluations de leurs livraisons.

- Pour se prouver quelque chose, se réaliser, améliorer son amour propre, avoir une bonne estime de soi, se sentir efficace, utile et compétent. Cette motivation serait particulièrement importante pour les chômeurs.
- Pour avoir l'impression que son avis est pris en compte, qu'on est consulté, qu'on a du pouvoir sur les choses, qu'on peut changer des choses dans le monde, de laisser son empreinte (Shirky, 2008), d'être auteur et acteur, par vanité.
- Ne pas être un consommateur passif d'informations mais un actif producteur de connaissances.

Motivations sociales, communautaires et collectivistes :

- Se sentir utile pour une communauté, un groupe, pour la société, servir son pays en valorisant son patrimoine, se mettre au service de la science, l'intérêt général, le bien public, agir pour une cause, pour des valeurs ou pour des idéaux (motivations principalistes). Avoir le sentiment de participer à une cause ou à un mouvement qui nous dépasse. Digitalkoot fait, par exemple, explicitement appel au patriotisme : "Start saving … Finnish culture here". De nombreux contributeur de OpenStreetMap on probablement l'impression de s'opposer à l'hégémonie de Google Map. Dans le jeu de puzzle fold.it, à l'instar des jeux où il faut sauver le monde des envahisseurs, on doit chercher à percer le secret des protéines (Good, 2011).
- Participer à la libre diffusion, l'utilisation et la conservation du patrimoine, participer à sa réutilisation et à sa valorisation. Les contributeurs ne souhaitent généralement pas que leur travail puisse être réutilisé commercialement par des sociétés. Ils tiennent également à travailler pour des organisation à but non lucratif comme les bibliothèques.
- Possibilité de rencontrer des personnes, d'échanger, d'être en interaction et être connecté à un réseau social

- Faire quelque chose de désintéressé, ni par bénéfice personnel, ni pour de l'argent, dans un esprit d'altruisme, de partage, de générosité, de charité et de philanthropie
- Se sentir redevable par rapport au service rendu par le site et se faire, en retour, un devoir de participer.

3.3.2- Les motivations extrinsèques

Rétributions immédiates :

- Motivations économiques (rémunération du travail fourni sous le forme de paiements). Ce type de motivation pourrait avoir un effet négatif sur d'autres types de motivations.
- Cadeaux, avantages. Certains projets offrent des cadeaux (stylos, livres, t-shirts…) ou des bons d'achats dans la bibliothèque pour de la numérisation ou de l'impression à la demande à ceux qui ont le plus contribué. D'autres organisent des événements et des banquets "in real life" réservés aux bénévoles ou leur financent le voyage pour visiter l'institution.

Rétributions futures :

- Améliorer sa popularité sur le web, son *e-reputation* en apparaissant sur Internet comme bénévole d'un projet culturel, faire de l'autopromotion, améliorer son statut social et satisfaire sa soif de reconnaissance sociale (notamment pour les personnes au chômage). Bénéficier de la prestigieuse fonction de conservateur et travailler pour une institution célèbre. Certains projets remercient individuellement leurs contributeurs par mails personnalisés et par remerciements publics dans la communication écrite et orale de l'institution sur son site web, sa newsletter ou sur les réseaux sociaux. Le projet scientifique Galaxy Zoo, a ainsi ajouté le nom des internautes dans la liste des auteurs des publications scientifiques produites au cours du projet. Dans les projets de folksonomie, le nom de la personne qui a ajouté les mots clés peut également être cité. Enfin, pur les projets de *crowdfunding* et de numérisation à la demande, le nom des mécènes et un lien

vers leurs sites web doivent être signalés afin de leur permettre un retour sur investissement en terme de trafic web si un livre génère beaucoup de visites (sur le modèle de Google Adwords). Sur YouTube, les contributeurs sont d'autant plus actifs que leurs vidéos génèrent du trafic web (Huberman, 2009).

- Rechercher une réciprocité (on recevra plus facilement de l'aide sur Internet si on a soi même déjà aidé les autres, "Je le fais parce que j'aimerais qu'on le fasse pour moi")
- Trouver un emploi ou connaître une évolution de carrière grâce à cette autopublicité
- Développement personnel pour une évolution de carrière

3.3.3- L'opposition entre les motivations intrinsèques et extrinsèques

Parmi toutes ces motivations, on distingue des motivations individualistes (accroître son propre bien être), des motivations altruistes (accroître le bien être de son prochain), des motivations collectivistes (accroître le bien être de son groupe) et des motivations principalistes (défendre un principe moral comme la liberté, l'égalité, la fraternité ou la justice).

Une étude sur les motivations des bénévoles des projets de *crowdsourcing* culturel (Brabham, 2010), réalisée à partir de données recueillies par messagerie instantanée auprès de 17 personnes en mars, avril et octobre 2008 révèle que les motivations intrinsèques (plaisir, amusement, résoudre des problèmes, améliorer ses compétences, addiction) prédomineraient sur les motivations plus extrinsèques (argent, opportunités professionnelles, amour de la communauté). L'altruisme des contributeurs serait toutefois discutable pour certains projets. Ainsi, pour le projet ACM Digital Library, la majorité des corrections effectuées en ligne sur des références bibliographiques serait tout simplement l'œuvre des auteurs eux-mêmes (Bainbridge, 2012).

D'après une autre enquête menée par (Dunn, 2012), 79 % des contributeurs actifs agissent pour des motivations à la fois pour eux-mêmes et pour les autres. Sur 59 personnes, 24 affirment agir par intérêt pour le sujet, 3 pour

aider les autres à apprendre, 2 pour contribuer à la science, 2 pour expérimenter le *crowdsourcing*, 1 pour s'impliquer dans le bénévolat et 1 pour la nouveauté. 1 seul estime qu'un algorithme informatique aurait pu être utilisé à la place du travail humain. D'après une analyse de 207 messages du forum du projet Galaxy Zoo (Raddik, 2010) a constaté que la motivation principale était l'astronomie (39%), suivie du désir de contribuer (13%) puis une préoccupation pour l'immensité de l'univers (11 %).

(Acar, 2011) a étudié l'impact des gratifications sur les travailleurs principalement stimulés par des motivations intrinsèques et semble trouver un effets négatif des gratifications sur ce type de personnes. Par ailleurs, la qualité des données produites serait améliorée avec le recours aux motivations intrinsèques comme le suggère (Rogstadius, 2011) qui a comparé la qualité des données produites gratuitement à partir de motivations intrinsèques avec celles de données produites contre rémunération à partir de motivations extrinsèques.

Malgré l'intérêt pour les motivations intrinsèques dans le cadre de projets culturels, l'expérience du projet TROVE invite toutefois à ne pas négliger les motivations extrinsèques. En effet, pendant les 6 premiers mois du projet TROVE, la moitié des contributions étaient anonymes et le fruit de motivations plutôt intrinsèques (intérêt personnel, altruisme). Six mois après le lancement du projet, seules 20 % des contributions étaient toujours anonymes. Les bénévoles ont donc probablement aussi besoin de reconnaissance. C'est la raison pour laquelle les motivations extrinsèques ont ensuite été développées sous la forme de classements statistiques par les porteurs du projet TROVE.

3.3.4- Les motivations spécifiques des projets de *gamification*

Les motivations mobilisées spécifiquement par les projets de *gamification* semblent être les suivantes :

- développement personnel (acquisition de compétences, résolution de problèmes)
- des récompenses (argent, prix, promotions, reconnaissance, responsabilités)
- amusement et distraction
- information (sur l'avancée du projet, le volume de leur propre contribution)

D’après (McCarthy, 2012), les joueurs masculins auraient d’avantage tendance à évaluer leurs performances que les joueurs féminins qui seraient plutôt attirées par le caractère relationnel des jeux. De manière générale, les femmes seraient d’avantage attirées par les jeux de gestion, de puzzle, de combat et d’aventure. Et pour leur part, le genre masculin aurait une préférence pour les jeux de sport, de tir, de stratégie ou pour les jeux de rôle.

D’après (Dunn, 2012), la *gamification* peut parfois aussi être un obstacle pour certains utilisateurs qui veulent s’engager ou qui s’intéressent à un sujet car le développement de connaissances peut y être moindre. Par ailleurs, certains joueurs risquent de produire un travail quantitatif de piètre qualité uniquement pour pouvoir être bien classés. Par exemple, dans le projet Old Weather de transcription de pages manuscrites de journaux de bateaux du 19è siècle contenant des observations météorologiques, certains risquent de délaisser la qualité pour passer plus rapidement du grade de la marine de cadet à celui de lieutenant puis à celui de capitaine de bateau ou aussi pour pouvoir conserver leur titre de capitaine de vaisseau. D’autres risquent de se démotiver et de renoncer à chercher à se mesurer à des joueurs trop bien classés et trop difficiles à détrôner (Eveleigh, 2013). Ainsi, sans provoquer une addiction autour des microtâches proposées, mais afin d’accroître la motivation des joueurs et (Von Ahn, 2008) suggère d’afficher, en plus du top des meilleurs joueurs, le top des meilleurs joueurs sur le mois, sur la semaine, dans la journée... afin d’encourager encore d’avantage la participation des joueurs qui, sur un semaine peuvent espérer monter sur le podium. Allant dans le même sens, (Ridge, 2011) indique que de nombreux joueurs sont stimulés à la fois par l’envie immédiate et locale de battre le joueur situé juste au dessus dans la classement et l’objectif à long terme de battre le score le plus élevé. Dans ces conditions, il suggère que la liste des meilleurs scores puisse non seulement être affichée par heure, jour, semaine, mois, année, tout le temps mais aussi par ville, région, pays et continent. En croisant ces deux variables, on pourrait afficher, par exemples, la liste des meilleurs joueurs par pays et par mois ou celle des meilleurs joueurs par ville et par année… Dans la

mesure où les joueurs semblent réagir différemment aux différents types d'objectifs, il serait ainsi possible d'afficher pour chacun le type d'objectif qui correspond le mieux à sa personnalité.

La possibilité de faire gagner un cadeau ou une somme d'argent aux meilleurs joueurs pourrait également être un excellent moyen d'accroître leurs contributions. On peut aisément imaginer qu'un tel gain attirerait bien plus que les jeux de hasard puisqu'il serait véritablement possible de gagner en fonction de son acharnement. La valeur produite par tous les joueurs servant à générer le gain. Un autre modèle pourrait être de rétribuer les joueurs à hauteur de leur contribution.

3.3.5- *Crowdsourcing* et récompenses

Dans la plupart des projets de *crowdsourcing* et pas exclusivement dans les projets de *gamification*, les contributeurs sont classés selon leur contribution, à la manière des jeux vidéos. Ainsi, les internautes, dans leurs espaces propres, ont généralement accès à leurs statistiques et les listes de documents sur lesquels ils ont travaillé. Cela peut être très bénéfique pour leur autopromotion, leur *e-reputation* et pour rechercher un emploi. Ainsi, des contributeurs du projet Gallaxy Zoo ont été remerciés et associés dans des articles issus du projet, un super contributeur a été invité à témoigner à une prestigieuse conférence publique organisée autour du projet Transcribe Bentham. Mais, au delà de récompenses sociales, des récompenses symboliques ou matérielles et en nature sont proposées par certains projets comme ArcHIVE, CONCERT et TROVE et de véritables rétributions financières octroyées aux internautes travaillant au service des bibliothèques sur l'Amazon Mechanical Turk Marketplace ou sur CrowdFlower. On parle dans ce cas de *crowdsourcing* rémunéré.

Il est très important de chercher à fidéliser les participants afin d'obtenir des données de participants expérimentés, donc des données de meilleure qualité. Pour cela, on peut afficher la liste des plus gros contributeurs, afficher le nom du contributeur pour chaque contribution, valoriser un contributeur particulier en mettant en avant sa biographie dans une newsletter, les remercier individuellement, les récompenser par des diplômes, des formations, des formations diplômantes

reconnues, des cadeaux, des abonnements, des livres, organiser des sorties, des événements, la participation à l'analyse des résultats... (Bauer, 2010).

Concernant, par exemple, le projet australien ArcHIVE, la récompense prend la forme d'une rétribution symbolique permettant d'échanger les points gagnés contre des fac-similés (*print on demand*), des objets, des marque-pages ou des posters. Ce modèle a également été adopté par le site de *crowdsourcing* culturel http://velehanden.nl (consulté le 23 juin 2016) qui permet de convertir les points accumulés en cadeaux, services ou rétribution financière (Djupdahl, 2013). Concernant le projet TROVE, Rose Holley a évoqué la possibilité d'offrir des cadeaux, des T-shirts, des livres, des diplômes, des formations, des cérémonies de remerciements publics sur le web, dans les réseaux sociaux, dans des bulletins ou des visites particulières des fonds de la Bibliothèque Nationale d'Australie aux meilleurs contributeurs (Holley, 2009). Dans la littérature, il est également mentionné, l'invitation à rencontrer le chef de la cartographie de l'institution dans le cadre du projet de *gamification* British Library Georeferencer project (Dunn, 2012), mais aussi des lecteurs MP3, des produits gratuits, l'accès à des fonctionnalités avancées sur une plateforme (Biella, 2015) ou même de petites récompenses financières offerts aux bénévoles (Rouse, 2010), ou enfin, des bons d'achats chez Amazon (Birchall, 2012). Pour finir, s'agissant des projets de *crowdfunding* comme Numalire, le retour sur investissement pour un mécène ou un internaute peut se mesurer en terme de trafic web généré par les livres qu'il a permis de numériser et la publicité faite autour de son nom ou du nom de son entreprise ou encore de son institution.

D'après une étude rapportée par (Ipeirotis, 2011), l'argent n'aurait pas d'impact sur la qualité des données produites mais en aurait généralement un sur la participation. Néanmoins, le fait de transférer une motivation intrinsèque vers une incitation plus extrinsèque peut aussi provoquer des effets négatifs. (Rogstadius, 2011) estime, par exemple, qu'une rémunération faible pourrait avoir un effet moins positif qu'une absence de rémunération. Ainsi, les pressions extérieures, et en particulier les récompenses extrinsèques pourraient avoir un effet négatif sur les motivations intrinsèques qui fondent par exemple la

gamification. Rémunérer les internautes pourrait donc avoir pour effet de leur donner le sentiment qu'ils perdent leur autonomie et leur liberté et diminuer paradoxalement leur volonté de jouer (Hamari, 2014).

Ainsi, comme le relate (Groh, 2012), des expériences ont montré que s'ils sont rémunérés pour dessiner, des enfants vont peut être dessiner d'avantage mais que leurs dessins plus nombreux seront de moins bonne qualité, et que les enfants auront perdu le goût pour le dessin, en particulier s'ils cessent ensuite d'être rémunérés. De la même manière, comme le rapporte (Rogstadius, 2011), une expérience de 1975 de Edward L. Deci aurait montré que des étudiants ayant été rémunérés pour jouer à des jeux de puzzles avaient également perdu tout intérêt pour cette activité si elle n'était plus rémunérée. Dans son livre, Homo economicus : Prophète (égaré) des temps nouveaux publié en 2012, Daniel Cohen, relate qu'un directeur d'un centre de transfusion sanguine décida d'offrir une prime aux donateurs de sang et que cette action eut pour résultat paradoxal d'en diminuer significativement la quantité. Comme l'auteur le dit :

> "*S'il ne s'agit plus d'aider les autres mais de gagner de l'argent, leur participation change de nature. Un autre lobe de leur hémisphère est sollicité. L'homme moral quitte la salle quand l'Homo economicus y entre*."

Michel Bauwens qualifie ce phénomène d'éviction par le terme de « crowding out ». (Bauwens, 2015). L'être humain est bien un être complexe doué d'un libre arbitre, il n'agit pas exclusivement par amour de la carotte et par peur du bâton.

3.3.6- Les autres théories sur les motivations

Plus généralement, d'après la théorie de Maslow, les besoins se répartiraient de manière hiérarchique et pyramidale de la manière suivante avec des besoins vitaux à la base (besoins physiologiques d'existence, de sécurité, d'appartenance) et des besoins d'ordre supérieur au sommet (besoins de relation, d'estime, de pouvoir, de progression, de réalisation) :

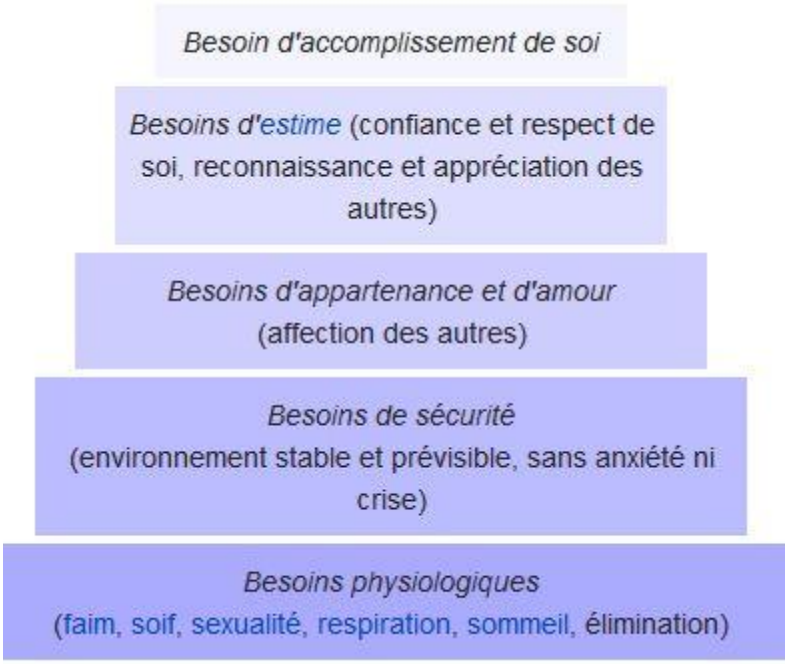


Figure 73. Pyramide des besoins de Maslow (d'après http://fr.wikipedia.org/wiki/Pyramide_des_besoins le 26 juin 2016)

Lorsque les besoins primaires sont satisfaits, de nouveaux besoins, supérieurs, apparaissent. Mais, par contre, ceux-ci ne peuvent exister tant que les besoins de bases ne sont pas déjà satisfaits. Dans ces conditions, il ne faudrait pas surestimer les récompenses au détriment de l'intérêt pour le travail lui même. La non prise en compte des besoins intrinsèques pourrait même déshumaniser le travail, le salaire n'étant pas réellement un facteur de motivation, mais plutôt un facteur de satisfaction. Il en est de même pour les projets faisant appel aux foules d'internautes.

(Gouil, 2014) évoque les travaux de Stéphane Debove qui estime que des raisons biologiques comme l'instinct de parenté, la nécessité de favoriser la transmission des gènes de nos proches, le besoin d'accroître par la collaboration les chances de survie de notre groupe, la possibilité d'améliorer notre réputation, de trouver un partenaire et de transmettre ainsi notre patrimoine génétique pourraient expliquer pourquoi les internautes coopèrent autant sur le web. De manière moins académique mais très efficace, sur son blog donneesouvertes.info, Simon Chignard, relate que lorsque des automobilistes font des appels de phares

à d'autres automobilistes afin de les prévenir de la présence d'un radar de gendarmerie, ils agissent à la fois par attente d'une réciprocité, ou parce qu'ils se sentent redevables d'avoir déjà été prévenus, par opposition au gendarme, ou par sentiment de solidarité avec la communauté des automobilistes.

Bien d'autres classifications des motivations existent. Ainsi, Herzberg identifie 5 facteurs de satisfaction : l'accomplissement, la reconnaissance de cet accomplissement, le travail en soi, la responsabilité et la progression socioprofessionnelle. Pour leur part, les théoriciens Porter et Lawle estiment que l'action doit être motivée par les 3 facteurs suivants : l'intérêt et l'enjeu de l'action, la contrepartie des relations sociales de l'acteur, la capacité de l'acteur à mener l'action. Pour McClelland (1961), la motivation est influencée par la variété des activités, les tâches qu'on peut réaliser entièrement et dont on peut revendiquer la paternité, le sens des tâches, l'autonomie, la possibilité de décider et, enfin, le retour que ces tâches peuvent apporter. Les individus obéissent aussi à des types de motivations diverses. Certains ont besoin de réalisation, d'autres de pouvoir, et d'autres enfin ont besoin d'appartenance ou d'affiliation. Douglas McGregor, quant à lui, estime qu'il existe 2 types de conceptions. Ceux qui pensent que les humains ont une aversion naturelle pour le travail et fuient tout type de responsabilité. Il serait, par conséquent, nécessaire de contrôler et d'éduquer les salariés pour obtenir du travail de leur part et d'utiliser la méthode de la carotte et du bâton pour les faire avancer par crainte de la sanction et recherche de la récompense. Cette théorie, dite "théorie X" est partagée à la fois par le grand capitaliste Ford et peut être aussi par certains certains marxistes comme Paul Lafargue avec son "Droit à la paresse". D'autres ont une théorie diamétralement opposée, la théorie Y. Selon eux, les individus aiment naturellement travailler, en retirent de la satisfaction et du plaisir et recherchent les responsabilités. Dans ces conditions, le travail est libérateur et épanouissant, il permet de se réaliser socialement et de se développer à l'instar des hobbies et des loisirs. Par conséquent, il faut favoriser la confiance, la responsabilité, l'autonomie, la liberté, le sens de l'initiative, la créativité des employés afin de les motiver et obtenir d'eux un résultat optimal. C'est sur ce type de conception des motivations que se fonde le *crowdsourcing* et

la motivation des bénévoles dépendra de la variété de leurs tâches, de leur autonomie, de leurs responsabilités, des informations, et du retour donné à ces tâches (feed back).

3.3.7- Les motivations des institutions culturelles et les pré-requis pour lancer un projet de *crowdsourcing*

D'après une enquête de 2010 rapportée par (Thuan, 2013) 10 % des entreprises auraient déployé une stratégie de *crowdsourcing*. Mais, comme le souligne (Alam, 2013), peu d'études se sont intéressées aux motivations des organisations et des institutions pour le *crowdsourcing*. L'article insiste sur le fait que les motivations qui poussent les institutions à avoir recours au *crowdsourcing* sont les mêmes que celles qui les conduisent à externaliser. Il s'agit, en particulier, de réduire les coûts et d'améliorer le rapport résultats / coûts (Lebraty, 2015). Concernant les bibliothèques, en particulier, il s'agira de diminuer les coûts dans un contexte de budget resserrés, d'accélérer les chantiers pour lesquelles elles disposent d'insuffisantes ressources humaines ou financières ou de lancer des projets qui ne pouvaient être envisagés pour ces raisons, d'accéder à des compétences et à des connaissances non disponibles en interne et allant au delà de celles d'une équipe limitée, de bénéficier des compétences d'érudits et de chercheurs, de mieux adapter ses services aux besoins et mieux faire connaître les activités des professionnels au grand public, de résoudre des problèmes impossibles à résoudre sans le *crowdsourcing*, d'améliorer la qualité des données, l'indexation des collections ou de les enrichir y compris avec de nouveaux types d'informations, de demeurer technologiquement pertinent dans une société en changement rapide, d'être innovant, de maintenir son leadership, d'accroître sa notoriété, d'utiliser de manière plus éthique et plus utile les budgets précédemment destinés à faire travailler des pays à bas coût de main d'œuvre et, last but not least, de rechercher de nouveaux types de relations avec les usagers.

En effet, au delà du besoin de recourir aux internautes pour capter de la force de travail gratuite sur le web, de sous-traiter des tâches qu'elle n'a plus les moyens de financer, ou encore d'initier des projets qu'elle n'aurait jamais pu

espérer développer sans l'aide des internautes, le *crowdsourcing* est aussi, et pour certains, surtout, le moyen d'étendre la mission des institutions culturelles, d'engager d'avantage le public au service des thématiques et des collections, en l'impliquant dans la conservation du patrimoine et de la mémoire publique afin de produire de nouvelles connaissances. Il permet aussi de changer la vision du public sur le Musée et la Bibliothèque qui ne sont à l'heure actuelle pas toujours considérés comme ludiques et amusants (Birchall, 2012). La constitution d'une communauté nouvelle, hors les murs, attachée à l'institution et/ou à ses collections devient ainsi aussi une finalité en soi. Il s'agira, pour la bibliothèque, de construire et d'animer une véritable communauté d'internautes autour de ses collections numérisées. L'usage du patrimoine numérisé par les internautes se fera ainsi moins superficiel, moins passif et pourra déboucher sur de véritables recherches. Au lieu de consommer de l'information, ils pourront devenir eux-mêmes producteurs d'informations. Au lieu de demander aux gens de travailler pour la bibliothèque, il s'agira plutôt de leur offrir la possibilité de participer à l'enrichissement du patrimoine commun. Sur son blog, Trevor Owen[31], considère ainsi que le *crowdsourcing*, dans sa meilleure forme, ne consiste pas à faire travailler des usagers, mais plutôt à leur offrir la possibilité de participer à la mémoire publique.

Il semble en tous cas, qu'il existe bien deux conceptions différentes de l'intérêt du recours au *crowdsourcing* pour les institutions, deux conceptions qui ne s'opposent d'ailleurs pas nécessairement. Certains semblent se focaliser sur l'intérêt des institutions en terme de coûts et d'autres sur l'engagement du public pour les collections. La question des coûts ne doit pas être négligée comme relevant d'une vision comptable et manquant de noblesse. Au contraire, les institutions doivent plus modestement reconnaître que l'investissement des internautes leur est nécessaire voir vital. Elles ne doivent pas considérer qu'elles font seulement plaisir aux internautes en leur permettant de s'exprimer. Elles ne doivent pas se contenter de faire du *crowdsourcing* dans une simple logique de communication institutionnelle autour d'un sujet à la mode et ne jamais réintégrer

[31] http://www.trevorowens.org (consulté le 23 juin 2016)

et réutiliser les données produites par les internautes comme c'est malheureusement encore trop souvent le cas. « *Il serait en effet dommage de n'utiliser les potentialités du web social que de façon « cosmétique* », *sans en faire véritablement bénéficier le signalement des collections et l'interface de recherche de la bibliothèque* ». (Moirez, 2013)

Quoi qu'on en dise, la principale force du *crowdsourcing* reste la diminution des coûts, et l'obtention de capacités de travail ou de compétences dont on ne dispose pas en interne. A l'instar de toute externalisation, le paiement (lorsqu'il ne s'agit pas de bénévolat) est basé sur les résultats et non sur le temps de travail effectué, ce qui présente un avantage certain par rapport au salariat.

Au delà des motivations des institutions, certaines conditions sont nécessaires au déploiement du *crowdsourcing* par les entreprises. (Thuan, 2013) et (Crowston, 2013) identifient ainsi le type de tâches (réalisables par internet, non confidentielles, pouvant être réalisées de manière indépendante et nécessitant peu d'interaction et de communication, pouvant être découpées en microtâches et pouvant être réalisées par des non experts), le type de main d'œuvre (foule plus nombreuse et diverse via le *crowdsourcing* qu'en interne où les ressources humaines et les compétences ne suffisent pas, présence de communautés déjà existantes de passionnés), le type de management (le budget dont on dispose est insuffisant et nécessite le recours au *crowdsourcing*, la présence de ressources humaines ayant expérimenté ou étant expertes en *crowdsourcing*, le niveau de qualité exigé) et enfin, l'environnement de travail (plateforme interne ou externe).

3.4- Sociologie des contributeurs et *community management*

Dans l'introduction conceptuelle, nous avons déjà évoqué la sociologie des contributeurs des projets de crowdsourcing de manière générale. Nous traiterons ici plus précisément des usagers dans le domaine des bibliothèques numériques en particulier et à la lumière des projets que nous avons analysés.

3.4.1- Sociologie des contributeurs

Voici une synthèse des informations sociologiques recueillies dans le cadre du diaporama des projets :

	Genre	**Âge**	**Niveau social**
Crowdfunding **(Numalire)**	Hommes (70 %)		Supérieur
Crowdfunding **(Ebooks on Demand)**	Hommes majoritaires		Supérieur
Correction OCR (TROVE)	Femmes (70 %)	Jeunes diplômés en recherche d'emploi et retraités, plus de 50 ans (65 %)	Diplômés
Transcription manuscrits (Transcribe Bentham)	Femmes (plus des deux tiers)	Retraités Jeunes diplômés	Supérieur
Taging **(VeleHanden)**	Hommes plus nombreux	La classe des plus de 50 ans est la plus nombreuse	
Gamification	Femmes aussi	Entre 25 et 44 ans	

(Digitalkoot)	nombreuses mais plus assidues (54 % des tâches), mais les 4 contributeurs les plus importants sont des hommes		
***Gamification* (Art collector)**	Hommes 10 % plus nombreux	25-34 ans	
***Gamification* (Muséum Games)**	Femmes plus nombreuse	Trentenaires	
***Crowdsourcing* rémunéré (Amazon Mechanical Turk Marketplace)**	Femmes américaines Hommes indiens	Plutôt jeunes	Supérieur
Enquête *crowdsourcing* (Dunn, 2012)	Parmi les répondants, 58 % d'hommes et 42 % de femmes	La plupart dans la tranche d'âge 35-45	
Enquête *crowdsourcing* (McKinley, 2013)		Domination des préretraités (classe des 56-65 ans)	

Tableau 17. Données collectées dans la littérature à propos de la sociologie des contributeurs des différents projets

Si on rencontre des différences de genres assez significatives pour chaque projet, il reste difficile d'en tirer des corrélations en fonction des types de projets en dehors de la numérisation à la demande par *crowdfunding* qui semble attirer d'avantage les hommes. De manière générale, le *crowdfunding* attire d'avantage les hommes de revenus élevés (Daudey, 2014). La *gamification* pourrait

également attirer d'avantage les hommes quand le *crowdsourcing* explicite attirerait d'avantage les femmes.

Le rapport de l'association Wikimedia (Beuth Hochschule für Technik, 2014) indique que 9 éditeurs sur 10 de Wikipédia sont des hommes et que cette proportion est même de 97 % pour le Wikipédia indien. Les femmes préféreraient consacrer d'avantage leur temps sur les réseaux sociaux comme Facebook (aux USA, leur niveau de participation à Facebook représenterait 71 % du total), elles disposeraient aussi de moins de temps libre que les hommes et n'apprécieraient pas le ton parfois assez virulent, polémique et agressif des discussions sur Wikipédia.

Concernant le niveau d'étude ou le niveau social, lorsque cette information a pu être récoltée, ce sont des personnes ayant un haut niveau d'études qui semblent être majoritaires parmi les contributeurs.

S'agissant de l'âge des contributeurs, on observe une nette domination des jeunes autour des projets de *gamification* et de *crowdsourcing* rémunéré alors que concernant la correction de l'OCR ou la transcription de manuscrits, ces activités semblent attirer des bénévoles plus âges. Cette observation va dans le même sens que (Daudey, 2014) qui rapporte que 44 % des 12-17 ans contribuent sur le web.

D'après Holley (2009), la plupart des bénévoles du projet TROVE sont des retraités, mais il y a aussi des jeunes diplômés en recherche d'emploi ou encore des chômeurs ou des salariés en arrêt maladie ou en congés qui contribuent. Et, concernant les fonctions à responsabilité comme la modération, elles sont plutôt prises en charge par des trentenaires ou des quarantenaires salariés à plein temps.

La domination des retraités passionnés de généalogie et d'histoire locale dans le *crowdsourcing* culturel classique pose la question de sa pérennité car rien n'indique que les générations futures de retraités auront les mêmes centres d'intérêts (Ayres, 2013) et peut être même, auront autant de temps libre.

3.4.2- *Crowdsourcing* ou *communitysourcing* ?

Les données récoltées dans le cadre du diaporama des projets ont révélé que, pour la plupart des projets, la majorité des données produites est le fait d'une petite minorité bien déterminée de participants et non d'une foule d'anonymes et ce, bien que ces projets s'adressent en théorie à un nombre illimité d'internautes.

Ces observations sont en adéquation avec ce qui est rapporté dans la littérature et notamment par (Brabham, 2012). Ainsi, 80 % du travail serait réalisé par à peine 10 % des volontaires les plus actifs. Certains sont susceptibles d'y consacrer autant de temps qu'ils le feraient pour un travail à plein temps voir d'éprouver une sorte d'addiction pour cette activité.

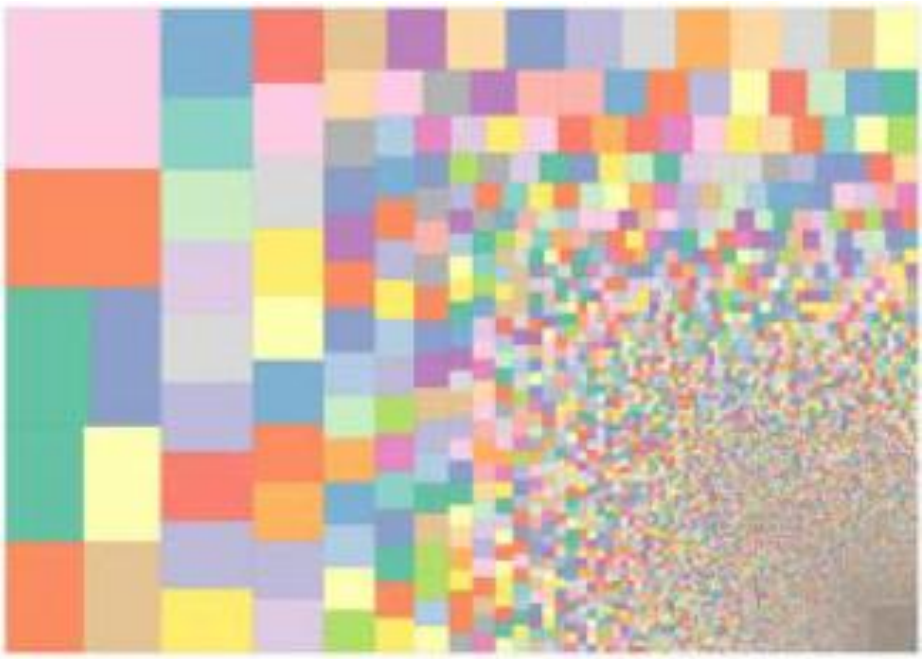


Figure 74. Diagramme illustrant qu'une poignée d'internautes est à l'origine de la majorité des contributions (Brumfield, 2013)[32]

Ainsi on devrait finalement plutôt parler de *communitysourcing* (ou *community-sourcing*) ou de *nichesourcing* (ou *niche-sourcing*) plutôt que de *crowdsourcing* concernant ces projets de bibliothèques numériques participatives.

[32] Dans ce diagramme, chaque carré représente un contributeur. La taille de chaque carré est proportionnelle à la quantité de ses contributions. On observe ainsi qu'une minorité de bénévole est à l'origine de la majorité des contributions. Cette observation est vérifiée par l'ensemble des projets ayant recours au bénévolat.

Le *nichesourcing* pourrait même ainsi être l'avenir du *crowdsourcing* (De Boer, 2012). En recrutant des petites communautés d'experts amateurs avec une forte diversité de profils, de parcours, et de points de vues, on obtiendrait aussi des groupes capables de prendre de meilleures, de plus intelligentes et de plus sages décisions (Surowiecki, 2005). Avec le *communitysourcing*, au lieu de confier des microtâches ou des tâches atomiques et répétables à une foule sans visage, on développe des communautés de pratique et d'intérêt, on rassemble des pairs ayant une identité, des affinités et surtout des objectifs communs. Leurs échanges réguliers entre ses membres engendrent progressivement la confiance sociale et accroît la réputation de chacun.

Ainsi, le Rijksmuseum a constaté qu'il avait plutôt besoin d'amateurs, d'experts, d'autodidactes, et de professionnels retraités que d'une véritable foule d'internautes et s'est orienté vers le *communitysourcing*. (De Boer, 2012). De la même manière, le projet MarineLives (ML) fait ainsi appel, non à n'importe qui, mais recrute des participants qui acceptent de travailler au moins 3 heures par semaine et qui s'engagent pendant 14 semaines minimum (Dunn, 2012).

3.4.3- Le travail des professionnels autour de ces projets et le *community management*

Comme le souligne (Ellis, 2014), considérer le *crowdsourcing* comme une simple source de travail gratuit est une grave erreur. Ça revient un peu à oublier que lorsqu'on fait l'acquisition gratuite d'un animal de compagnie, il faudra le nourrir, l'entretenir, le promener, le soigner… Sans un véritable management de la foule, le résultat obtenu risque d'être dramatique. Le travail gratuit des internautes sera largement compensé par d'autres coûts. L'étude de (Oomen, 2010), qui rend compte de l'expérimentation de l'Institut néerlandais du son et de l'image avec le jeu Waisda? souligne également l'importance de soutenir et d'encadrer les bénévoles et regrette que ce facteur ait parfois été sous-estimé par les porteurs de projets. Un rapport publié par l'OCLC (Smith-Yoshimura, 2012), recommande ainsi le recrutement d'un *community manager*. Une enquête menée dans ce même

cadre montre quelles sont les activités des équipes qui encadrent des projets de *crowdsourcing* :

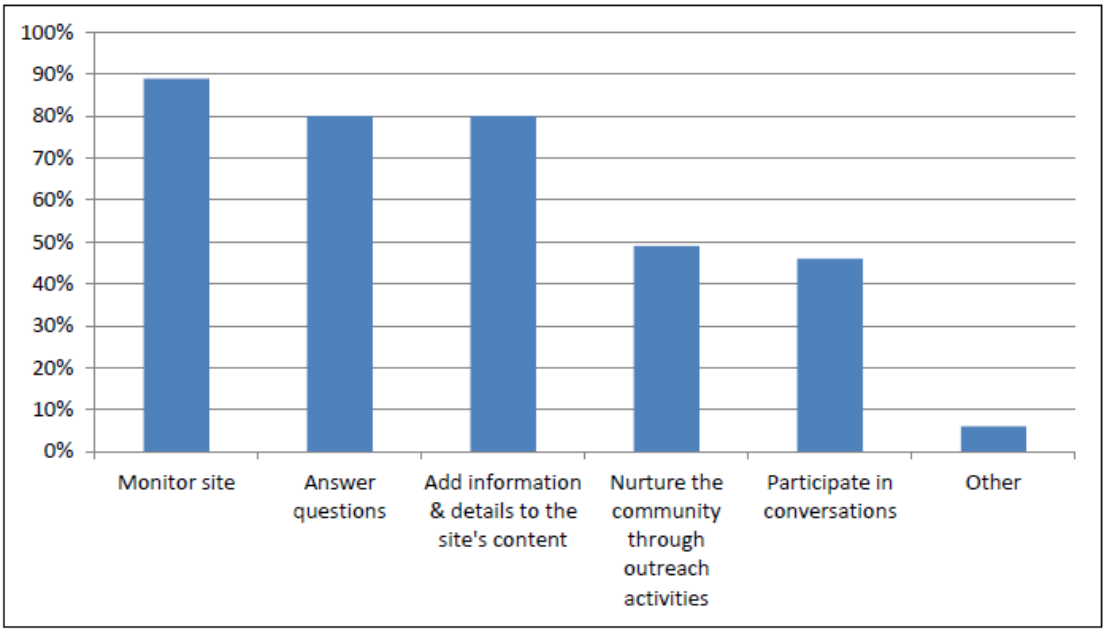


Figure 75. Répartition des activités du personnel manageant des projets de *crowdsourcing* d'après (Smith-Yoshimura, 2011)

Au delà de l'administration et du paramétrage de la plateforme, de la réponse aux questions des bénévoles, de la participation aux discussions, de l'ajout d'informations et d'actualités, de la formation des bénévoles, nous pourrions ajouter, afin de décrire une fiche de poste de *community manager* plus complète, la modération des contributions et des forums d'entraide, le contrôle de la qualité, la collecte de données statistiques, le développement de fonctionnalités, la réintégration des données produites, la gestion de projets, la communication, le recrutement, la motivation et la conservation des volontaires, la rédaction de blogs, de manuels, de guides, de tutoriels, de foires aux questions (FAQ), d'aides contextuelles, la définition de règles, la réalisation de vidéos de démonstrations screencast, le développement de "bacs à sable", la gestion d'une hotline, d'un helpdesk (Zastrow, 2014) et autres activités qui ont un coût qui risque d'atténuer, de compenser, voire de dépasser le travail gratuit récolté au travers du projet.

D’après l’enquête déjà mentionnée, ce sont souvent des professionnels qui consacrent une partie de leur temps au *community management* ou de professionnels à plein temps sur les projets de *crowdsourcing*, moins souvent des professionnels non spécialisés dans le domaine des technologies de l’information et encore moins souvent, des volontaires formés par des professionnels et qui participeront aussi à la rédaction des procédures, des règles et des manuels. Pour 23 % des répondants à cette enquête, les volontaires formés par des professionnels jouent un rôle dans le management du site web de *crowdsourcing*.

Afin d’aider les internautes à se former à s’autogérer, des outils peuvent être proposés aux contributeurs et rédigés également avec eux de manière collaborative. Parfois, les bénévoles peuvent même être chargés de modérer les contributions et de coordonner le travail des autres bénévoles grâce à des outils participatifs comme les wikis (Holley, 2009). Les bénévoles doivent pouvoir s’autoformer facilement avec l’aide de tutoriels, de vidéos screencast, de forums sur lesquels ils doivent pouvoir demander de l’aide. Les contributeurs ont tous les attributs de travailleurs à distance. Ils doivent, par conséquent, pouvoir se former à distance et les dispositifs classiques de travail à distance et de e learning (formation à distance) doivent donc être mobilisés pour eux.

Le temps que les professionnels consacrent aux plateformes semble être très variable d’une institution à l’autre comme le montre le diagramme suivant :

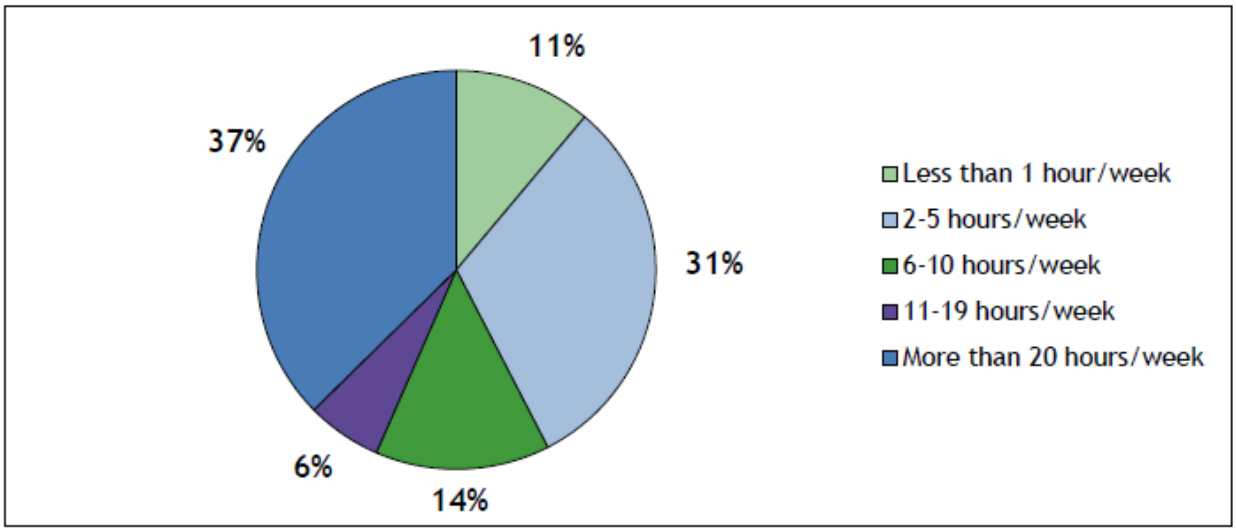


Figure 76. Le temps de travail du personnel des projets de *crowdsourcing* d'après (Smith-Yoshimura, 2011)

Le temps consacré aux projets de *crowdsourcing* se répartit de la manière suivante parmi les répondants à l'enquête :

Activity	Percentage of Time Spent										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Maintaining the site	3	12	7	4	1	0	0	0	2	1	0
Adding new content	1	9	4	3	3	3	2	2	0	2	0
Moderation	2	16	1	4	1	0	0	0	0	0	0
Incorporating user generated content	9	5	4	1	0	0	0	0	0	0	0
Adding new features or modifying the site's interface	5	4	6	3	2	0	0	0	0	0	0
Planning and administration	1	9	9	1	3	0	0	0	0	0	0

Tableau 18. La répartition du temps de travail du personnel des projets de *crowdsourcing* en fonction des activités et des missions d'après (Smith-Yoshimura, 2011)

Avec le développement du *crowdsourcing* en bibliothèque, le métier de bibliothécaire pourrait connaître une évolution et passer de catalogueur-indexeur à *community manager*.

Comme nous l'avons vu au chapitre des motivations, il peut exister une diversité importante de motivations différentes d'un individu à l'autre. Le *community manager* doit savoir jouer sur ces divers ressorts. Afin de fédérer les contributeurs dans leur diversité, et afin de susciter une adhésion, il est nécessaire que les internautes puissent facilement s'approprier le patrimoine numérisé, en ayant accès selon des critères classiques de métadonnées (titre, auteur, date, sujet, zone géographique...), ils doivent pouvoir facilement choisir le type de documents sur lesquels ils souhaitent contribuer (époques, thématiques, auteurs...) mais également avoir accès aux documents à traiter en fonction de leurs niveaux

de difficultés, par degré d'avancement, ou tout simplement par document au hasard.

La participation des internautes doit également être régulièrement entretenue en ajoutant périodiquement de nouveaux contenus à traiter, le site doit être éditorialisé afin d'accroître l'activité des bénévoles. Cette mise en ligne périodique de nouveaux contenus est préférable au fait de tout mettre à la fois, ce qui pourrait avoir pour effet, de décourager les bénévoles. Ainsi, 59 % des sites de *crowdsourcing* culturel étudiés dans le cadre de l'enquête de (Smith-Yoshimura, 2011) affirmaient effectuer une mise en ligne de nouveaux contenus au moins une fois par semaine :

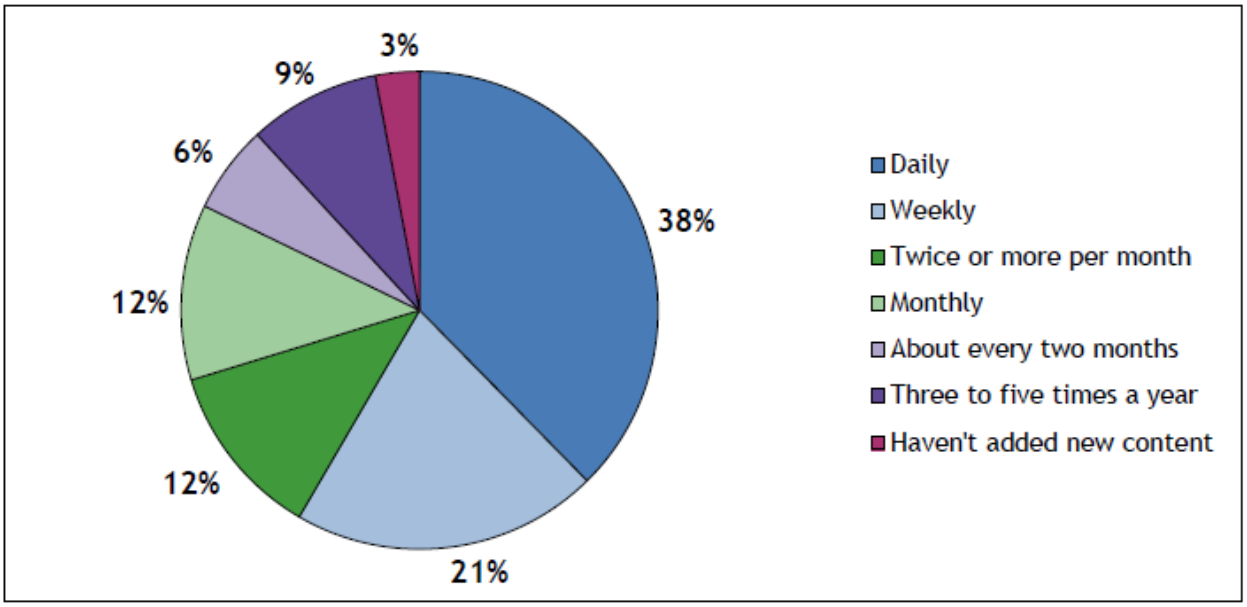


Figure 77. Fréquence avec laquelle les sites mettent en ligne du nouveau contenu d'après (Smith-Yoshimura, 2011)

Afin de mieux "entretenir la flamme", il faut également profiter des événements d'actualité, des occasions spéciales, des anniversaires historiques pour inciter les bénévoles à contribuer d'avantage.

Enfin, d'après (Moirez, 2013), certains projets proposeraient une organisation structurée et hiérarchisée des contributeurs en communautés. Ce serait le cas de COoperative eNgine for Correction of ExtRacted Text (CONCERT)

qui propose des activités en fonction de l'expérience et des compétences des internautes, de Monasterium qui fait intervenir des administrateurs experts, ou encore de Transcribe Bentham qui classe les contributeurs en fonction de la quantité de contributions. D'autres projets ne proposeraient pas de structuration de la communauté des contributeurs comme Ancient Lives, ArcHIVE, Digitalkoot, Do it yourself history, TROVE ou What's on menu?

3.5- La question de la qualité des contributions

Comme cela a été largement évoqué dans le chapitre conceptuel, une objection centrale des opposants à la mise en place du *crowdsourcing* porte sur la qualité des contributions.

La loi de Sturgeon prétend ainsi que 90 % des idées récoltées auprès de foules sont médiocres et que seules 10 % d'entre elles peuvent être de qualité équivalente à celles des spécialistes. (Roth, 2016). Pour certains, le travail gratuit serait nécessairement un travail de piètre qualité. D'ailleurs, outre atlantique également, dans le New York Times du 27 décembre 2010, des propos peu rassurants de Daniel Stowell, Directeur des archives Abraham Lincoln, sont rapportés par Patricia Cohen à ce sujet. Selon Daniel Stowell, le nombre d'erreurs présentes dans les contributions nécessiterait finalement de consacrer d'avantage de temps et d'argent à leur correction que dans le cadre d'un travail classiquement produit par des professionnels.

Nous verrons toutefois qu'il existe des systèmes permettant d'assurer une bonne qualité des contributions et d'en évaluer la qualité et qu'il existe des études qui ont cherché à comparer la qualité des données produites par les amateurs et par les professionnels. Enfin, nous aborderons le sujet de la réintégration des données produites par les internautes et le statut juridique de ces données.

3.5.1- Les systèmes d'évaluation et de modération des contributions

Dans le rapport commandé en 2012 par l'OCLC (Smith-Yoshimura, 2012) au sujet des métadonnées sociales pour les bibliothèques les archives et les

musées, il est recommandé de ne pas trop se soucier du spam ou du vandalisme, le Spam pouvant, par exemple, être filtré automatiquement avec l'aide de Captcha permettant de vérifier l'origine humaine des contributions. Il reste aussi possible de demander aux internautes de se loguer pour pouvoir contribuer afin de pouvoir limiter et repérer toute malveillance ou d'enregistrer leurs adresses IP et de demander aux internautes de surveiller eux mêmes la qualité des contributions comme le fait Wikipedia pour se prémunir du vandalisme en complément de l'utilisation de robots. Allant encore plus loin que le rapport de l'OCLC, (Holley, 2009) recommande de cesser de supposer que tout ira mal et de gaspiller un temps précieux à mettre des systèmes en place pour empêcher le vandalisme. Elle recommande de faire confiance aux internautes, de supposer qu'ils vont faire de leur mieux, et de leur donner un maximum de liberté. Elle observe que le projet australien TROVE est fondé sur ces principes de confiance et n'a jamais souffert de vandalisme. D'autres estiment que l'apparition du vandalisme dans un projet signifie qu'il est sorti de la phase du prototype pour rentrer dans une phase de maturité.

Mais, au delà de la simple question du vandalisme, celle de la qualité des contributions se pose et est d'ailleurs régulièrement mise en avant par ceux qui expriment une méfiance via à vis des dispositifs participatifs. Comme le souligne (Moirez, 2013), afin de garantir une qualité suffisante des contributions, il faut former les bénévoles, les assister, les évaluer, leur proposer des tâches en fonction de leurs compétences, confronter leurs contributions et en contrôler la qualité.

A partir du panorama des projets des projets, de la littérature (Quinn, 2011 et Kleka, 2014), mais aussi de blogs comme celui du développeur, Ben W. Brumfield (http://manuscripttranscription.blogspot.fr, consulté le 26 juin 2016), nous proposons la typologie suivante des systèmes de contrôle qualité des contributions :

Il existe également des systèmes de contrôle de la qualité :

- Pas de contrôle qualité et la confiance dans l'autorégulation des participants.

- Le contrôle qualité par des experts, des professionnels, des bibliothécaires, une méthode efficace mais très ingrate et très coûteuse.
    - La révision à durée déterminée ou indéterminée par des experts qui verrouillent et publient les contributions une fois un bon niveau de transcription atteint (exemples : Transcrive Bentham, Scripto, Do it yourself history, Monasterium, What's on menu)
    - L'insertion volontaire d'erreurs et de pièges dans des documents afin de vérifier que le contrôle qualité a bien été réalisé et afin de s'assurer de la vigilance et de la qualité du contributeur. Cette méthode a notamment été utilisée par un projet du panorama ayant eu recours à du *crowdsourcing* rémunéré pour de la transcription de manuscrits.
    - L'utilisation de tests de bienveillance. Utilisé par COoperative eNgine for Correction of ExtRacted Text (CONCERT)
- Le contrôle qualité par d'autres bénévoles, par la communauté de bénévoles, à la manière de Wikipédia ou en soumettant au vote les contributions.
    - Le contrôle qualité par division du travail. Dans la chaîne des microtâches, la tâche précédente est ainsi contrôlée et vérifiée par l'opérateur suivant. Le workflow de Wikisource utilise, par exemple, cette méthode, la même personne ne peut pas être en même temps, celle qui va valider une ligne et celle qui va valider une page. Celui qui corrige une page verra son travail évalué par celui qui valide une page qui verra son travail évalué par celui qui valide un texte.
    - Le contrôle qualité de bénévoles par d'autres bénévoles. Un groupe d'internautes vérifie le travail d'un premier groupe. Dans les marchés de numérisation, il arrive par exemple qu'un prestataire fasse le contrôle qualité d'une premier prestataire.
    - Le contrôle qualité par confrontation des saisies. Cette méthode est notamment utilisée par les sociétés qui travaillent dans la correction humaine de l'OCR, à Madagascar, par exemple. On fait transcrire le même texte par 2 opérateurs puis on confronte les différences de saisies afin d'obtenir un texte transcrit d'une qualité optimale. Le

procédé utilisé par certains projets de *crowdsourcing* est assez similaire. C'est l'un des méthodes les plus efficaces et des plus éprouvées afin de garantir la qualité des contributions. Au sein de cette méthode, plusieurs sous-méthodes existent :

- présenter le même mot à corriger à 2 contributeurs différents. En cas de divergence, c'est soit un expert ou un gros contributeur qui tranche, soit le mot est de nouveau présenté à 2 nouveaux contributeurs. C'est, par exemple, cette méthode qui est utilisée par le site de *gamification* Digitalkoot.
- présenter le même mot à corriger à plusieurs contributeurs différents (3 minimum). La majorité l'emporte. C'est, par exemple, cette méthode qui est utilisée par reCAPTCHA pour le projet Google Books.
- Le contrôle qualité par des moteurs, des algorithmes, des méthodes de filtrages statistiques. Cette méthode est notamment utilisée par Wikipédia.

Concrètement, d'après l'enquête conduite par l'OCLC (Smith-Yoshimura, 2011), 75 % des répondants (27 sur 36) affirment modérer les contributions. 36 % des répondants approuvent chaque contribution avant de la poster et 50 % des répondants peuvent éditer les contributions. Ce résultat a relativement surpris les experts chargés de conduire cette étude. En effet, cette activité de contrôle systématique peut être très chronophage et coûteuse. Par ailleurs, 53 % des répondants affirment que leur site web nécessite une identification, 36 % utilisent un système de type Captcha, 36 % utilisent l'adresse mail du contributeur. Dans seulement 31 % des cas, aucune identification n'est requise. Cela étant, pour seulement 6 % des répondants, le spam pose problème, pour 69 % ce problème n'est pas rencontré et pour 25 % occasionnellement. Et seulement 36 % des répondants (13 réponses), ont déjà été confrontés à des utilisateurs malveillants cherchant à ajouter des contributions inappropriées.

Parmi les stratégies utilisées par les institutions pour garantir la qualité des contributions, l'enquête de l'OCLC a identifié les suivantes :

- L'Institution conserve le droit de modifier, réutiliser, ou supprimer le contenu généré par l'utilisateur sans préavis (57 %)
- Les utilisateurs qui violent la politique peuvent être bloqués à partir du site (31 %)
- La propriété du contenu généré par l'utilisateur est conservée par le site / institution (31 %)
- Une charte présentant les lignes directrices du projet et son mode de fonctionnement (14 %)
- Pas de politique spécifique (11 %)
- Les utilisateurs de confiance peuvent contribuer sans modération (liste blanche) (9%)

D'après (Moirez, 2013), certains sites nécessitent une authentification obligatoire et préalable afin de pouvoir contribuer comme COoperative eNgine for Correction of ExtRacted Text (CONCERT), Monasterium, Transcribe Bentham et Digitalkoot (authentification via Facebook) tandis que pour d'autres l'authentification est facultative comme Do it yourself history, TROVE, What's on menu ou Wikisource. Comme le souligne également le directeur des Archives départementales du Cantal dans son article (Bouyé, 2012), certains Archives laissent une totale liberté aux internautes souhaitant contribuer et qui n'ont même pas besoin de s'inscrire, tandis que d'autres encadrent, contrôlent et font même passer des tests préalables de paléographie. C'est cette première voie qui a été choisie dans le Cantal.

Les documents sont parfois aussi classés selon leur difficulté et attribués selon les compétences des contributeurs qui a, elle-même, été évaluée. Des épreuves d'évaluation sont parfois même soumises aux volontaires afin de déterminer leur niveau et leur attribuer des tâches adaptées à leurs facultés.

Au delà de la classique validation par des experts, l'enregistrement de l'historique de chaque modification et la possible restauration d'une version antérieure est un moyen utile pour assurer la qualité des contribution et éviter le

vandalisme. C'est ainsi que fonctionnent, par exemple, TROVE ou Wikisource. Les internautes doivent également pouvoir facilement signaler des erreurs. Google Books permet, par exemple, ainsi à ses usagers de signaler que tel ou tel document a mal été numérisé.

Une autre manière efficace de garantir la qualité du travail dans le cadre de *crowdsourcing* rémunéré, en particulier, peut être de concevoir des tâches de sorte qu'il n'est pas plus facile de tricher que de faire réellement la tâche ou de jouer sur la réputation des contributeurs dont le travail est évalué à la manière des vendeurs ebay (Quinn, 2011). Ainsi, sur l'Amazon Mechanical Turk Marketplace, les statistiques des travailleurs sont visibles et il est possible de connaître le nombre de tâches validées et rejetées de chacun. Il est également possible de sanctionner des internautes malveillants en les bannissant (compte bloqué, contributions supprimées), ce qui peut nuire à leur e-*reputation* et même avoir une incidence sur leur vie sociale et professionnelle (Dunn, 2012). On parle ainsi de "public shaming".

Enfin, (Eickhoff, 2012) observe que les plateforme de *crowdsourcing* rémunéré seraient infestées de travailleurs malintentionnés qui tentent de maximiser leurs profits grâce à la tricherie. Il observe qu'ils sont particulièrement répandus chez certaines nationalités et suggère donc de restreindre l'offre de travail à certaines autres nationalités parmi lesquels les tricheurs seraient moins nombreux. Mais cela pourrait toutefois poser un problème à la fois éthique et peut être également juridique.

3.5.2- Comparaison entre la qualité des données produites par les amateurs et celles produites par les professionnels

Les théoriciens de la sagesse des foules, comme James Surowiecki, estiment, comme nous l'avons vu, dans le chapitre conceptuel, que la diversité des profils contenus dans une foule avait tendance à donner de bien meilleurs résultats que l'avis des meilleurs spécialistes d'un domaine pour ce qui concerne les décisions. La "loi des grands nombres" permettrait en tous cas, dans le domaine des sciences citoyennes, de neutraliser les erreurs individuelles dans la masse des données justes apportées par les foules (Boeuf, 2012). On connaît aussi les

études mises en avant par Wikipédia[33] et qui concluent à une qualité égale ou supérieure de l'encyclopédie participative par rapport aux encyclopédies classiques à comité de lecture restreint et qui ont finalement plus de chances de laisser passer des erreurs que sur une encyclopédie où le monde entier peut les corriger. Dans le domaine des sciences participatives, les joueurs du jeu fold.it, pourraient également, dans bien des cas produire de meilleures résultats concernant les protéines que le programme Rosetta selon certains auteurs (Good, 2011). Malgré ces arguments et malgré tous les dispositifs que nous venons d'évoquer dans le chapitre précédent et qui permettent de garantir la qualité des données produites, la comparaison du résultat obtenu par les amateurs avec celui des professionnels peut être légitime. Elle a d'ailleurs fait l'objet de multiples études.

Ainsi, dans une étude universitaire (Thogersen, 2012), on a cherché à comparer la qualité des indexations d'images obtenues par des amateurs via des mécanismes de *crowdsourcing* sous forme de *gamification* et par des professionnels via un fonctionnement plus traditionnel. Il semble que le *crowdsourcing* aurait tendance à privilégier, de manière subjective, le contenu, c'est à dire ce qui est représenté dans l'image tandis que les professionnels seraient plus objectivement attachés à la forme, aux objets. Une autre étude menée par J. Trant en 2009, et rapportée par (Paraschakis, 2013) et par (Ridge, 2011), révèle que, sur un échantillon de 36 981 termes proposés par les internautes du projet steve.museum, 86 % d'entre eux sont différents des vocabulaires contrôlés et des thesaurii employés par les professionnels, 70,2 % correspondent partiellement à des termes du Art and Architecture Thesaurus (AAT) en particulier et 88,2 % de ces 36 981 termes ont été estimés comme utiles par ces mêmes professionnels. Une expérimentation au Rijksmuseum d'Amsterdam (Oosterman, 2014) qui comparait le travail de corrections, de traductions, et d'identifications d'espèces de fleurs entre des experts et des crowdworkers a également conclue à la pertinence du recours au *crowdsourcing* rémunéré pour ce type de tâches. Une étude comparative (Rorissa, 2010) sur 4

[33] Giles, Jim (2005). Internet encyclopaedias go head to head. Nature , 438, p. 900-901

441 tags sur 1000 photos Flickr comparés à 3 709 descripteurs sur 996 photos de l'archive photographique de la Bibliothèque de l'Université de St. Andrews conclue, elle aussi, à la complémentarité de l'approche professionnelle et de l'approche folksonomique. Cette étude incite les professionnels à s'inspirer de la richesse du vocabulaire utilisé par les internautes dans la construction des thesaurii. Elle affirme qu'un professionnel indexe généralement un document sur un sujet qu'il maîtrise mal, avec l'aide d'un thésaurus d'un usage complexe, et qu'il cherche à le faire afin de permettre à un usager de retrouver une information dans une démarche hiérarchique, de haut en bas. Au contraire, dans le cas d'une indexation libre ou du tagging, l'usager décrit très simplement un document dont il connaît généralement bien le sujet puisque sa navigation ne l'a généralement pas amené à le consulter par hasard et il le tague d'avantage pour son propre intérêt dans une démarche bottom up, de bas en haut. Si les termes qu'il utilise sont susceptibles d'être ambigus, polysémiques, synonymes et moins précis, ils risquent cependant d'être plus riches et plus proches des mots clés saisis dans le cadre d'une recherche. Au delà de la simple production de données, (Blasco, 2013) a cherché à comparer les résultats entre une équipe de développeurs organisée de manière traditionnelle et une équipe de développeurs auto-organisée parmi la foule des développeurs dans le cadre d'une compétition avec gratification financière entre les équipes. Les compétiteurs issus de la foule des développeurs auraient proposé plus de solutions, plus de solutions qui fonctionnent, et des solutions de meilleure qualité.

A la différence de ces études qui concluent à l'intérêt de recourir au travail des amateurs, d'autres études sont plus nuancées. Ainsi, (Bar-Ilan, 2008) a cherché à comparer l'indexation libre issue du tagging amateur avec une indexation structurée réalisée par des professionnels à partir de thesaurii. 47 étudiants en sciences de l'information ont participé à l'expérience. Ils ont été divisés en 2 groupes. Un groupe pour l'indexation libre et un groupe pour l'indexation structurée à partir de champs à compléter. Les mêmes images leur ont été données pour être indexées. Des informations plus détaillées ont été obtenues par le second groupe qui structurait l'information sous forme de champs. Cette

expérience prouve néanmoins surtout que le fait de saisir des métadonnées dans des champs permet de ne pas oublier des types de métadonnées. Par contre, la qualité des mots clés issus d'indexation libre et d'indexation structurée n'a pas été comparée par l'étude.

Une autre étude apparaît aussi comme relativement nuancée (Snow, 2008). Ainsi, la qualité des annotations en langage naturel proposées par des professionnels et par des amateurs de l'Amazon Mechanical Turk marketplace a été comparée. Selon les auteurs, les annotations obtenues via les amateurs sont plus abondantes mais aussi plus bruyantes et moins pertinentes que celles produites par des experts. Individuellement, la qualité des contributions sont évidemment meilleures du coté des experts. Par contre, en confrontant le travail de 4 amateurs en moyenne (2 minimum et 9 maximum), on obtiendrait une qualité assez similaire.

Enfin, une étude est nettement moins enthousiaste concernant la qualité des données produites par les amateurs. (Oomen, 2010) s'est penché sur la qualité de l'indexation produite pour les documents audiovisuels de l'Institut néerlandais du son et de l'image via le jeu Waisda?. Il en ressort que seulement 5,8 % des tags ont une occurrence dans le thésaurus de l'institut et que seulement 23,6 % d'entre eux sont présents dans la base Cornetto des mots de la langue néerlandaise. De la même manière, concernant la qualité des tags sur Flickr The Commons, une étude de (Guy & Tonkin, 2006) et rapportée par Earle estime, que sur un échantillon, seuls 40 % des tags avaient une occurrence dans le dictionnaire Open Source Aspell.

A la lumière de toutes ces études dont les résultats demeurent contradictoires, il reste donc difficile d'arrêter un point de vue définitif sur le sujet.

3.5.3- La réintégration des données produites

En fonction de la qualité des données récoltées, on décidera ou non d'intégrer les données produites par les amateurs dans le système d'information, catalogue ou bibliothèque numérique. Il existe cependant, en la matière, deux philosophies. Les institutions qui recherchent simplement à engager le public autour de leurs collections ou à améliorer leur image dans le cadre d'une

communication institutionnelle autour d'un sujet à la mode. Ces institutions auront tendance à ne pas exploiter les données produites par les internautes. D'autres institutions ont réellement besoin de l'aide des internautes. Celles-ci auront donc d'avantage tendance à utiliser le travail qui leur a généreusement été offert.

Dans le rapport commandé par l'OCLC, (Smith-Yoshimura, 2012), le rédacteur relate que, à la suite d'une enquête menée par l'OCLC sur 76 sites, il s'avère que seule la moitié des sites utilisant de *crowdsourcing* réutilisaient les tags produits par les internautes et qu'un peu plus du tiers des répondants affirmaient ne pas indexer les métadonnées produites par *crowdsourcing*. Ce résultat, à première vue surprenant, provient bien du fait que le *crowdsourcing* n'a pas vraiment été envisagé comme un moyen d'externaliser des tâches ou encore que la qualité des contributions reste dépréciée par la profession.

Table 7: Uses of social metadata (n=36)

Survey Question (Section 8)	Yes	No
Are you concerned with how the content of your site is used or repurposed?	28%	72%
Have you incorporated metadata (including tagging) created by users into your own metadata and description workflow?	39%	61%
Do you incorporate other user-contributed content (e.g., photographs, documents) into your site?	44%	56%
Does your system index user-supplied metadata?	61%	39%
Do you perform any spell-checking of user content or de-duping of tags submitted by users (e.g., differences in capitalization or spelling, singular vs. plural, etc.)?	19%	81%

Tableau 19. Utilisation faite par les institutions culturelles des métadonnées sociales d'après l'étude OCLC (Smith-Yoshimura, 2011)

(Stiller, 2014) estime que les données produites par les utilisateurs sont précieuses et doivent être valorisées au même titre que les données produites par les institutions. Il constate aussi que les professionnels demeurent encore réticents à accepter ces contributions qui restent souvent séparées du contenu institutionnel par crainte d'une dévaluation de leurs contenus ou par crainte d'abus de la part des utilisateurs. Selon l'auteur, ces appréhensions sont souvent arbitraires, surtout si une communauté de travail existe qui peut surveiller le contenu produit.

Pour finir, l'exemple de la Bibliothèque nationale de France qui a eu le plus grand mal afin de réintégrer dans Gallica, les textes corrigés par les bénévoles de Wikisource illustre bien les difficultés auxquelles s'expose un projet qui aurait insuffisamment envisagé la réutilisation des données produites par les internautes. En effet, la BnF a finalement été contrainte de faire développer des fichiers XML ALTO afin de géoréférencer chaque mots des textes corrigés par les internautes, ce qui aura probablement finalement été d'avantage coûteux qu'une simple externalisation du travail de correction de l'OCR à Madagascar.

3.5.4- Le statut juridique des contributions : *crowdsourcing* et web sémantique

Dans une intéressante publication, (Djupdahl, 2013), l'expérience du projet YEAH! de *crowdsourcing* appliqué au domaine des archives et d'utilisation des technologies du web sémantique et du Linked Open Data est relatée. Le YEAH project est financé par la Swedish Governmental Agency for Innovation Systems (VINNOVA), en partenariat avec NordForsk, the Icelandic Centre for Research (RANNIS), et le Estonian Ministry for Economic Affairs and Communication. Dans cette publication, les auteurs rappellent notamment que les contributions produites gratuitement par des volontaires ne sauraient moralement être interdites à la réutilisation et verrouillés avec des licences restreignantes. En toute logique, les données produites bénévolement doivent pouvoir, librement et sans restriction, être réutilisées par tous sur le web y compris par d'autres systèmes d'information, via des technologies linked Open Data. Cette réutilisation possible est d'ailleurs un argument supplémentaire pour les projets afin de stimuler les contributions.

Dans tous les cas, au moment où les volontaires créent un compte sur le site, il est nécessaire de leur faire approuver un contrat énonçant quel sera le statut des données qu'ils vont produire et leur licence de diffusion.

3.6- L'évaluation des projets de *crowdsourcing*

D'après l'enquête conduite par l'OCLC (Smith-Yoshimura, 2011), 91 % (30 réponses) des répondants estiment que leur projet de *crowdsourcing* est un succès. Cette même enquête identifie les critères de succès tels que perçus par les porteurs de projets :

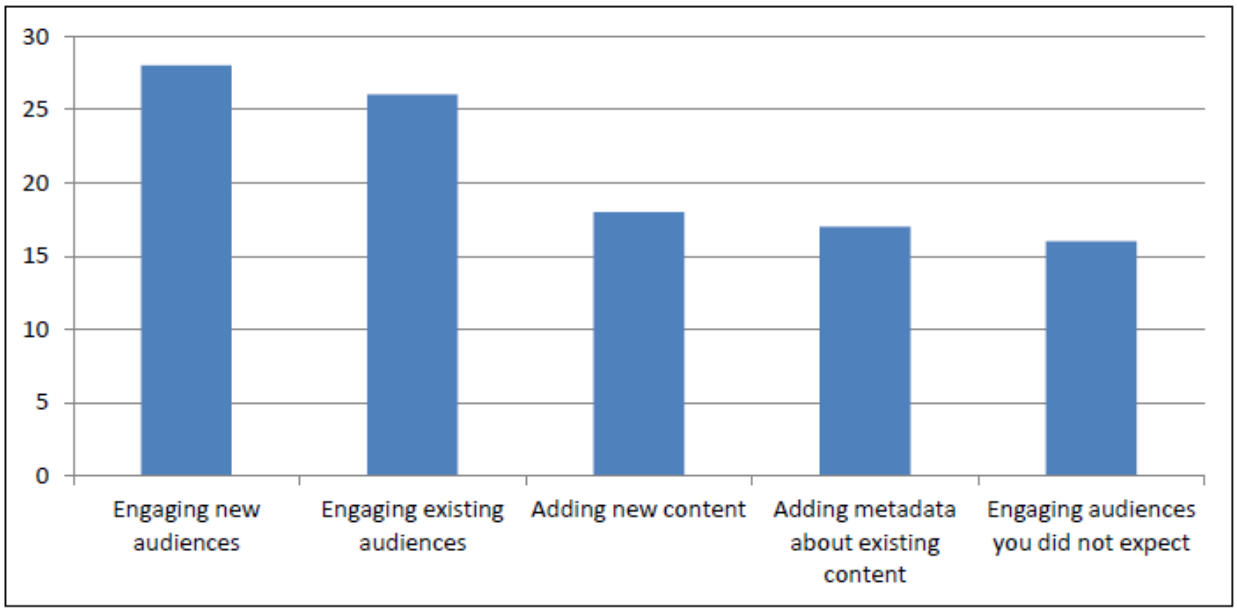


Figure 78. Les critères de succès d'après (Smith-Yoshimura, 2011)

En tête, on trouve le fait d'engager de nouveaux publics et des publics existants, devant même le fait d'obtenir de nouveaux contenus ou d'ajouter des métadonnées aux contenus déjà existants. Il semble donc que la philosophie qui consiste à lancer des projets de *crowdsourcing* afin de changer la relation avec le public et d'améliorer l'image de l'institution soit dominante au détriment de l'utilité réelle du travail des volontaires et de vraies finalités pour les projets. Il apparaît donc nécessaire d'évaluer les projets de manière qualitative mais aussi quantitative, un exercice auquel les institutions ne se prêtent malheureusement

pas toujours. Pour cela, les outils comme Google Analytics, les enquêtes et les entrevues sont indispensables (Birchall, 2012).

Le *crowdsourcing* semble avoir un impact sur le trafic web des bibliothèques numériques. Ainsi, Nicole Saylor, responsable de la bibliothèque numérique de l'Université de l'Iowa a rapporté que, grâce au *crowdsourcing*, au 9 juin 2011, la bibliothèque numérique était passée de 1000 hits maximum par jour à plus de 70 000 hits[34].

Website traffic at CDNC before / after implementing crowdsourcing

	before crowdsourcing 11-Jun-2011 / 12-Jul-2011	after crowdsourcing 11-Jun-2012 / 12-Jul-2012	change
visits	17,485	21,488	+22.9%
unique visitors	11,381	13,376	+17.5%
visit duration	9m 24s	11m 7s	+18.3%
bounce rate	51.3%	44.5%	-6.8%
pages per visit	14.9	11.7	-21.5%

Tableau 20. Statistiques avant et après *crowdsourcing* pour la California Digital Newspaper Collection (d'après Geiger, 2012)

[34]L'utilisation du hit afin de mesurer le trafic web d'un site ne nous semble pas être l'indicateur le plus pertinent dans la mesure où le nombre d'objets par page web peut biaiser les résultats et donner l'impression qu'une page web peu consultée génère finalement plus de trafic web qu'une page beaucoup plus consultée mais possédant moins de documents à afficher. Néanmoins si on compare le même site web à deux moments différents, en dehors de changement importants de contenus sur ce site, on peut effectivement conclure à une forte augmentation de son trafic web comme le fait Nicole Saylor.

Au delà de ce simple exemple, d'après l'enquête de l'OCLC (Smith-Yoshimura, 2011), le nombre de visiteurs uniques mensuels déclarés par les institutions culturelles se situe de la manière suivante :

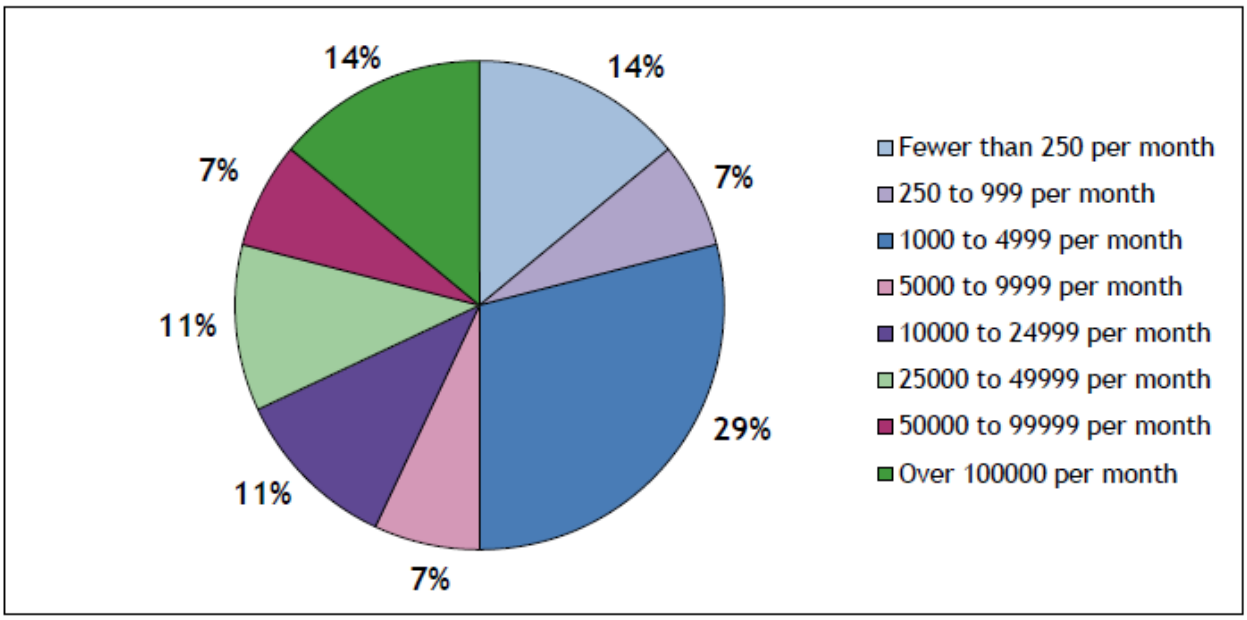


Figure 79. Nombre de visiteurs uniques par mois pour les projets de *crowdsourcing* d'après (Smith-Yoshimura, 2011)

Et d'après cette même enquête, le nombre de visiteurs contributeurs déclarés par les institutions culturelles est le suivant :

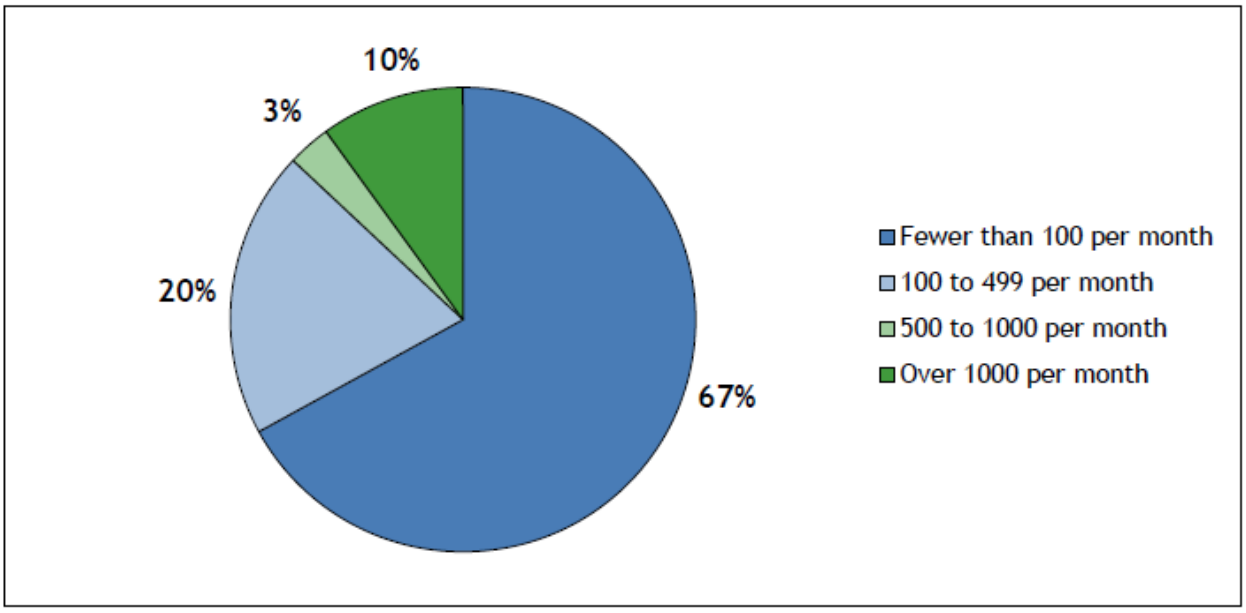


Figure 80. Nombre de contributeurs par mois pour les institutions culturelles d'après (Smith-Yoshimura, 2011)

3.6.1- les facteurs de réussite et d'échecs

En s'inspirant de (Sharma, 2010), de la thèse de (Brokfeld, 2012) et des études de (McKinley, 2015) qui se sont intéressés à la question des facteurs de réussites des projets de *crowdsourcing* et sans revenir à tous les éléments déjà évoqués, on peut énumérer les clés du succès suivantes :

- Une vision, des objectifs, des finalités, un challenge clairement définis, énoncés dès la page d'accueil et facilement compris par des contributeurs potentiels. Pour reprendre la métaphore déjà employée précédemment, le tailleur de pierre doit savoir à quoi serviront ses efforts, il doit savoir qu'ils serviront à bâtir une cathédrale. Il ne doit pas être utilisé comme un simple moyen n'ayant pas à connaître le contexte et la finalité du travail qui lui est demandé. Il doit être convaincu que son travail est indispensable et ne saurait être automatisé. Les instructions de travail doivent être claires et efficaces.
- Une communication efficace afin de recruter des bénévoles en explicitant les diverses raisons de contribuer et en jouant sur différents types de motivations. Le fait d'être au service de l'intérêt général est un avantage pour les institutions culturelles publiques. Pour être crédible et apparaître comme sincère aux yeux des

internautes, le projet doit afficher des sponsors reconnus, montrer l'équipe du projet, ses résultats, le nombre de contributeurs. Les données produites par les internautes doivent être librement accessibles et réutilisables par le public.

- Un capital humain disponible et motivé. Comme nous l'avons vu au chapitre des motivations, la motivation peut être stimulée par un tableau des résultats, par un diagramme de la progression vers la réalisation du projet, par un classement des plus gros contributeurs, par des remerciements et par des récompenses...
- Une infrastructure susceptible de recevoir le travail fourni avec un environnement de travail intuitif, ergonomique et fiable. Les utilisateurs doivent pouvoir se loguer directement à partir de leurs comptes Facebook ou Google+ et inviter facilement à participer les membres de leurs réseaux sociaux.
- Des activités faciles, amusantes, intéressantes, formatrices. De nouveaux contenus sur lesquels travailler régulièrement ajoutés. Les nouveaux contributeurs doivent pouvoir s'entraîner sur un "bac à sable". Ils doivent ensuite pouvoir choisir ce sur quoi ils vont travailler selon leurs centres d'intérêts, leurs niveaux d'expertises et le temps qu'ils peuvent consacrer au projet.
- La confiance et le lien social. Les utilisateurs doivent pouvoir interagir et former une communauté. Il faut donner du pouvoir de décision aux bénévoles, leur faire confiance et susciter leur confiance. La plupart des projets qui ont réussi ont considéré les internautes comme des partenaires du projet et non comme une source de travail gratuit. Ils ont réellement utilisé et réintégré les données produites.
- Une conduite du changement en interne et un environnement extérieur favorable (situation économique et sociale)

A l'inverse, les projets qui échouent font généralement appel à des tâches trop complexes, à des connaissances trop spécialisées, nécessitent de suivre des tutoriels ou des formations préalables difficiles, sont trop vagues dans leurs objectifs, ne donnent pas assez de retours aux contributeurs, ne leur montrent pas quelle utilisation est faite des données produites (Ridge, 2013).

Plus généralement, les facteurs d'échecs d'un projet de *crowdsourcing* sont globalement les mêmes que pour toute externalisation avec la dilution de la responsabilité. Avec l'externalisation, les conséquences des décisions se faisant moins sentir, cela peut générer une perte d'attention, de contrôle, et un envol des délais. Le principal risque du *crowdsourcing* serait ainsi qu'une société faisant appel à ce type d'externalisation surestime ses capacités à en assurer le management et que celui-ci lui finisse par lui échapper complètement. Elle pourrait aussi surévaluer les profits qu'elle pourrait tirer du *crowdsourcing*.

Le recours aux foules peut aussi générer de vives tensions parmi le personnel en interne. En effet, les plus sceptiques ont tendance à considérer que de permettre au profane d'indexer, c'est ouvrir le patrimoine au courrier des fans, aux faux souvenirs, à la manipulation des mémoires voire à l'incivilité. La perte de contrôle du patrimoine est également un objet de craintes, notamment pour ce qui concerne leur réutilisation dans des cadres non éthiques ou leur réutilisation commerciale. Le fait d'être submergé de commentaires auxquels il faudra répondre peut aussi être un sujet de préoccupation pour les bibliothécaires. Ainsi, d'après une enquête suisse auprès des bibliothèques (Estermann, 2014), 72 % des sondés estiment que la mise en place du *crowdsourcing* nécessitera un temps considérable pour le suivi et que ce temps est difficile à estimer (70 %) pour des résultats imprévisibles (61 %). Enfin, le *crowdsourcing* peut aussi être considéré comme une remise en question du travail d'indexation des professionnels, si le travail est réalisé gratuitement par des bénévoles.

3.6.2- Évaluation quantitative des projets de *crowdsourcing* et de leurs coûts

De la même manière qu'il est essentiel, pour une organisation, de savoir si l'externalisation d'une activité est moins coûteuse et plus productive que si elle est maintenue en interne, son externalisation auprès de foules d'internautes doit être comparée avec son externalisation auprès d'un sous-traitant plus classique et avec son automatisation. On comparera, en particulier, les coûts du *crowdsourcing* en communication, gestion de projet et *community management*, contrôle qualité, récompenses ou rémunération avec les coûts tels qu'ils existent quand ils sont pris en charge en interne. Mais, selon (Lebraty, 2015), le *crowdsourcing* serait

“particulièrement compétitif en termes de coûts” et “dans tous les cas [...] moins cher qu’une autre forme d’externalisation”.

Voici des indicateurs d’évaluation quantitative que nous proposons pour chaque objectif que peut se donner une institution qui se lance dans un projet de *crowdsourcing*, avec des données pour des projets lorsque nous sommes parvenus à les collecter dans le cadre du diaporama des projets :

1- Corriger l’OCR ou transcrire les manuscrits	Nombre de mots corrigés ou transcrits par mois	Nombre de lignes corrigées ou transcrites par mois
reCAPTCHA	86 millions de mots validés par jour, soit **2,58 milliards de mots validés par mois**	inconnu
TROVE	inconnu	**2 724 671 lignes corrigées** (sur le mois d’avril 2015)
Digitalkoot	6 461 659 mots corrigés en 1 an, en février 2012 soit **538 472 mots corrigés par mois**	inconnu
Transcribe Bentham	Du 28 janvier 2012 au 2 novembre 2012, en moyenne 51 manuscrits (25 500 mots) par semaine, soit **102 000 mots par mois**	inconnu
California Digital Newspaper Collection (CDNC)	inconnu	578 000 lignes en 2012, soit **48 167 lignes par mois**

2- Indexer les contenus	Nombre de tags apportés par mois
TROVE	**68 167 tags** (sur le mois d'avril 2015)
Flickr: The commons	2 millions de tags en 5 ans soit **33 333 tags par mois**
steve.museum	468 120 tags entre mars 2007 et fin 2010 soit 468 120 / 46 = **10 175, 5 tags par mois**

3- Faire financer la numérisation du fonds par les internautes grâce à un service *crowdfunding* de numérisation à la demande	Nombre de livres numérisés par mois	Argent récolté par mois
Ebooks on Demand en 2011 (27 bibliothèques européennes)	1781 commandes sur l'année 2011 **= 148, 5 livres par mois**	75 512, 3 € sur l'année 2011 = **6542, 69 € par mois**
Le livre à la carte, Phénix éditions	Une **centaine de commandes par mois** début 2011 à la Bibliothèque Municipale de Troyes	inconnu
Adopter un livre sur Gallica	**7,64 numérisations par mois**	inconnu
Numalire en 2014 (8 bibliothèques parisiennes)	36 livres sur 8 mois = **4,5 livres par mois**	inconnu

Au delà des données que nous avons récoltées au cours du diaporama des projets, il serait intéressant de produire les indicateurs suivants :

4- Convertir des consommateurs passifs de la bibliothèque numérique en producteurs actifs	Rapport entre le nombre de visiteurs et le nombre de contributions
5- Augmenter la fréquentation de la bibliothèque numérique	Nombre de visites par mois et par document
6- Faire de la communication institutionnelle	Nombre d'articles évoquant le projet dans la presse locale, nationale, internationale, les revues scientifiques, les blogs, les réseaux sociaux...
7- Diminuer les coûts	Évaluer à partir d'un salaire horaire donné, l'argent que la foule a apporté au projet en temps de travail.

Les sciences citoyennes sont coûteuses en temps et en énergie afin de manager les bénévoles, les sites web et les bases de données. Pourtant, la question des coûts de développement des projets de *crowdsourcing* n'est que très rarement évoquée dans la littérature. Comme le souligne également (Sagot, 2011), le développement de plateformes et le contrôle de la qualité n'est que rarement évalué dans les études. Le *crowdsourcing* appliqué au patrimoine culturel demeure dans une phase expérimentale et les initiatives ne sont pas toujours rentables. (McKinley, 2015)

L'une des rares publications à mentionner cet aspect est celle de (Causer, 2012). Cet auteur donne des indications précises de coûts pour développer le projet Transcribe Bentham (262 673 livres sterling, 2 Research Associates à plein temps, 1 curator, 1 consultant). Il estime aussi que les 2 Research Associates à plein temps auraient finalement pu transcrire 5000 manuscrits à plein temps sur 12 mois soit d'avantage que ce que les internautes ont produit. Il considère ainsi que, en 2012, le projet Transcribe Bentham aurait permis de ne pas dépenser l'équivalent de 6 mois d'un éditeur à plein temps. Mais, comme ce même auteur le rapporte, les éditeurs des Papers of Abraham Lincoln ont finalement estimé qu'ils

passaient plus de temps à corriger les corrections qu'à effectuer les transcriptions eux-mêmes.

Pour tous ces projets, il serait très intéressant de comparer les coûts (main d'œuvre, développements, hébergement, conduite de projet, communication....) nécessaires à leurs développements avec les "bénéfices" qu'ils ont générés en convertissant le temps de travail fourni en valeur.

Dans le chapitre consacré à reCAPTCHA, nous avions estimé à 146 millions d'euros par an ce que Google n'avait pas à dépenser chaque année grâce au travail involontaire des internautes. En complétant ces calculs avec ceux de (Ipeirotis 2011), de (Geiger et Zarndt 2012) et de (Zarndt 2014), nous estimons que les projets suivants n'ont pas eu à dépenser, pour la correction humaine de l'OCR (s'ils avaient fait appel à des prestataires) les coûts suivants :

Projet	**Type de *crowdsourcing***	**Coût non dépensé**
California Digital Newspaper Collection (fin 2011-)	*Crowdsourcing* explicite	53 130 $ cumulés en juin 2014
TROVE (août 2008-)	*Crowdsourcing* explicite	2 580 926 $ cumulés en mai 2014
Digitalkoot (février 2011-)	*Gamification*	Entre 31 000 et 55 000 € cumulés en octobre 2012
Google Books et reCAPTCHA (2008-)	*Crowdsourcing* implicite	146 millions d'euros par an (au rythme de 2008)

Calcul de ce que la correction de l'OCR aurait coûté sans le recours au *crowdsourcing* pour quelques projets représentatifs (d'après Andro, 2015, 2)

Ces coûts non dépensés mériteraient toutefois d'être comparés aux coûts qui ont été consentis en développement, administration de plateformes, communication, *community management* et réintégration des données produites par les bénévoles.

Concernant le projet de numérisation à la demande par *crowdfunding* Numalire, nous avons constaté que les bibliothèques avaient consacré 207 heures d'un travail assez ingrat (descriptions matérielles de documents afin de produire des devis) au projet pendant les 8 mois de l'expérimentation et que l'on pouvait considérer que cela leur avait coûter 6210 € en temps de travail pour seulement 36 livres dont la numérisation a été numérisée, soit une moyenne de 172 € par livre. Dans ce cas, les bénéfices du projet compensent donc difficilement ses coûts.

3.7- La conduite du changement

Tout changement dans la manière de travailler a tendance à générer une résistance et une volonté presque naturelle de conserver l'ancien mode de travail. Pour les bibliothèques, le *crowdsourcing* est un changement majeur puisque les tâches qui étaient auparavant prises en charge par des professionnels vont pouvoir l'être par des amateurs dont les activités vont changer.

La résistance au changement peut prendre les formes suivantes :

- L'inertie, la procrastination, la résistance larvée
- La résistance argumentée
- L'action contre le changement
- Le sabotage ou l'excès de zèle

Dans les bibliothèques ces types de réticences sont à prévoir :

- Peur d'avoir une charge de travail trop importante pour assurer le contrôle de la qualité des données produites par les internautes ou, dans le cadre de numérisation à la demande, pour communiquer les documents au prestataire et effectuer les constats d'états avant et après numérisation.
- Hostilité idéologique et culturelle vis à vis du privé, des amateurs, des délégations de service publique (refus de sous-traiter des services publics par des entreprises privées) et du mécénat (très peu pratiqué en bibliothèque).
- Incrédulité : personne n'acceptera de travailler gratuitement pour nous ou de participer au financement de la numérisation des livres, d'autant que la numérisation d'un livre est encore très coûteuse.

Ces réactions de résistance au changement sont naturelles car tout individu a peur de quitter ce qu'il connaît pour ce qu'il ignore encore mais elle peut être très variable d'un individu à l'autre. Un changement de SIGB pourra être perçu comme un changement faible pour un informaticien alors qu'il pourra être perçu comme très difficile pour un bibliothécaire contraint de changer ses habitudes et d'apprendre à travailler avec une nouvelle interface.

La durée et l'intensité de la résistance au changement diminuent avec une meilleure compréhension du projet et une participation à sa mise en œuvre. La mise en œuvre de se changement doit être accompagnée par une solide conduite du changement. Sinon, elle sera un échec. Cette sous-estimation serait d'ailleurs l'une des raisons majeures de l'échec de certains projets de *crowdsourcing*.

On aura donc intérêt à avoir une typologie des individus et des groupes d'individus impliqués dans un projet en fonction des critères suivants :

- Degré d'adhésion ou d'opposition au projet
- Pouvoir d'influence sur le projet (qui détient le savoir ? les compétences ? qui décide ? qui contrôle les règles ? …)

Comme le soutient Danièle Imbault, la majeure partie des individus seront généralement à classer comme attentistes-passifs ou comme hésitants. Quelques uns sont des pionniers partants ou inconditionnels. D'autres sont d'hostiles opposants ou partagés. Il est inutile de gaspiller ses forces en communication sur ces derniers hormis afin d'obtenir la liste des arguments contre le projet. Par contre, il est nécessaire de concentrer ses efforts sur les hésitants et les passifs ayant de l'influence sur les autres (hommes clés, leaders d'opinion, représentants du personnels) afin de les convaincre, de les mobiliser autour du projet et de remporter l'adhésion et entraîner les passifs par contamination. Sinon, ce sont les opposants qui s'en chargeront. Il vaut donc mieux se montrer plus organisés qu'eux et limiter leur pouvoir d'influence en mobilisant les hésitants et les passifs.

Pour cela, plusieurs ressorts peuvent être utilisés : la communication sur les moyens, les objectifs et les finalités du changement, les formations, l'implication dans le changement, le soutien d'experts indépendants ou de leaders d'opinions,

la reconnaissance de l'implication des agents et l'exemple de réussite de certaines personnes (communication interne, fête...) et la prise en compte du point de vue des bénévoles, dans le cadre de réunions ou sous la forme de forums où ils peuvent rester anonymes, les déplacements réguliers sur le "terrain" afin d'entendre les agents et "évangéliser", les entretiens individuels, le recrutement de personnes favorables au projet, l'évaluation autour du projet. Comme le suggère François Dupuy dans "La sociologie du changement" (Dunod, 2014), « en matière de changement, le mieux est l'ennemi du bien et le perfectionnisme tatillon est un puissant facteur d'immobilisme ».

Dans les bibliothèques, le changement est rendu difficile par le fait que "les cadres ont une culture administrative et d'encadrement peu propice aux initiatives et à l'innovation" (Delaine, 2014). Les bibliothèques disposent également d'avantages comme le sentiment, propre au service public, d'être au service de l'intérêt général. Néanmoins, les bibliothèques sont habituées au changement, elles ont déjà connu le développement du libre accès, des périodiques électroniques et la numérisation et bénéficient donc d'une expérience importante en conduite du changement.

Le changement prend un certain temps qui peut être assimilé à une période de deuil pendant laquelle les individus renoncent définitivement à l'ancien mode de fonctionnement en passant par les phases suivantes :

- le refus de comprendre et le déni, surtout si l'individu est satisfait par le mode de fonctionnement en vigueur et qu'il faut changer. Il va donc ignorer l'information.
- la résistance au changement, la colère, la sidération, la négociation : elle permet néanmoins de mieux prendre conscience du changement et a donc son utilité. Elle permet surtout d'expliciter des arguments contre le changement et qui sont susceptibles, en fonction de leur pertinence, de le renforcer.
- La pré-intention : par manque d'information, par peur, manque de confiance ou d'intérêt, la personne n'envisage pas de changement immédiatement d'habitude.
- l'intention : elle hésite et envisage de changer de comportement prochainement
- la préparation : elle a décidé de changer et cherche à se former
- l'action : elle se résigne et modifie ses habitudes

- le maintien : avec quelques tentations de retour en arrière
- l'acceptation et la résolution : la personne est satisfaite du changement et nouveau mode de fonctionnement

La conduite du changement comportera, par conséquent 3 phases :

- Diagnostic
- Accompagnement
- Consolidation

Le dirigeant idéal est l'intégrateur qui porte à la fois un fort intérêt pour les résultats et pour ses collaborateurs. L'une des principales fonctions des cadres et des directeurs de bibliothèques devrait être d'avoir une vision de l'avenir de la bibliothèque, de fixer le cap, de donner du sens aux changements, d'expliquer les projets, de persuader du bien-fondé des objectifs, des finalités, d'encourager et de motiver la participation de ses équipes. La direction doit savoir leur donner le sentiment d'être utiles, gagnants dans le changement, connaître leurs responsabilités, leurs marges de manœuvre.

Comme le rapporte (Lebraty, 2015), les entreprises ont beaucoup de difficultés à recourir au *crowdsourcing* pour des tâches qu'elles réalisaient auparavant en interne. Par contre, elles ont beaucoup plus de facilité à mettre en œuvre une démarche de *crowdsourcing* dans le cadre du déploiement d'une activité nouvelle.

3.8- Les grandes étapes d'un projet de *crowdsourcing*

Pour conclure ce chapitre et, en nous inspirant de (Tweddle, 2012), voici les principales étapes d'un projet qui résument les points qui ont précédemment été mentionnés :

- Avant de commencer, définir la finalité et choisir le type de *crowdsourcing* à utiliser.
- En début de projet, recruter l'équipe du projet et les ressources humaines, définir les objectifs, identifier les sources de budget, identifier les types de participants

- En phase de développement, concevoir l’architecture, identifier les prérequis techniques, de données, de stockage, améliorer le fonctionnement de manière itérative.
- En phase industrielle, mener une campagne de communication autour du projet, intégrer les données et apporter un retour rapide aux contributeurs.
- En phase d’analyse, réintégrer les données produites, les rendre réutilisables et évaluer le projet

Chapitre 4- Expérimentations autour de la correction participative de l'OCR et autour du *crowdfunding*

La finalité de notre travail de recherche était d'évaluer les avantages et les inconvénients d'un recours au *crowdsourcing* pour les bibliothèques numériques. Afin de répondre à cette question, une analyse de la littérature a donné lieu à quelques apports conceptuels au sujet des origines historiques du *crowdsourcing* dans les modèles économiques du « juste à temps » et du « à la demande », de la taxonomie du *crowdsourcing*, de son rapport à la loi de la valeur de Ricardo, des freins culturels dans les bibliothèques de France, de la comparaison concrète entre les coûts et les bénéfices de sa mise en œuvre, et enfin, de la conception du *crowdsourcing* tantôt comme moyen démocratiser le patrimoine ou tantôt comme moyen de produire de résultats concrets

Au delà de l'analyse de littérature actualisée grâce à un dispositif de veille, et afin de répondre à la question de la pertinence du *crowdsourcing* en bibliothèque, nous avons également procédé à de l'observation participante dans le cadre de nos fonctions et à des expérimentations pragmatiques dont les résultats ont été objectivés grâce à des entretiens et à des enquêtes.

L'expérimentation principale porte sur le *crowdfunding* qui n'est qu'une des formes de *crowdsourcing*. La thèse pourrait ainsi paraître déséquilibrée entre la partie analyse de la littérature portant sur le *crowdsourcing*, *crowdfunding* inclus et la partie expérimentale portant sur l'une de ses formes. Nous avons toutefois fait le choix de concentrer notre attention sur cette forme qui nous semblait être la moins étudiée et peut être la plus spécifique. Ce choix est également lié à l'objectif de tester et de consolider un modèle économique viable pour développer éventuellement une micro-entreprise pérenne.

Dans le texte qui suit, nous emploierons souvent la première personne du singulier afin de relater notre parcours personnel et nos expérimentations de *crowdsourcing* appliqué à la numérisation, car c'est, sans doute, la forme la moins artificielle pour ce type de récit.

4.1- Première expérimentation autour de Wikisource à l'Ecole Nationale Vétérinaire de Toulouse en 2008

En 2006, alors que j'étais responsable de la Bibliothèque d'Ichtyologie au sein du Muséum national d'Histoire naturelle, j'avais recherché un mécénat pour une opération de numérisation. C'est la Fondation Total qui avait accepté le projet. Il s'agissait de financer la numérisation des 22 volumes de l'Histoire naturelle des poissons de Cuvier et Valenciennes, ouvrage de référence pour les ichtyologistes. Cette première expérience d'externalisation auprès d'un prestataire et de partenariat avec le secteur privé me fit prendre conscience du potentiel qu'il pouvait y avoir à externaliser certaines missions de service public auprès d'acteurs privés et des possibilités de collaborations autour d'intérêts communs.

En 2008, devenu Directeur de la Bibliothèque de l'Ecole Nationale Vétérinaire de Toulouse, suite à un concours d'un grade supérieur, j'avais sollicité auprès de l'association Wikimédia France un nouveau mécénat afin de faire numériser 95 thèses conservées à l'Ecole Nationale Vétérinaire de Toulouse et tombées dans le domaine public. Ces thèses ont été numérisées par la société Azentis et diffusées sous Wikisource [35] et leur OCR brute a été corrigée bénévolement par des internautes. Cette expérience a, à ma connaissance, été l'une des premières expérimentations de *crowdsourcing* dans une bibliothèque française. Mais, en dehors d'un communiqué de presse de Wikimedia France du 9 avril 2009[36], elle a malheureusement très peu été évoquée par la suite, y compris dans les rares publications française sur le sujet. En 2010, un partenariat assez similaire mais à une échelle plus large (1416 livres) a été noué entre la Bibliothèque nationale de France et Wikimédia France comme nous l'évoquons dans le chapitre consacré à Wikisource dans le panorama des projets. Les Archives nationales, les Archives départementales du Cantal et les Archives départementales des Alpes Maritimes participent depuis également à Wikisource.

[35] https://commons.wikimedia.org/wiki/Category:ENVT_thesis_digitized_by_Wikimedia_France (consulté le 23 juin 2016)
[36] http://www.wikimedia.fr/sites/default/files/CP_WMFR_these_ecole_veterinaire_toulouse.pdf (consulté le 23 juin 2016)

Cette expérimentation à l'Ecole Nationale Vétérinaire de Toulouse, pionnière et contemporaine du projet australien TROVE, sur une échelle beaucoup plus modeste, m'a convaincu de la faisabilité et de l'intérêt du *crowdsourcing* appliqué aux bibliothèques numériques et encouragé à poursuivre sa mise en œuvre sur une échelle plus vaste. L'occasion m'en fut donnée à la Bibliothèque Sainte-Geneviève suite à la réussite à un concours d'un grade supérieur.

4.2- Le projet de plateforme mutualisée et participative du PRES Sorbonne Paris-Cité

En 2009, j'ai constaté avec étonnement, en tant que chef des projets de numérisation de la Bibliothèque Sainte-Geneviève, que la majeure partie des documents numérisés par les bibliothèques en France, en dehors de la Bibliothèque Nationale, dormaient sur des CD Rom où des disques durs externes, que la durée de vie de ces supports était beaucoup plus faible que ce qu'on imagine, et que ces documents numériques risquaient donc d'être perdus. Par ailleurs, les mêmes imprimés risquaient également d'être numérisés plusieurs fois par plusieurs bibliothèques, dans la mesure où, n'étant pas diffusés en ligne, il était difficile de savoir qu'ils avaient déjà été numérisés. Plusieurs causes peuvent expliquer pourquoi les bibliothèques renoncent ainsi à diffuser sur Internet ce qu'elles ont numérisés avec des budgets non négligeables (il faut compter entre 60 et 150 € pour la numérisation d'un livre).

La première des causes qui expliquent cette situation est que Gallica ne permet pas d'héberger des documents numérisés en dehors des chaînes de numérisation de la Bibliothèque nationale de France car Gallica est adossé sur les métadonnées du seul catalogue de la BnF, qu'il fonctionne avec un workflow rigide. Une ouverture relative a été proposée par la BnF grâce au moissonnage OAI-PMH d'autres bibliothèques numériques mais ce moissonnage s'arrête aux métadonnées et surtout, elle nécessite que les bibliothèques numériques moissonnées existent et disposent d'un entrepôt OAI. Cette possibilité ne s'adresse donc pas aux bibliothèques qui ne sont pas parvenues à diffuser leur numérisation sur le web. Une autre tentative d'ouverture a été proposée autour du

projet “Gallica marque blanche” avec un service de “tiers archivage” mais elle se solda par la simple livraison du logiciel Gallica et fut expérimentée par exemple, à la Bibliothèque Nationale et Universitaire de Strasbourg (BNUS). Or, il existait déjà de multiples logiciels libres ou commerciaux pour développer des bibliothèques numériques et le projet Gallica marque blanche ne permit donc pas l’ouverture de Gallica aux contributions de bibliothèques extérieures sur le modèle de ce qui est possible avec Internet Archive.

Outre l’absence de débouché public national, une dernière raison expliquant pourquoi de si nombreuses bibliothèques ne mettaient pas en ligne ce qu’elles avaient numérisé vient de ce que les rares bibliothèques étant parvenues à diffuser le résultat de leur numérisation ont développé des plateformes très coûteuses en licences, en infrastructures et en ressources humaines pour une visibilité très faible, une pérennité non garantie, et des fonctionnalités archaïques, obligeant, par exemple, leurs lecteurs à consulter des livres sur des écrans de PC et non sur leurs tablettes liseuses.

Dans ces conditions, à la Bibliothèque Sainte-Geneviève, au lieu de développer une énième petite bibliothèque numérique coûteuse, peu adaptée aux lecteurs et sans grande visibilité, j’ai proposé de nous orienter vers Internet Archive, 2ème plus importante bibliothèque numérique au monde portée par une organisation internationale à but non lucratif, bénéficiant d’une visibilité et d’une pérennité importantes, mais aussi de fonctionnalités avancées comme l’archivage pérenne et la lecture sur tablettes des fichiers EPUB et MOBI pour Kindle. La Bibliothèque Sainte-Geneviève fût ainsi la première bibliothèque de France à participer gratuitement à Internet Archive. Elle fût suivie ensuite par d’autres institutions comme Sciences-Po Paris et le sera très probablement par d’autres institutions prestigieuses. Une étude comparative des statistiques de consultation ainsi qu'un rapport de l'Inspection Générale des Bibliothèques nous ont permis, par la suite, de valider cette décision :

> « *Quant au projet d’une Bibliothèque numérique PSL, la situation des bibliothèques numériques existantes ne nous semble pas plaider, à ce jour et sous réserve de plus ample examen, pour la création d'une bibliothèque*

numérique supplémentaire, de petite taille face aux géants déjà en place (Google books, Internet archive, Hathi Trust, et assez loin derrière déjà, Gallica). En effet, elle exigerait des coûts d'investissement et de fonctionnement importants pour une visibilité probablement relative, et ne pourrait finalement être qu'un assemblage d'ensembles disparates, sans réelle cohésion documentaire. Aussi, s'il demeure quelques programmes communs à trouver, l'existant est déjà considérable et risquerait de n'amener qu'à du doublonnage. Une perspective intéressante et à moindre coût, et véritablement utile au public, nous semblerait plutôt être celle suivie par la bibliothèque Sainte-Geneviève qui consiste à numériser et à verser dans Internet archive. »[37]

En parallèle, des études comparatives et benchmarking au sujet des solutions logicielles existantes avaient été publiées par mes soins sous la forme d'un livre aux éditions de l'ADBS, et de nombreux échanges avaient été développés avec d'autres bibliothèques recherchant également une solution de bibliothèque numérique. L'idée de mutualiser une bibliothèque numérique participative avait rapidement été proposée au Service Commun de la Documentation de l'Université Paris 8, à la Bibliothèque Mazarine, à la Bibliothèque de la Sorbonne et à la Bibliothèque Inter-Universitaire de Pharmacie. Un regroupement informel de ces institutions avait été organisé en 2008 afin de préparer des visites communes d'institutions ayant développé des bibliothèques numériques, de partager de l'information, une veille sur un site web, et d'organiser des présentations de logiciels. Ce groupe avait également travaillé à la rédaction d'un cahier des charges pour le développement d'une bibliothèque numérique mutualisée et participative dès 2009, cahier des charges qui avait été retravaillé fin 2009 par les consultants de Six et Dix à partir d'un financement de la Bibliothèque Sainte-Geneviève. Malheureusement, les bibliothèques de ce regroupement

[37] Jean-Luc Gautier-Gentès, Benoît Lecoq (2012). Le volet documentaire de Paris Sciences Lettres. L'occasion de l'exemplarité. Rapport de l'Inspection générale des bibliothèques - n° 2012-004, 135 p.

informel, qualifié parfois de “brigade volante” appartenaient à des tutelles et à des Pôles de Recherche de l’Enseignement Supérieur (PRES) différents et il était difficile de trouver un financement à ce projet.

Néanmoins, à l’occasion d’une réunion de Directeurs de bibliothèques du PRES Sorbonne Paris-Cité à laquelle mon Directeur m’avait convié, j’émis l’idée de développer cette bibliothèque numérique mutualisée et participative principalement autour de 2 idées : la correction participative de l’OCR par *crowdsourcing*, le financement participatif de la numérisation à la demande par *crowdfunding* et l’impression à la demande grâce à l’acquisition d’une Espresso Book Machine. Cette bibliothèque numérique serait ouverte aux bibliothèques du précédent regroupement informel ainsi qu’à toutes les bibliothèques de France ayant renoncé à diffuser le fruit de leurs campagnes de numérisation faute de solutions satisfaisante de diffusion. L’objectif était précisément de leur offrir enfin un débouché. Cette bibliothèque numérique avait donc une ambition régionale et nationale allant bien au delà du PRES Sorbonne Paris-Cité. Il s’agissait de construire un Internet Archive français pour les livres. Cette bibliothèque numérique devait permettre de partager les coûts de développement et de maintenance, d’accroître la visibilité des documents numérisés, d’améliorer la qualité et la pérennité des projets de numérisation tout en conservant l’identité de chaque institution sur le modèle en marque blanche des collections de l’archive ouverte nationale HAL (graphisme de la vitrine, nom de domaine, statistiques de consultation). Le projet fût rapidement validé par le PRES en 2010 et le cahier des charges fût retravaillé principalement en collaboration entre la Bibliothèque Sainte-Geneviève et Sciences Po dès juin 2010 puis élargi sous la forme de groupes de travail aux autres institutions du PRES, en particulier l’Université Diderot Paris 7 et la Bibliothèque Universitaire des Langues et Civilisations (BULAC). Le projet fut proposé aux élus de la Ville de Paris et obtint un financement de 850 000 €. Le calendrier suivant avait été fixé et diffusé sur le site web du PRES :

“*- Finalisation de la rédaction du CCTP avec l’assistance d’une AMO : 1er semestre 2011*

· Publication de l’Appel d’Offre : juin 2011

· Choix du prestataire : 1er octobre 2011

· Mise au point du logiciel : entre octobre 2011 et février 2012

· Acquisition et installation des serveurs : entre janvier et février 2012

· Recette du logiciel sur les serveurs et corrections : entre mars et avril 2012

· Mise en production: mai 2012"

Malgré ce calendrier, un cahier des charges très développé, un budget conséquent et une finalité claire, le projet fût néanmoins continuellement retardé par des auditions de sociétés, par la constitution de groupes de travail, puis par des événements prévisibles comme le départ à la retraite d'Axel Kahn de l'Université Paris 5 en décembre 2011, le lancement des Idex en février 2012 et beaucoup moins prévisibles comme la mort de Richard Descoings, le Président de Sciences Po en avril 2012, des événements politiques avec les élections présidentielles puis législatives en avril-juin 2012, des événements liés à l'Enseignement Supérieur avec les Assises de la Recherche en septembre 2012...

4.3- Participation au lancement du projet de *crowdfunding* Numalire (Yabé)

4.3.1- Présentation du projet

Numalire est un projet français de numérisation à la demande par *crowdfunding* qui a été expérimenté en 2013 et pendant 8 mois, à compter du 7 octobre 2013 et jusqu'en mai 2014, autour de quelques bibliothèques parisiennes dont la Bibliothèque Sainte-Geneviève et les bibliothèques de l'Institut National de la Recherche Agronomique, mais aussi la Bibliothèque des Arts Décoratifs, la Bibliothèque Historique de la Ville de Paris, la Bibliothèque Forney, la Bibliothèque de l'Hôtel de Ville de Paris, la Bibliothèque de l'Académie nationale de médecine et la Bibliothèque Marguerite Durand. Le projet a été porté par la société YABé fondée par Filippo Gropallo et Denis Maingreaud, cadres chez Orange. Cette société a été soutenue par le Labo de l'édition à Paris. C'est autour de ce projet original qu'a porté l'essentiel de nos expérimentations. 500 000 notices de documents libres de droits conservés dans les 8 bibliothèques participantes ont été diffusés en ligne afin d'en proposer la numérisation aux internautes qui y accèdent directement via les moteurs de recherche, via des liens depuis les catalogues des bibliothèques ou encore via le site Numalire. Ces 8 bibliothèques ne disposant généralement pas de services de reprographie, ce service leur permettait d'offrir un service nouveau à leurs usagers sans avoir à en supporter le coût et de faire financer une partie de leurs programmes de numérisation par les internautes.

Les internautes qui souhaitent financer la numérisation d'un livre conservé par l'une de ces bibliothèques s'authentifient sur le site Numalire et la bibliothèque reçoit une demande de devis. Ensuite, le personnel de la bibliothèque se charge de vérifier que le document peut être numérisé (pas déjà numérisé, pas sous droits, complet, bon état) et valide la demande de devis sous 48 h avant de décrire matériellement le document (nombre de feuillets, format, angle d'ouverture), un travail relativement ingrat, mais qui aurait aussi été fourni pour préparer un

programme classique de numérisation. A partir du coût de numérisation ainsi calculé, une souscription est ouverte sur le site encourageant l'internaute à s'appuyer sur ses réseaux sociaux afin de partager le financement de la numérisation. Si la souscription aboutit, un prestataire vient chercher l'ouvrage pour le numériser selon un cahier des charges précis. Une copie numérique est ensuite livrée aux souscripteurs et à la bibliothèque. Un fac-similé peut ensuite également être commandé sous la forme de *Print on Demand*.

Au départ, l'idée de vendre les EPUB produits a été envisagée puis abandonnée au profit de bénéfices réalisés sur la vente de *Print on Demand*. D'ailleurs les fichiers numérisés sont diffusés sous licence Public Domain Mark. Pour leur part, les bibliothèques bénéficient aussi d'une commission de 15 % du volume total de pages numérisées à utiliser afin de réaliser des numérisations de leurs choix.

En complément de cette présentation du projet, on trouvera également des éléments d'informations en annexe dans la partie destinée au projet Numalire du complément au panorama des projets.

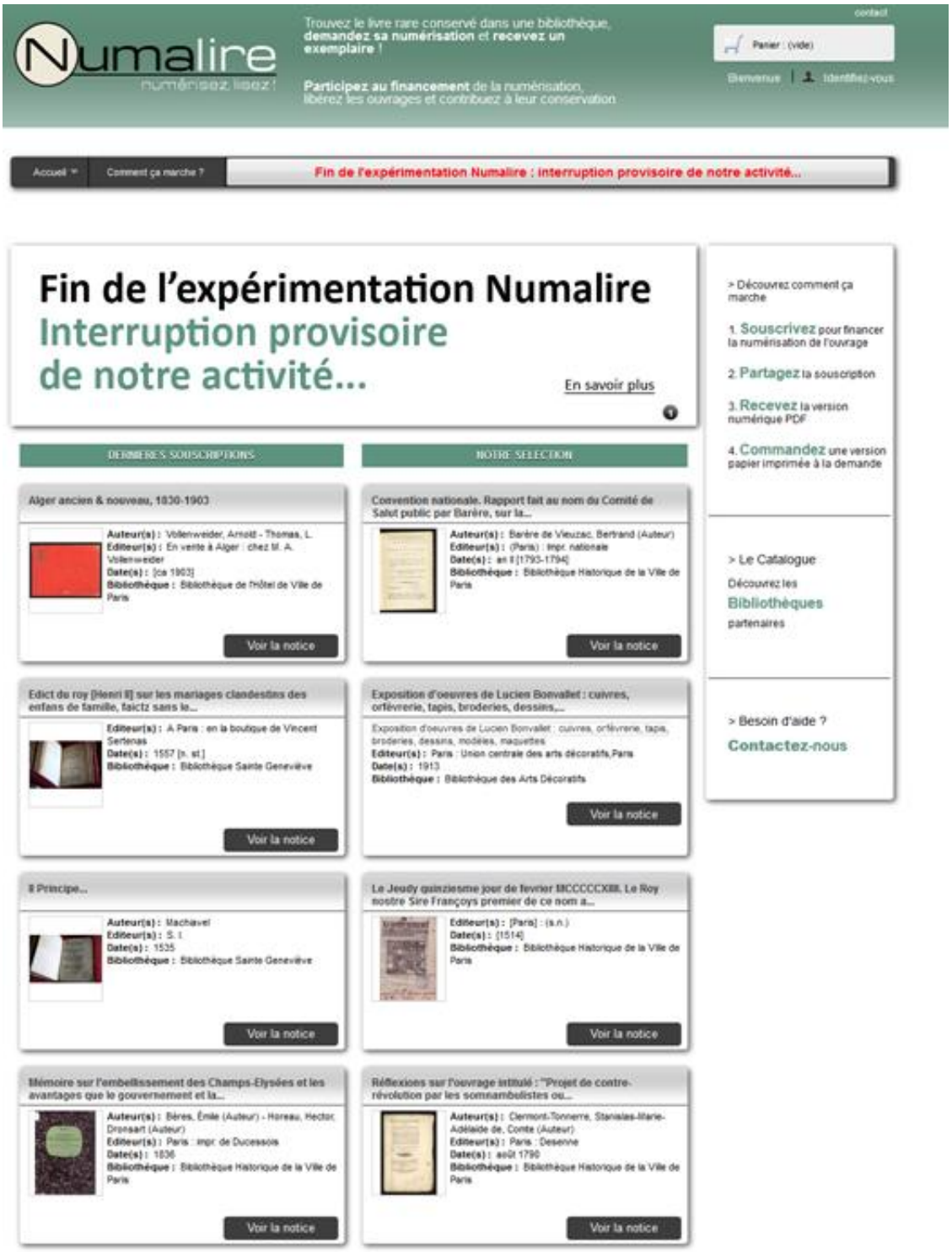


Figure 81. Page d'accueil du site www.numalire.com

Figure 82. Ouvrage dont la numérisation a été financée sur le site www.numalire.com

Dans le cadre de ce travail de thèse doctorale, nous sommes intervenus dans le projet Numalire à la fois en tant que partenaire d'une institution participante, porte parole des bibliothèques participantes et collaborateur du projet. Notre rôle a surtout été un rôle de veilleur, et de consultant spécialiste du *crowdsourcing* ayant

proposé des pistes d'évolutions. Cette initiative a servi d'expérimentation principale pour ce travail de thèse. La problématique a été de tester ce modèle économique, d'en évaluer les avantages et les inconvénients. Pour cela, nous avons eu recours à des entretiens avec les responsables de bibliothèques et à une enquête auprès des clients. Nous avons ensuite formulé plusieurs propositions aux porteurs du projet afin d'en améliorer le fonctionnement.

4.3.2- Référencement web et profil des visiteurs du site numalire.com

Le projet a officiellement été lancé le 7 octobre 2013 pour une expérimentation qui a duré 8 mois. Au jour du lancement, le site avait généré 85 sessions de 71 visiteurs uniques, puis, suite à une campagne de référencement naturel à partir du 22 octobre 2013, le nombre de visiteurs s'est rapidement accru sur tout le mois de novembre qui a réunit 34 847 sessions de 31 549 visiteurs uniques jusqu'à atteindre 1426 sessions de 1370 utilisateurs uniques le 18 novembre 2013 à son apogée. Mais, en décembre il a commencé à décliner avec 21 460 sessions de 19 393 utilisateurs uniques. En janvier 2014, il n'y avait plus que 3 971 sessions de 3 463 visiteurs uniques sur le site.

Au total sur la période de 8 mois de l'expérimentation le site a généré 77 212 sessions pour 66 825 utilisateurs uniques et 115 176 pages vues, soit une moyenne de 6434 sessions par mois, 326 sessions par jour.

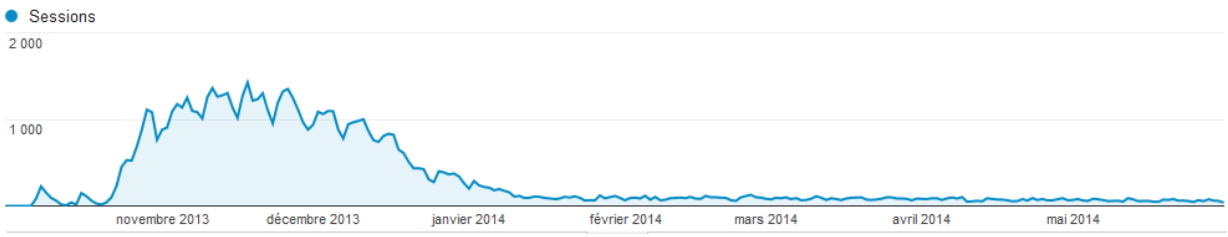


Figure 83. Statistiques journalières de consultation du site numalire.com entre le 1er octobre 2013 et le 31 mai 2014 d'après Google Analytics

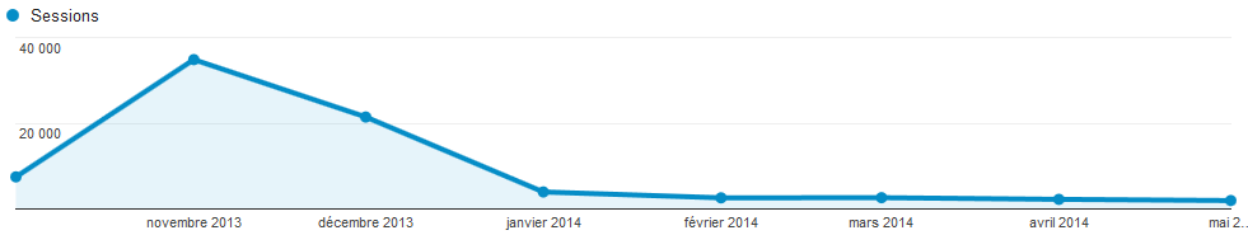


Figure 84. Statistiques mensuelles de consultation du site numalire.com entre le 1er octobre 2013 et le 31 mai 2014 d'après Google Analytics

Cette diminution significative du nombre de visiteur dès la fin de l'année 2013 ne saurait s'expliquer par un "essoufflement médiatique", c'est à dire par la diminution du nombre de visiteurs ayant entendu parler du projet via la campagne de communication et ayant décidé de visiter le site. En effet, le nombre de personnes ayant saisi Numalire dans Google afin de se rendre directement sur le site, n'est pas très important (seulement 144 sessions sur la période, soit 0,19 % des sessions), seulement 3 966 session (5,14 %) proviennent d'une saisie directe de l'adresse du site dans leur navigateur, seulement 2 688 sessions (4,48 %) proviennent d'internautes ayant cliqué sur le lien vers numalire.com depuis un autre site (site de bibliothèque par exemple) et seulement 504 sessions (0,64 %) viennent des réseaux sociaux. Au total, ce sont donc seulement 7302 sessions (9,46 %) qui ne proviennent pas d'une recherche directe dans un moteur de recherche (91 %) et leur évolution temporelle ne présente pas une diminution contrastant avec celle de l'ensemble du site et qui puisse par conséquent l'expliquer. Un "essoufflement médiatique" ne saurait donc expliquer la diminution du nombre de visiteurs.

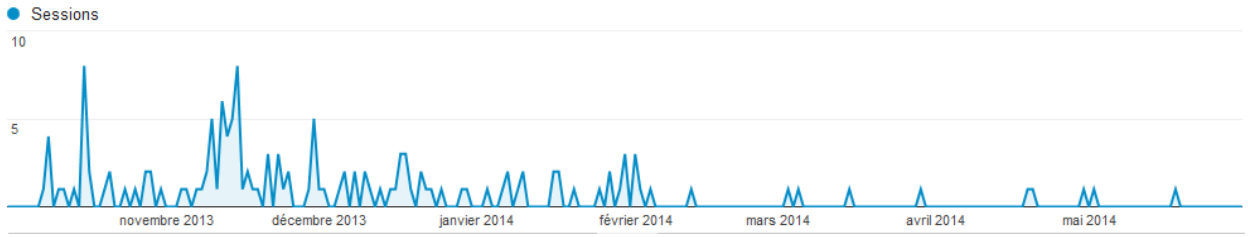


Figure 85. Evolution du nombre de sessions sur le site numalire.com ayant pour origine la saisie du mot numalire dans Google d'après Google Analytics

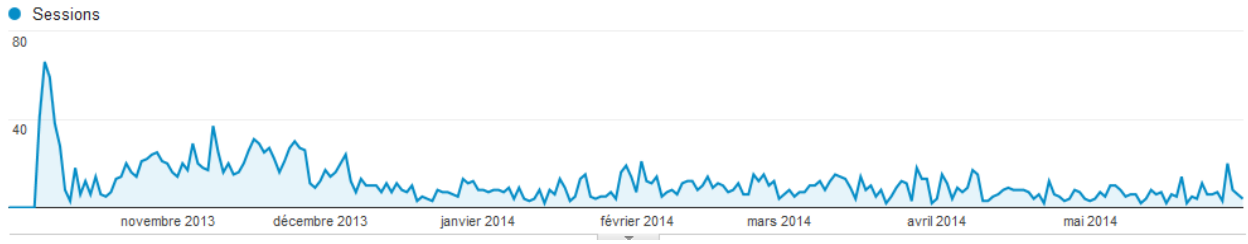


Figure 86. Evolution du nombre de sessions sur le site numalire.com ayant pour origine le clic sur un lien vers numalire.com depuis un site web d'après Google Analytics

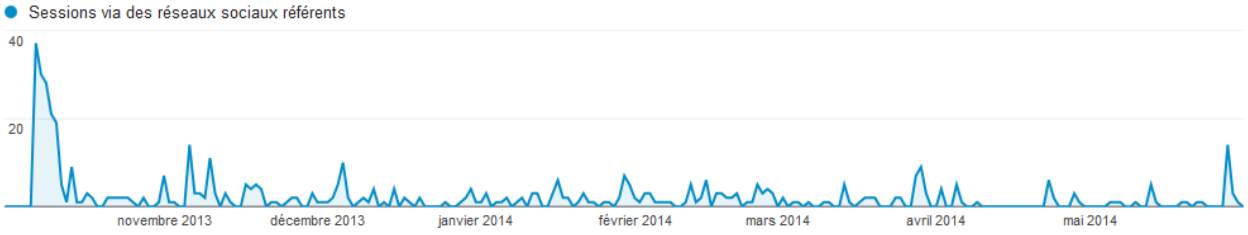


Figure 87. Evolution du nombre de sessions sur le site numalire.com ayant pour origine les réseaux sociaux d'après Google Analytics

Cette diminution du nombre de visiteurs pourrait être liée au fait que les internautes qui se sont rendus sur le site cherchaient à consulter un document numérisé et qu'ils auraient donc progressivement intégré que Numalire ne proposait pas de livres numérisés mais proposait seulement d'en financer la numérisation. Dans ces conditions, les internautes ne viendraient sur le site qu'une seule fois. On constate ainsi que 86,57 % des visiteurs du site ne l'ont visité

qu'une seule fois et que 87,38 % des sessions n'ont pas duré plus de 10 secondes ce qui pourrait accréditer cette idée. Néanmoins, cette hypothèse ne saurait expliquer à elle seule la brutalité de cette évolution qui semble donc bien aussi avoir pour origine un changement de l'algorithme PageRank de Google.

Concernant la démographie des visiteurs, on apprend, grâce à Google Analytics que ce sont plutôt de jeunes hommes de la région Ile de France (33,5 % dans la tranche de 25-34 ans, 54,15 % d'hommes). En gros, ⅓ des visiteurs proviennent de Paris, ⅓ des régions de France et ⅓ de l'étranger (Belgique, Canada, USA, Italie, Suisse, Algérie, Allemagne, Maroc, Espagne…). Cette observation reste peu surprenante dans la mesure où les bibliothèques participant au projet proviennent toutes de la région Ile de France.

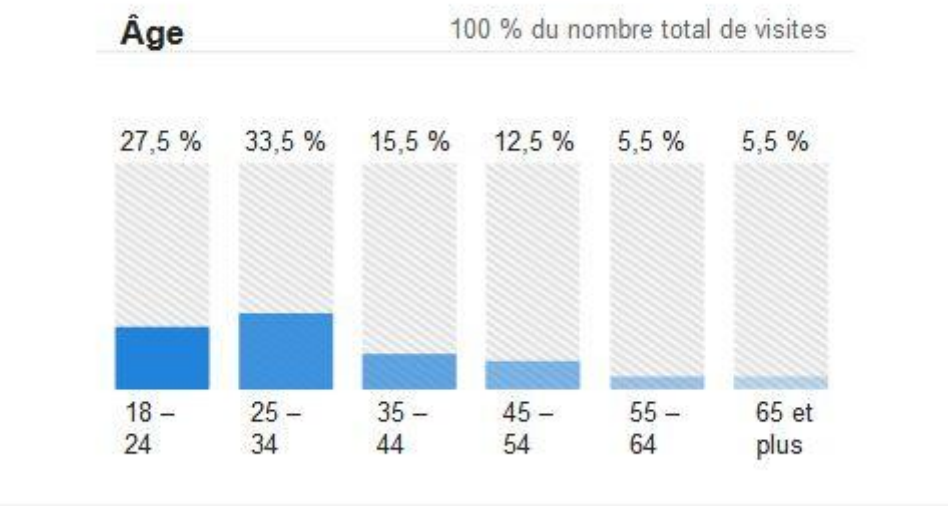


Figure 88. Âge des visiteurs du site numalire.com d'après Google Analytics

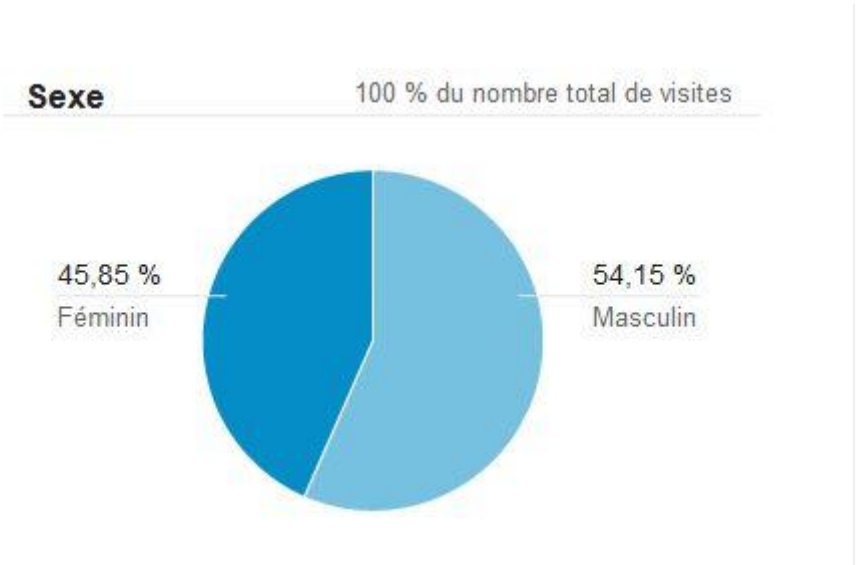


Figure 89. Genre des visiteurs du site numalire.com d'après Google Analytics

Nombre de visites par pays

Pays/Territoire	Visites
France	48 402
Belgium	2 366
Canada	1 599
Switzerland	1 414
Algeria	1 269
United States	1 157
Italy	1 127
Morocco	912
Germany	791
Spain	651

Figure 90. Origine géographique des visiteurs du site numalire.com d'après Google Analytics

Figure 91. Origines régionales des visiteurs français du site numalire.com d'après Google Analytics

4.3.3- Propositions d'améliorations

4.3.3.1- Numériser sans devis et favoriser l'achat impulsif

Le principal risque d'échec du projet avait été identifié dès 2009, bien avant son lancement. En effet, le fondateur de Chapitre.com, Juan Pirlot de Corbion, et que nous avions rencontré à de multiples reprises nous avait fait part d'une expérimentation autour de 2004 qu'il avait mené avec la Bibliothèque nationale de France. Il avait exposé environ 400 000 notices bibliographiques du catalogue de la Bibliothèque Nationale sur un site afin de proposer aux internautes de financer la numérisation des livres correspondants. Environ 50 demandes de devis étaient parvenues chaque jour de la part des internautes, d'après les informations informelles que nous avons collectées. Pour leur part, les agents de la BnF devaient effectuer des descriptions matérielles des documents afin d'évaluer la possibilité ou non de numériser et les coûts de numérisation. Ils devaient ainsi compter le nombre de feuillets, mesurer les angles d'ouvertures, évaluer l'état de chaque livre afin de déterminer la volumétrie, le type de matériel de numérisation à mobiliser. Or, ce surcroît important d'une activité particulièrement ingrate n'avait

probablement pas suffisamment été anticipé. Les devis parvenaient tardivement aux internautes et un nombre très faible d'entre eux aboutissaient à un financement, les internautes ayant généralement changé d'avis entre temps, après avoir reçu leur devis. L'impression de faire un travail aussi ingrat qu'inutile et le coût humain du projet aboutit à son abandon alors que la demande semblait pourtant suffisamment forte.

Dans le cadre du projet Numalire, cette étape de devis reste coûteuse en temps humain pour les bibliothèques. Elle rend moins rentable le projet pour elles dans la mesure où l'argent récolté pour numériser leurs fonds est, en grande partie, compensé par l'argent dépensé sous la forme de temps de travail afin de produire ces descriptions matérielles. Du point de vue du client, cette étape, ralentit considérablement la livraison de la numérisation commandée. Enfin, du point de vue du projet lui-même, une trop faible quantité de demandes de devis aboutissent à des commandes de numérisation car l'achat d'impulsion est impossible avec ce mode de fonctionnement.

L'achat d'impulsion est un achat immédiat, spontané, non planifié, non prévu, précipité, personnel, affectif en réaction à un stimulus (définition inspirée de Bonnefont, 2000). Dans le cas de Numalire, l'internaute recherche généralement un livre numérisé qu'il souhaite consulter ou un livre ancien qu'il souhaite acheter, il découvre qu'il n'est pas déjà numérisé, et, en arrivant par hasard sur le site de Numalire, il se voit proposer la possibilité d'en obtenir la numérisation et d'être également celui qui sera à l'origine de la rediffusion du livre à un public élargi. C'est ce stimulus qui doit être particulièrement soigné en communiquant sur des phrases du type : "avant qu'un autre le fasse avant vous, soyez le généreux mécène qui aura permis à cette œuvre majeure d'être de nouveau accessible au monde et ayez ainsi un lien éternel avec elle en lui ayant permis de passer à l'ère du web", ou encore "soyez le premier à avoir accès à ce trésor de la littérature oublié au fond d'une bibliothèque" mais aussi "Aidez-nous à créer un accès libre et gratuit au patrimoine imprimé des Bibliothèques et devenez un mécène qui contribue à la conservation du patrimoine, y compris pour quelques euros". Afin que le client connecté, c'est à dire, celui dont on connaît déjà le nom, puisse déjà

s’imaginer déjà en situation et ait envie de passer commande, on pourrait aussi écrire son nom sur une plaque de marbre virtuelle comme mécène officiel ayant permis à cette œuvre majeure de passer à l’ère du web. Dans tous les cas, afin de rendre proche le résultat, on pourrait afficher une photographie d’un livre ayant pour page de titre dynamique le titre du livre dont on suggère à l’internaute de financer la numérisation. Mais pour l’heure, dans le cadre du projet, l’achat d’impulsion est rendu impossible par le système de devis.

Au cours de l’expérimentation de 8 mois, sur 414 demandes de devis seuls 36 devis ont abouti au financement d’une numérisation, soit seulement 11 % des devis traités. Cela signifie qu’il faut compter 414 / 36 = 11,5 demandes de devis pour que l’une d’entre elles se convertisse en commande ferme. Cette proportion est insuffisante surtout si on considère le temps de travail nécessaire à la production des devis par les bibliothèques.

Signalons toutefois que, dans 25 % des cas, les bibliothèques ont estimés que le document ne pouvait être numérisé. Paradoxalement il est arrivé régulièrement que le document commandé soit déjà numérisé et déjà accessible gratuitement. Dans ce cas, il s'agissait probablement d'internautes venus tester Numalire, ayant choisi un livre sur une thématique qui les intéressait et n'ayant pas cherché sur le web s'il existait déjà sur support numérique. Il semble, en effet, difficilement concevable qu'un internaute à la recherche d'un livre en particulier puisse en financer la numérisation si celui-ci est déjà accessible gratuitement en ligne. Parmi les autres raisons de refus des bibliothèques, on trouve aussi le fait que l'ouvrage soit sous droit. Afin d'éviter ce cas de figure ou le limiter considérablement, il suffirait de ne pas proposer de documents antérieurs à 1890 ou à 1900. Il reste encore d'autres raisons de refus des bibliothèques comme le fait que les ouvrages soient trop fragiles, manquants ou incomplets. La numérisation d'un ouvrage abîmé devrait, au contraire, être considéré comme d'autant plus nécessaire que, avec l'acidité du papier qui poursuit son œuvre de destruction, le document risque de ne plus du tout être consultable dans quelques décennies. Dans ces conditions, il est probablement préférable de le sauvegarder en le numériser quitte, effectivement, à accélérer sa lente et inexorable destruction.

Mais, malgré ces 25 % de refus, il reste que le rapport de 11,5 % entre le nombre de demandes de devis et le nombre de commandes est trop faible si l'on considère les coûts pour les bibliothèques afin de produire les descriptions matérielles de livres. En effet, on peut considérer qu'il faut compter une bonne demi heure pour effectuer les vérifications (déjà numérisé, droits patrimoniaux) extraire le document au sein des collections et en effectuer la description matérielle. Dans ces conditions, le projet aurait globalement coûté 207 heures de travail aux bibliothèques. En moyenne, il faut donc compter, 11,5 x 0,5 = 5 h 45 de travail pour obtenir une numérisation. Ce temps de travail de 5 h 45 heures est un investissement non négligeable d'autant qu'il s'agit d'un travail assez ingrat. Il faut néanmoins considérer que ce travail aurait également été fourni pour préparer un programme classique de numérisation ou, dans une moindre mesure, avec la consultation des documents par des lecteurs. Converties en coûts, sur la base de 17,88 € de l'heure (catégorie C de la fonction publique), ces 207 heures représentent environ 3700 € pour l'État. On peut donc considérer que la numérisation participative des 36 livres a coûté 3700 / 36 = 102,78 € par livre, ce qui équivaut finalement au prix de la numérisation d'un livre. Les bibliothèques ne sont, dans ces conditions, pas vraiment bénéficiaires avec ce modèle économique dans la mesure où l'opération revient finalement à convertir du temps de travail d'agents de la fonction publique en financement en numérisation. Ce système de devis reste donc le principal inconvénient du projet du point de vue de sa faisabilité.

Aussi, nous proposons de supprimer cette étape de devis. Afin d'éviter que les bibliothèques aient systématiquement à compter le nombre de pages à numériser, le prix pourrait être fixé sur la base des informations contenues dans la notice déjà existante et qui comporte des dimensions imprécises et un nombre de pages toujours en deçà du nombre de feuillets réellement à numériser car les normes de catalogage et de description bibliographique stipulent qu'il faut renseigner le dernier numéro de page folioté et écrit, non le nombre réel de feuillets. Afin d'évaluer le nombre réel de feuillets, on pourrait se baser sur les statistiques récoltées dans le cadre des marchés de numérisation en comparant le

nombre de photographies livrées avec le nombre de feuillets qu'on aurait pu calculer à partir de la notice bibliographique.

Dans le cadre de l'expérimentation Numalire, nous avons constaté qu'il y avait en moyenne et dans la réalité 15 % de feuillets en plus par rapport aux indications contenues au sein des notices bibliographiques sur les 414 demandes de devis. Si sur une notice bibliographique on a "XIV-143 p.", on pourrait très facilement calculer que l'ouvrage comporte 14+143=157 pages et ensuite, comparer avec le nombre réel d'images livrées par le prestataire de numérisation d'un programme de numérisation classique. C'est ce que nous avons fait à la Bibliothèque Sainte-Geneviève sur 515 livres du 19è siècle numérisés pour lesquels nous avons comptabilisé le nombre de pages pouvant être calculé à partir des notices bibliographiques (123 639 feuillets) et l'avons comparé avec le nombre de vues réellement livrées et facturées par le prestataire (134 049 vues). Il s'avère qu'il y avait, en moyenne, 8,42 % de vues à numériser en plus par rapport au nombre de pages qui apparaissent décrites sur les 211 notices bibliographiques. Mais, ces résultats statistiques sont trop divergents et portent sur des échantillons trop restreints pour pouvoir être suffisamment fiables. Nous devrons donc, dans un premier temps, nous contenter de cette indication d'environ 10 % de feuillets en plus par rapport aux indications qui figurent sur les notices bibliographiques. Toutefois, en partant de ce chiffre de 10 %, et en l'affinant au fil de l'accroissement de l'expérience du projet, nous pourrions proposer un calcul automatique relativement fiable de prix sur la base du nombre de feuillets estimé, de la date de publication et du format. Les prestataires de numérisation sauront que cette estimation n'est pas contractuelle et que elle pourra varier légèrement, tantôt en leur faveur, tantôt en leur défaveur. Ils devront ainsi anticiper sur de possibles erreurs ponctuelles, y compris dans les notices bibliographiques elles-mêmes et proposer des tarifs en conséquence.

L'immense avantage de ce prix calculé sans passé par l'intermédiaire d'un devis serait que 100 % des demandes aboutiraient à une commande, que le temps d'attente pour le client serait considérablement réduit et que les bibliothèques réduiraient considérablement leurs coûts en temps de travail ingrat. Au cours de

l'expérimentation de 8 mois, cette proposition a partiellement été mise en œuvre en affichant une estimation de coûts non contractuelle. Bien que l'étape devis ait été maintenue, nous avons constaté que la simple présence de ce calcul avait permis de diminuer de 30 % environ le nombre de demandes de devis et donc, d'augmenter le ratio nombre de commandes réelles / nombre de demandes de devis.

Les principales difficultés pour la mise en œuvre d'un calcul automatique de devis restent les erreurs dans les notices bibliographiques et l'absence d'indications relatives aux angles d'ouvertures et qui déterminent quel scanner doit être utilisé et donc, in fine, quel sera le coût de la numérisation. Néanmoins, nous avons vu que nous pouvions, à partir de statistiques réduire ces incertitudes et les lisser sur l'ensemble des commandes. Dans tous les cas, et pour conclure au sujet de la question des devis, il semble absolument préférable d'alléger au maximum l'intervention des bibliothèques qui ne doivent avoir éventuellement qu'à accepter ou refuser les demandes de numérisation ou seulement de manière exceptionnelle, en cas de notices ne comportant pas des données suffisantes.

Pour le calcul des coûts, il pourrait enfin être envisagé de s'appuyer sur "Digitization cost calculator" développé par la Digital Library Federation (DLF)[38]. Et, pour la vérification des droits, l'API développé par le "Public domain calculator[39]" pourrait être envisagée.

4.3.3.2- Diminuer les coûts de numérisation par son "ubérisation"

Les prix pratiqués dans le cadre du projet Numalire peuvent sembler relativement élevés comparés aux tarifs pratiqués par le réseau européen Ebooks on Demand (53 € en moyenne d'après (Mühlberger, 2009) et parfois aussi par rapport au prix du document original chez un bouquiniste. Cette différence s'explique tout d'abord par le fait que Ebooks on Demand a généralement recours à des ateliers de numérisation internes aux bibliothèques participantes, c'est à dire financés avec de l'argent public qui ont la possibilité de ne pas faire de bénéfices

[38] https://duke.qualtrics.com/jfe/form/SV_3OtqSEAbpl2QDl3 (consulté le 23 juin 2016)

[39] http://outofcopyright.eu (consulté le 23 juin 2016)

et de ne pas répercuter la totalité des coûts sur les usagers. Les prestataires privés, au contraire, sont contraints d'être plus rentables pour conserver leur existence. Cette différence de tarifs s'explique aussi, en partie, par le choix d'un prestataire de numérisation prestigieux et ayant la meilleure réputation auprès de la communauté des conservateurs, mais c'est un prestataire qui a également la réputation d'être cher, bien que nous ayons pu personnellement constater que ce n'était pas toujours le cas. Enfin, il faut considérer que la numérisation à la demande, c'est-à-dire à l'unité, sera toujours plus coûteuse que la numérisation de masse. La numérisation en masse permet de définir des trains d'ouvrages en fonction de caractéristiques physiques similaires (formats, angles d'ouvertures, ancienneté…) et de lisser ainsi les coûts liés aux réglages pour chaque type de document sur un nombre important d'entre eux. A l'inverse, avec la numérisation à la demande, il est nécessaire de consacrer beaucoup plus de temps de préparation à la numérisation de chaque livre.

Si ces prix importants peuvent s'expliquer, comme nous venons de le faire, ils représentent néanmoins un frein évident à l'achat du service et à la réussite du projet. Dans ces conditions, plusieurs solutions pourraient être proposées afin de réduire les prix.

Il pourrait être proposé de faire appel à un prestataire moins coûteux ou de mettre en concurrence les prestataires, éventuellement pour chaque prestation, afin de diminuer les coûts de numérisation, en permettant, par exemple, aux clients de choisir le prestataire de numérisation de leur choix. Une autre possibilité, diamétralement opposée, serait, à l'instar de Ebooks on Demand, de faire appel à un service public, en numérisant sur place les documents avec l'aide d'un agent public se déplaçant de site en site ou d'un atelier de numérisation itinérant dans un camion. La numérisation in situ permettrait en outre de numériser des livres très précieux qui ne peuvent sortir des bibliothèques en raison de leurs trop importantes valeurs d'assurance. Pour les livres moins précieux, il pourrait être proposé d'avoir recours à des services de courtiers et de transport de colis afin d'expédier les livres dont la numérisation a été commandée vers l'atelier de numérisation. Le transport de colis fragiles et précieux comme les œuvres d'art est

largement utilisé par des sociétés comme DHL et pourrait ainsi permettre de gagner en efficacité et en coût. Il pourrait également être possible de faire appel aux marchés de numérisation en cours dans les bibliothèques et aux ateliers de numérisation ainsi installés in situ pour leur faire traiter ces numérisations à la demande.

Afin de motiver les personnes privées ou morales à financer les numérisations en diminuant leurs coûts, des reçus donnant droit à des déductions fiscales de 66 % pourraient être systématiquement proposés à l'image de ce que pratiquent les Amis de la Bibliothèque nationale de France.

Une solution susceptible de mécontenter les prestataires consisterait à proposer à des bénévoles de se charger gratuitement de la numérisation au bénéfice des internautes en ayant fait la demande, y compris, pourquoi pas, dans le cadre de mécénats de compétences d'entreprises. Ainsi, il arrive que des Wikipédiens numérisent bénévolement des fonds dans les bibliothèques et l'association Wikimédia, que nous avons rencontrée dans le cadre du projet, pourrait offrir un relais et une meilleure visibilité au projet. En partenariat avec Wikipédia, on pourrait ainsi afficher les documents dont la numérisation est demandée et des bénévoles pourraient se porter volontaires pour en numériser certains gratuitement.

Enfin, une autre solution beaucoup plus ambitieuse mais susceptible de mécontenter à la fois les sociétés privées et les bibliothécaires consisterait à "ubériser" la prestation de numérisation, c'est à dire à permettre à des auto-entrepreneurs ou à des personnes privées de se charger de la numérisation commandée en échange d'une rémunération pour chaque livre numérisé via le *crowdfunding*. Une marketplace pourrait mettre en relation très librement ceux qui commandent une numérisation avec ceux qui l'effectuent sur le modèle du très libéral modèle économique de Uber. Ainsi certains internautes en rémunéreraient d'autres pour effectuer la numérisation. Dans ce cas, sur la base du prix fixé par la plateforme selon le calcul automatique, la mise en concurrence pourrait se faire sur les prix. Cette mise en concurrence entre eux des prestataires et des prestataires bien établis avec de simples personnes privées mobiles et

indépendantes à le recherche de compléments de revenus pourrait permettre de diminuer considérablement les prix, et peut être aussi les délais, et d'augmenter considérablement l'efficacité du projet. Afin d'"ubériser" encore d'avantage le modèle économique, il pourrait même être proposé de numériser des collections de particuliers au delà de celles qui sont conservées dans les bibliothèques. Des collectionneurs pourraient ainsi proposer et se faire financer la numérisation de leurs propres livres en les partageant sur le web. Mais, à en juger, par le sort réservé à la société Uber en France, il reste fort probable qu'il soit difficile de l'expérimenter sans une conduite du changement conséquente.

4.3.3.3- Élargir à d'autres types de documents que le livre imprimé

L'expérimentation a porté exclusivement sur des livres imprimés. Or, l'expérience des chefs de projets de numérisation en bibliothèques et de prestataires privés montre que ce marché est en train de se restreindre. En effet, avec la numérisation de masse et, en particulier, avec le seuil de 30 millions de livres numérisés par Google, il devient difficile de trouver des livres imprimés qui méritent encore d'être numérisés, la plupart l'ayant déjà été ou ne pouvant pas encore l'être en raison du Code de la propriété intellectuelle. Dans ces conditions, il ne reste que quelques miettes à numériser et le projet Numalire pourrait finalement arriver trop tard, la numérisation étant déjà de l'histoire ancienne. Par exemple, sur plus de 16 000 ouvrages du 19è siècle conservés à la Bibliothèque Sainte-Geneviève, seuls 400 d'entre eux présentaient tous les critères pour être numérisés.

Il serait donc particulièrement opportun d'élargir l'offre à d'autres types de documents, en particulier la numérisation d'articles qui font partie des documents les plus demandés en bibliothèques. La numérisation d'articles de quelques pages serait, par ailleurs, beaucoup moins onéreuse que la numérisation de volumes de livres et pourrait, par conséquent, attirer une clientèle avec des capacités financières plus modestes.

4.3.3.4- Communiquer d'avantage sur les réseaux sociaux et s'appuyer sur les investisseurs, les mécènes, les libraires

Au delà des recommandations générales classiques déjà développées dans le chapitre précédent et sur lesquelles nous ne reviendront pas ici, il pourrait être opportun de proposer aux clients la création de leurs comptes directement via leurs comptes Facebook, Twitter ou Google+. Cela faciliterait non seulement la création de compte sur le site, mais permettrait surtout de proposer ensuite systématiquement aux internautes de partager leurs lancements de souscriptions avec les membres de leurs réseaux sociaux, de l'afficher sur leurs "murs" et permettrait de mieux cibler ainsi des personnes susceptibles de contribuer. Ainsi, le projet pourra être diffusé de façon virale sur les réseaux sociaux, d'autant que les souscripteurs peuvent légitimement avoir envie de faire savoir qu'ils contribuent à la sauvegarde du patrimoine. Il s'agit donc bien ici d'externaliser aussi auprès des internautes la fonction marketing dans une logique de *crowdsourcing* assez implicite.

Afin d'augmenter la visibilité du projet, un partenariat pourrait être proposé à des grandes sociétés de *crowdfunding* comme KissKissBankBank ou Kickstarter afin de synchroniser le contenu de Numalire avec ces sociétés. Ainsi, chaque document dont le financement de la numérisation est proposé sur Numalire pourrait être affiché sur ces grandes plateformes qui bénéficient d'une visibilité bien plus forte via un moissonnage périodique des métadonnées de notices de livres.

Par ailleurs, pour le moment, seul le cas de figure de l'internaute qui souhaite financer la numérisation d'un livre en particulier a été prise en compte, non celle de l'internaute qui souhaite plutôt aider une institution dans ses programmes de numérisation. Afin de faciliter ce cas d'usage, il serait nécessaire de permettre aussi un affichage des livres en fonction des bibliothèques qui les conservent. Ainsi, les contributeurs pourraient choisir de financer la numérisation d'un document parmi ceux conservés dans l'institution de leurs choix.

Pour le moment, le projet a surtout été pensé comme un service de numérisation à la demande permettant de satisfaire le besoin en documentation

des clients. Il serait judicieux d'élargir la demande aux mécènes, en proposant systématiquement à ceux qui ont permis de financer une numérisation, la possibilité d'ajouter leur logo, et surtout d'insérer un lien vers leur site web ainsi qu'un texte de leurs choix. Cela permettrait ainsi de mentionner, par exemple, "cet ouvrage a été financé grâce au soutien de la Fondation Total". A partir de là, il pourrait également être intéressant d'afficher la liste des documents numérisés grâce à telle ou telle fondation. Pour cela, une campagne de communication à destination des mécènes serait nécessaire.

Le projet pourrait également s'adresser à des investisseurs, c'est à dire à des personnes qui financeraient la numérisation de tel ou tel livre dans l'espoir d'obtenir un retour sur investissement en terme de trafic web. En effet, si l'on considère le prix des campagnes de publicité Google Adwords et Google Adsense qui permettent de faire apparaître un lien publicitaire vers un site web depuis une page de résultats de requête Google ou directement sur un site web, il pourrait être très intéressant de financer la numérisation d'un livre et de bénéficier d'un lien vers un site web pour une durée indéterminée. Un livre numérisé générant plus de 50 000 visites par an et générant une centaine de clics sur le lien vers le site de l'investisseur ayant financé sa numérisation pourrait s'avérer être un bon investissement. Une fonctionnalité de comptage de clics pourrait appuyer une communication à destination d'investisseurs potentiels. Ainsi, un modèle économique harmonieux gagnant-gagnant et innovant serait proposé. En effet, la numérisation d'un livre coûte entre 40 et 70 €. Cela correspond au coût généralement facturé pour la visualisation d'une publicité par environ 6000 internautes. Dans ces conditions, on peut considérer que la numérisation d'un livre par une société est rentabilisée dès qu'un nombre proche de 6000 internautes ont vu son nom. Certains livres mettront plus de 200 ans pour obtenir ce trafic et resteront donc probablement à la charge des bibliothèques, de généreux mécènes ou d'usagers en ayant besoin, mais d'autres ne mettront que quelques années à être vus par 6000 visiteurs et peuvent donc être numérisés avec de l'argent privé par des sociétés ou par des particuliers investissant dans la numérisation de tel ou tel livre afin d'en obtenir du trafic web ou une publicité en retour. Cela permettrait à

l'argent public de se concentrer sur les documents d'intérêt patrimonial, historique ou scientifique qui n'intéressent ni le grand public ni le secteur privé et de laisser ces derniers se charger de financer la numérisation des livres susceptibles de susciter l'intérêt du public ou de générer du trafic web. Sur Internet Archive, par exemple, le livre le plus téléchargé (amusements in Mathematics / Dudeney) l'a été 2 624 587 fois en 5 ans. L'investisseur qui aurait investi dans sa numérisation aurait donc bénéficié d'un excellent retour sur investissement avec son nom et son lien affichés plus de 40 000 fois par mois. Il faut, par ailleurs, considérer que ces publicités permettront de très bien cibler leurs publics. Par exemple, la numérisation d'un livre sur tel ou tel sujet intéressera plutôt tel mécène ou tel annonceur en particulier.

Outre le fait de proposer de faire figurer le nom du particulier (s'il le souhaite) ou du mécène ayant financé la numérisation du livre (à la manière de publicités Google Adwords), on pourra imaginer aussi un modèle économique permettant à ces personnes de bénéficier d'un retour sur investissement en leur permettant de bénéficier d'une partie des recettes de la vente d'impressions à la demande (à la manière de YouTube qui rémunère les internautes dont les vidéos ont généré un important trafic web). Au delà de l'usager qui a simplement besoin de la numérisation d'un livre, nous pourrions donc nous adresser au mécène qui veut que son nom apparaisse sur un livre, à l'investisseur qui espère que le document dont il aura financé la numérisation génère un nombre important de visites et de clics vers son site web. Nous pourrions également imaginer le cas d'un investisseur qui souhaite commercialiser l'accès au livre numérisé et/ou son EPUB et/ou son POD pendant une durée déterminée pendant laquelle il pourra bénéficier d'une exclusivité d'exploitation et au delà de laquelle, le livre sera accessible gratuitement et mis sous licence Public Domain Mark. Ainsi, dans le cas d'un investisseur qui souhaite financer la numérisation d'un document pour en commercialiser le document électronique pendant une durée de 3 ans, il faudrait offrir une interface pour le paiement en ligne de l'accès ou du téléchargement du document électronique et pour le versement des droits payés à l'investisseur. L'investisseur pourrait ensuite déterminer librement le prix de vente du document

dont il a financé la numérisation. Par exemple, un livre qui aura coûté 60 € à la numérisation sera amorti au bout de 10 ventes s'il est vendu 6 € le PDF ou au bout de 20 ventes si le PDF est vendu 3 €. Des petites communautés de chercheurs et d'érudits se chargeraient ainsi eux-mêmes de la valorisation et du bon référencement de la plateforme, leurs intérêts convergeant avec ceux de Numalire.

Un autre élargissement du domaine d'application pourrait être le traditionnel prêt entre bibliothèques auquel la numérisation à la demande pourrait permettre un rajeunissement. Enfin, il serait peut être judicieux de ressusciter l'ancienne pratique des souscriptions telle qu'elle existait aux 19è et 20è siècles. A l'époque, les libraires et leurs clientèles pouvaient se mobiliser sous la forme de souscriptions afin d'éditer des livres, dans le cas présent afin de rééditer des livres anciens en les numérisant.

4.3.3.5- Améliorer le référencement du site en multipliant les liens vers ses notices et en créant une bibliothèque numérique

Comme nous l'avons vu, la fréquentation du site a connu une diminution sensible passant de 31 549 visiteurs uniques en novembre 2013 à seulement 3 463 en janvier 2014.

Au delà du recours à des sociétés de référencement et afin d'améliorer le PageRank du site qui n'est que de 2/10 d'après http://www.pagerank.fr en août 2015, il serait judicieux de multiplier les liens qui pointent vers le site en mobilisant les réseaux sociaux, les partenaires mais aussi et surtout en ajoutant des boutons vers le site depuis chaque notice des catalogues des bibliothèques partenaires ou même depuis le SUDOC comme cela est désormais possible pour Ebooks on Demand. En effet, le PageRank de Google, qui détermine la position d'un site dans la liste des résultats à une requête dans Google, est calculé en grande partie en fonction du nombre de liens qui pointent vers tel ou tel nom de domaine et du PageRank des sites d'où proviennent ces liens.

Voici, par exemple, ce qui est pratiqué par le réseau Ebooks on Demand, il s'agit ici d'un lien-bouton depuis le catalogue en ligne de la Bibliothèque Inter-Universitaire de Santé vers EOD.

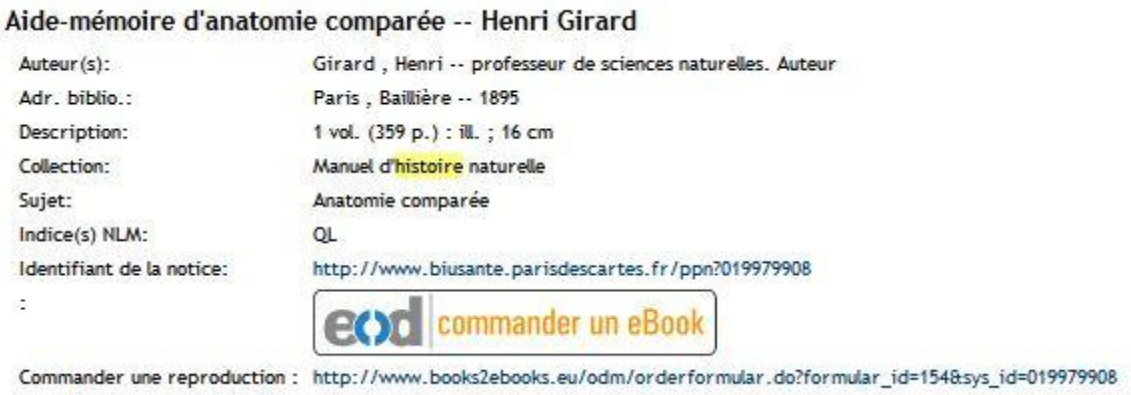


Figure 92. Capture d'écran d'un bouton Ebooks on Demand

Pour le moment, dans la liste des résultats à une requête Google, les métadonnées de livres sur le site Numalire apparaissent ainsi :

Figure 93. Capture d'écran de la manière dont Google affiche une notice Numalire

Si on écrivait plutôt en titre de la page web "Catalogue des livres en très petit nombre... **(livre numérique)**", peut être que ça donnerait d'avantage envie aux internautes de cliquer. Mais peut être aussi que les gens seraient déçus de ne pas trouver le livre numérisé mais la seule proposition d'en financer la numérisation... Il est d'ailleurs également possible, comme nous l'avons vu, que ce soit l'une des raisons de cette baisse de trafic. Les gens savent désormais, après l'avoir expérimenté, que sur Numalire, ils ne trouveront pas les documents numérisés et ils ne cliquent plus. Peut être qu'il pourrait donc être judicieux de mettre également du contenu numérisé sur Numalire afin que les internautes continuent à cliquer dans l'espoir de tomber sur le livre numérisé et pourquoi pas, se laissent tenter un jour par la perspective du *crowdfunding*. Dans tous cas, mettre en ligne du contenu devrait être bénéfique pour le référencement du site.

Pour cela, nous pourrions systématiquement diffuser les documents numérisés par *crowdfunding* sur Internet Archive, 2ème bibliothèque numérique au monde, portée par une organisation internationale à but non lucratif, proposant des fonctionnalités avancées et pérennes, gratuit pour ses participants et bénéficiant d'un excellent PageRank (8/10) et d'une forte visibilité sur le web. Une collection sur le modèle de celui créé par la Bibliothèque Sainte-Geneviève pourrait être proposée : https://archive.org/details/bibliothequesaintegenevieve (consulté le 23 juin 2016)

Sur chaque ouvrage numérique, un lien pourrait pointer vers le site numalire.com et améliorer ainsi son PageRank, sa visibilité et donc son trafic web.

Mais ce n'est pas tout. Légalement et techniquement, il est tout à fait possible d'afficher sur des pages web de Numalire de documents diffusés sur d'autres bibliothèques numériques. Ainsi, http://www.bibnum.education.fr (consulté le 23 juin 2016) affiche des ouvrages numérisés et diffusés sur Scribd grâce à des fonctionnalités de type "lecteurs embarqués" ou "embed". Ces pages web, très faciles à réaliser, pourraient drainer un trafic web supplémentaire, surtout si on affiche les documents en langue française les plus consultés de Internet Archive[40] et de Gallica[41] sans avoir à en supporter les coûts réels d'hébergement et de diffusion.

4.3.3.6- Élargir à l'international et devenir partenaire du réseau européen Ebooks on Demand (EOD)

Le modèle économique, une fois bien établi en France, il pourrait être exporté dans d'autres pays. Il existe en particulier le réseau Européen Ebooks on Demand dont les résultats sont bons et dont il serait improductif et vain de rentrer en concurrence. Sans revenir sur ce réseau largement évoqué dans le panorama des projets, nous pouvons décrire les principales différences de fonctionnement

[40] https://archive.org/search.php?query=%28%28language%3Afre%20OR%20language%3A%22French%22%29%29%20AND%20mediatype%3Atexts&sort=-downloads (consulté le 23 juin 2016)

[41] http://gallica.bnf.fr/topdocs?lang=FR (consulté le 23 juin 2016 mais la ressource n'était plus en ligne)

entre Numalire et EOD. EOD propose un service de numérisation à la demande qui, à la différence de Numalire, ne permet pas à plusieurs contributeurs de financer la numérisation du même livre de manière participative sous la forme d'une souscription *crowdfunding*. Mais, la principale différence est que les bibliothèques participantes à EOD ont généralement recours à leurs propres services de numérisation en interne. Il est donc bien difficile à des bibliothèques qui ne disposent pas d'un atelier de numérisation de participer à EOD. Au contraire, Numalire fait appel à un prestataire privé de numérisation, la société Arkhênum et a d'ailleurs pour bibliothèques partenaires, des bibliothèques qui, précisément, ne disposent pas d'un service de numérisation et souhaitent justement, via Numalire, en proposer un à leurs lecteurs sans avoir à en supporter le coût. D'une certaine manière, à l'échelle internationale, Numalire pourrait donc compléter utilement Ebooks on Demand en s'adressant justement aux bibliothèques qui n'ont pas d'ateliers de numérisation.

Ainsi, pour EOD, de nouvelles bibliothèques pourraient ainsi désormais participer à son réseau désormais y compris lorsqu'elles n'ont pas d'ateliers de numérisation car, dans ce dernier cas, la commande pourrait être automatiquement renvoyée vers Numalire. Cela permettrait à EOD d'élargir son réseau aux bibliothèques qui ne numérisent pas en interne et à Numalire d'élargir son réseau à l'international. Mais cela impliquerait de réfléchir à des partenariats avec des prestataires de numérisation à l'étranger et, techniquement, à développer une passerelle entre EOD et Numalire.

4.3.3.7- Numériser le contenus des bibliothèques sans convention préalable et changer de statut

Jusqu'à présent, les 8 bibliothèques ayant participé à l'expérimentation ont toutes été amenées à signer une convention. Afin d'élargir le nombre de bibliothèques participantes, d'augmenter le nombre de titres proposés à la numérisation et ainsi d'améliorer la visibilité du site, son trafic web mais aussi d'augmenter le nombre de commandes, il pourrait être envisagé de moissonner le contenu des bibliothèques via les entrepôts RDF de métadonnées mis à

disposition par les catalogues nationaux SUDOC[42] et CCFR dans le cadre de politiques de développement du web sémantique et d'ouverture des données publiques.

Le catalogue de Numalire, une fois élargi à l'ensemble du patrimoine imprimé national, il resterait pour chaque commande dans une nouvelle bibliothèque à lui demander si elle accepte qu'on lui numérise sur place un des documents qu'elle conserve pour satisfaire les besoins d'un lecteur. Elle recevra en échange le PDF du document sous une licence Public Domain Mark qu'elle pourra diffuser dans sa propre bibliothèque numérique si elle le souhaite et qui sera dans tous les cas diffusé sur Internet Archive sous la même licence la plus permissive possible. Il est fort probable que les bibliothèques acceptent et que leurs tutelles les encouragent fortement à le faire.

Néanmoins, le fait que Yabé ait un statut de société peut générer une forme de méfiance de bibliothèques peu habituées à ce type de partenariat. Dans ces conditions, le changement de statut est déjà fortement envisagé par ses fondateurs afin de devenir une association ou un fonds de dotation. Cela permettrait d'obtenir d'éventuels financements publics, un détachement d'un agent public et permettrait aux clients d'obtenir une défiscalisation. Un fond de dotation est une Association reconnue d'utilité publique pour poursuivre une mission d'intérêt général. Ce statut permettrait à Numalire d'avoir le soutien de mécènes, d'être mieux accepté par les bibliothèques qui pourront être représentées au sein du fond de dotation et faciliterait la possibilité de trouver un financement public par appels à projets de ministères ou de la Commission Européenne ou une subvention du Centre National du Livre.

4.3.4- Enquêtes auprès des bibliothèques et auprès des clients

Nous avons sollicité les professionnels et les internautes afin de recueillir leurs réactions face à ces propositions ainsi que, dans une démarche de « *user innovation* », leurs idées d'évolution.

[42] http://www.abes.fr/Acces-direct-a/Pour-reutiliser-les-donnees/Presentation-des-jeux-de-donnees-reutilisables (consulté le 23 juin 2016)

4.3.4.1- Auprès de bibliothèques

Dans le cadre de l'expérimentation et de la thèse, des entretiens téléphoniques ont été conduits auprès des bibliothèques partenaires.

Au début du projet, certaines réticences ont pu s'exprimer au sujet de la faisabilité du projet y compris sur le plan juridique, de son caractère non prioritaire ou inutile, de l'inexistence d'une réelle demande, mais aussi d'une crainte que le commerce privé se substitue aux financements publics et aux bibliothécaires. Mais les bibliothèques ont finalement été séduites par le projet ou parfois ont été décidées par leur hiérarchie. Ainsi Jean-Louis Missika Adjoint au Maire de Paris a eu une influence positive pour le projet. Dans tous les cas, n'ayant ni budget ni atelier pour numériser, beaucoup ont eu le sentiment de n'avoir rien à perdre à tenter cette expérimentation. D'ailleurs certains agents plutôt réticents en début de projet, ont finalement souvent proposé le service à leurs lecteurs.

Au cours du projet, les bibliothèques ont été assez surprises de constater que de nombreuses demandes concernaient des livres déjà numérisés, mais probablement mal référencés sur le web. Le prix d'une numérisation n'étant pas anodin, on peut s'étonner que certains soient prêts à financer un livre ayant déjà été numérisé et on pourrait supposer que le travail consistant à vérifier si tel ou tel livre a déjà été numérisé aurait pu ainsi être externalisé auprès du grand public avec d'avantage d'efficacité (c'est d'ailleurs aussi une forme de *crowdsourcing* implicite). Toutefois, on peut supposer qu'il s'agit non d'internautes arrivés sur le site en recherchant un livre particulier sur un moteur de recherche, mais plutôt d'internautes étant arrivés directement sur le site et ayant décidé de tester ou de soutenir le projet. Dans ce cas, ils ont pu tout à fait passer à coté du fait que le livre choisi avait déjà été numérisé.

Concernant les délais de fonctionnement et les tâches à effectuer, les bibliothèques se sont montrées globalement satisfaites. Toutefois, le souhait d'automatiser certaines tâches et de diminuer les coûts pour les bibliothèques a été émis. Par contre, concernant les prix, l'ensemble des professionnels interrogés les ont trouvés beaucoup trop élevés. C'est d'ailleurs, de leur point de vue, le principal inconvénient du projet. A titre de comparaison, l'un d'entre eux a signalé,

au cours d'un entretien téléphonique, que le livre "Les Victimes du lait et du régime lacté" est évalué à 290 € via Numalire et à seulement 59,80 € via EOD et qu'il était même parfois plus cher de financer la numérisation d'un livre que d'acheter l'original chez un bouquiniste. Dans ces conditions, il est probable que le client bibliophile ne réfléchira pas longtemps pour choisir le livre imprimé d'origine.

S'agissant, à présent, de l'évolution du nombre de demandes, les bibliothèques ont constaté une diminution au fil du projet et se demandent si l'effet nouveauté lié au projet pourrait être déjà passé et pourrait expliquer la diminution de l'engouement des internautes. Elles ont constaté que le financement d'un livre était généralement l'œuvre d'une seule personne et qu'il n'était que rarement partagé entre plusieurs.

Néanmoins, tous les partenaires ont annoncé leur volonté de poursuivre le projet à condition toutefois d'améliorer le taux de conversion des demandes de devis en commandes et de diminuer ainsi les coûts pour les bibliothèques, d'autant que ce travail de description matérielle de livres peut être ingrat et décourageant. Du point de vue de multiples professionnels interrogés, la question du devis automatique semble toutefois difficile à mettre en place en raison de notices bibliographiques de qualité insuffisante, de la présence de dépliants non signalés dans les notices, de types de documents trop singuliers ou encore des angles d'ouvertures qui n'y sont pas mentionnés. Or, c'est une indication nécessaire pour déterminer quel matériel sera utilisé pour numériser et une indication malheureusement inexistante dans les notices bibliographiques. Selon certains collègues le devis automatique doit donc être limité aux ouvrages à faible valeur d'assurance. Il risque de générer des erreurs de calculs. Une simple indication sous la forme de fourchette serait, par contre et de ce point de vue, utile afin d'améliorer le taux de conversion des demandes de devis en commandes.

L'idée de s'ouvrir aux mécènes et aux investisseurs ne convainc pas d'avantage. L'idée de moissonner périodiquement les métadonnées des bibliothèques intéresse les professionnels interrogés, mais celle de solliciter les bibliothèques sans convention préalable les laisse parfois dubitatifs. Eux mêmes préfèrent qu'il y ait au moins une convention à la première sollicitation.

4.3.4.2- Auprès des clients

Une enquête a été réalisée en juin 2014. Un questionnaire a ainsi été envoyé à 380 personnes possédant un compte sur le site Numalire. 118 d'entre elles (31 %) y ont répondu. On trouvera en annexe les données tirées de cette enquête.

En résumé, concernant la sociologie des contributeurs, on observe une nette domination des hommes (70, 59%) venant de milieux sociaux favorisés et bénéficiant d'un haut niveau d'études. Ceci peut s'expliquer en raison du coût important de la numérisation comme des sujets spécialisés dont traitent les livres qui n'ont pas déjà été numérisés et diffusés sur le web. D'ailleurs, 48 % de ceux qui n'ont pas donné suite à la souscription l'on fait en raison du prix à payer.

La majeure partie d'entre les internautes (72,73%) est arrivée directement via un moteur de recherche à partir d'un titre de livre recherché.

4.3.5- Résultats et conclusions de l'expérimentation

Au total, sur 8 mois, le projet a généré 414 demandes de devis. Parmi elles, 30% de documents ont été déclarés non numérisables (déjà numérisé, sous droit, disparu, trop ancien ou trop mauvais état), mais 270 souscriptions ont été ouvertes et ont permis 36 numérisations effectives, soit un taux de 11 % de demandes de numérisations qui se concrétisent en commandes réelles.

Sur ces 36 numérisations, seulement 14 % ont fait l'objet d'un financement participatif sous la forme d'un achat collectif.

D'après nos calculs, la participation au projet a demandé environ 207 heures de travail aux bibliothèques soient 3700 € d'argent public. Dans ces conditions, la numérisation des 36 livres a finalement été financée par de l'argent public sous la forme de temps de travail à effectuer les descriptions matérielles des livres nécessaires aux devis. Pour que le projet soit rentable, il serait nécessaire d'automatiser ce calcul de devis ou, au moins, de diminuer le nombre de devis nécessaires pour obtenir une commande.

Avec un chiffre d'affaire de seulement 2000 €, le projet est loin d'être viable. Néanmoins, le taux de conversion du nombre de visiteurs en commandes étant stable et comme il n'existe pas de concurrents dans les moteurs de recherche, en augmentant le contenu et le nombre de visiteurs, les revenus du projet ne peuvent qu'augmenter mécaniquement, ce qui milite fortement pour un élargissement de l'échelle de l'expérimentation.

Plus largement, cette expérimentation nous a conforté dans l'idée que le *crowdfunding* était bien une forme de *crowdsourcing* faisant appel à l'aide des internautes non pas sous la forme de travail mais sous la forme de contributions financières. A l'instar des microtâches, cette participation reste toutefois limitée dans un cadre défini par l'institution et ne permet pas aux internautes d'influer sur la politique des projets en dehors du fait de peser collectivement la politique d'acquisition des bibliothèques numériques.

Il aurait probablement été intéressant de comparer les résultats de cette expérimentation avec d'autres projets de numérisation à la demande. Néanmoins, le projet Ebooks on Demand (EOD) s'appuie sur des ateliers de numérisation internes aux bibliothèques alors que le projet Numalire s'adresse, au contraire, aux bibliothèques qui n'en disposent pas. Les deux modèles étant probablement complémentaires mais néanmoins très différents, leur comparaison ne nous a pas semblé pertinente.

Ce chapitre expérimental a fait l'objet d'un article (Andro, 2014, 2)

Conclusion de la thèse

Au cours de ce travail de recherche, nous avons été conduit à étudier la définition, l'origine, la philosophie conceptuelle politique, économique et managériale du *crowdsourcing* en bibliothèques, puis à réaliser un panorama relativement détaillé et complet des projets existants, à analyser ce mouvement sur le plan de la taxonomie des projets, de la motivation des internautes, du *community management*, de la qualité des données produites, de l'évaluation des projets, et de la conduite du changement. Au cours de ce travail, nous avons été amenés à développer nos propres concepts en tant que contributions à la connaissance du *crowdsourcing* en bibliothèques avec des apports originaux au sujet des origines historiques du *crowdsourcing*, de sa taxonomie, de la loi de la valeur, et des analyses personnelles au sujet de ses difficultés de mise en œuvre dans les bibliothèques françaises et des conceptions diamétralement opposées qui y conduisent. Nous avons également été amenés à conduire nos propres expérimentations originales et, au fur et à mesure de l'avancée de ce travail de recherche à publier des articles qui sont évoqués en annexe.

Notre hypothèse de départ était que le crowdsourcing et le crowdfunding en particulier pouvaient être profitables aux bibliothèques qui y ont recours. Notre analyse de la littérature nous a permis d'identifier de multiples avantages tant au niveau des coûts que des résultats d'un point de vue à la fois quantitatif et qualitatif. Elle nous a permis aussi d'identifier des inconvénients principalement autour de la question de la qualité des données produites, d'un parfois faible retour sur investissements par rapport aux coûts, et plus généralement, autour de la question du remplacement des professionnels par des bénévoles ou des travailleurs sous payés. Notre observation participante et notre expérimentation nous ont permit également de vérifier ces hypothèses et de valider la faisabilité d'une numérisation à la demande financée par crowdfunding à la condition d'automatiser le calcul du devis de numérisation de chaque document.

Les limites de cette thèse sont liées au périmètre restreint de cette expérimentation, du caractère nécessairement limité de notre implication dans le projet. Il n'a pas été possible non plus d'expérimenter un calcul automatique du

prix des numérisations. Or, cette piste qui nous avait été suggérée par le fondateur de Chapitre.com, Juan Pirlot de Corbion, bien avant la thèse et qui n'a pas pu être expérimentée, reste probablement le meilleur moyen de rendre viable le modèle économique que nous avons expérimenté. Entre le commencement de ce travail de recherche en décembre 2012 et sa finalisation début 2016, la situation du crowdsourcing dans le bibliothèques en France a légèrement évolué. Quelques premières publications et interventions dans des conférences ont vu le jour. La Bibliothèque nationale de France a expérimenté le projet Correct et quelques archives ont eu recours à Wikisource. Nous souhaitons que ce sujet soit encore développé, en particulier en tant que sujet de recherche et espérons y avoir contribué par la publication de cette thèse et d'un certain nombre d'articles.

Au-delà de cette thèse, nous souhaitons avoir la possibilité de mener prochainement de nouvelles expérimentations autour du financement par *crowdfunding* de la publication Gold Open Access de livres ou d'articles, de la *gamification* des tâches les plus ingrates des projets de numérisation, du recours au *crowdsourcing* rémunéré pour extraire des données de textes annotés par *text mining* et enfin, du développement de jeux afin de construire des vocabulaires d'annotation :

Gold Open Access et *crowdfunding*

Une expérimentation d'édition sous un modèle gold open access avec un financement participatif de la libération open access des publications pourrait être mise en œuvre au delà de la thèse avec les éditions Quae auxquelles participe l'Institut National de la Recherche Agronomique. Ce financement, donnant à des réductions fiscales pourrait s'adresser au grand public ou à des mécènes. D'un modèle auteur payeur qui est celui du gold open access, nous passerions ainsi à un modèle mécène payeur.

***Gamification* des tâches ingrates autour des projets de numérisation**

Dans le domaine des bibliothèques numériques, un jeu pour identifier les documents à numériser et contrôler la qualité des livraisons pourrait être proposé.

On pourrait, par exemple, imaginer une pateforme comportant un jeu avec plusieurs parties :

1- Trouver un trésor dans la bibliothèque

Afficher au joueur internaute, de manière aléatoire, des notices bibliographiques d'imprimés antérieurs à 1944. Demander à l'internaute si le document a déjà été numérisé et, si l'ouvrage est postérieur à 1880, quelle est la date de décès de l'auteur ou des auteurs ? Les réponses doivent être données le plus rapidement possible, le joueur est chronométré et voit soumis à son expertise des livres jusqu'à ce qu'il identifie un trésor qui mérite d'être numérisé. Les réponses des joueurs sont confrontées entre elles afin d'obtenir des données de qualité qui viendront enrichir les notices bibliographiques (date de décès, URL du livre numérisé ou date à laquelle le livre numérisé était introuvable) et permettre d'identifier les documents qui méritent d'être numérisés. Une fois un trésor identifié, on passe à la partie suivante.

2- Trouver un indice dans le livre

Afficher au joueur internaute un livre numérisé et lui demander de contrôler chaque page en cochant pour chacune, les points de contrôle classiques (page manquante, page tronquée, page dans le désordre, page en partie ou totalement illisible). Le joueur est de nouveau chronométré et ses données confrontées à celles d'autres joueurs afin d'en assurer la qualité. Son nombre de points dépend aussi à la fois du temps qu'il va mettre et de la coïncidence avec les saisies des autres joueurs.

3- Ajouter le livre dans sa collection

Au cours de cette partie, en un temps limité, il faut cataloguer un maximum de livres numérisés qui se présentent en saisissant également des mots clés pour en décrire le contenu. Les données des internautes sont confrontées afin d'en assurer la qualité. La rapidité et la qualité du travail des joueurs donnent lieu à des points.

4- Rééditer le livre

Au cours de cette partie, on s'inspirera du jeu Digitalkoot ouvert à la fois aux imprimés et aux manuscrits avec un jeu portant non pas sur des taupes cherchant à traverser un pont, mais sur une autre idée

On fera gagner des cadeaux aux meilleurs joueurs.

Crowdsourcing rémunéré et *text mining*

Nous souhaiterions également prolonger ces expérimentations autour du *crowdsourcing* en comparant les résultats obtenus, pour extraire des données au sein d'un corpus de textes, entre une méthodologie classique de *text mining*, une méthodologie de *crowdsourcing* rémunéré et une méthodologie ayant recours au *text mining* pour annoter les textes puis au *crowdsourcing* pour en extraire des données. Notre hypothèse est que les technologies de *text mining* ne sont pas suffisantes pour obtenir de réels résultats et nécessitent l'apport du *crowdsourcing*. Les algorithmes ne peuvent interpréter aussi finement que l'intelligence humaine.

Le *text mining* ou *text data mining* ou encore *text analytics* peuvent être considérés comme relevant d'une démarche inverse de celle qui préside aux projets de *crowdsourcing*. Au lieu de faire appel aux millions de petites mains sur le web, on cherche à utiliser l'intelligence artificielle des machines afin d'extraire du sens au sein de corpus trop vaste pour les travailleurs.

Elle s'en rapproche néanmoins dans le sens où on renonce à confier ce travail à un salarié humain, la tâche dépassant les capacités humaines. D'un point de vue économique, il s'agit d'exploiter des données et des informations déjà produites afin d'en tirer des connaissances nouvelles.

Une comparaison entre les résultats obtenus avec le *text mining* est intéressante.

Le *text mining*, qui permet de structurer de l'information textuelle qui ne l'est pas, peut être utilisé pour catégoriser, clusteriser, extraire des données, produire des taxonomies, analyser les sentiments, résumer des documents, modéliser les relations ou découvrir des relations cachées. Ces applications au domaine de la numérisation sont donc prometteuses.

La relation entre *crowdsourcing* et *text mining* est dialectique. Avec le *crowdsourcing*, on obtient une annotation manuelle des textes et des images sans a priori. Avec le *text mining*, on obtient une annotation automatique des textes et

des images à partir de ressources et de vocabulaires qui devraient a priori se trouver dans les corpus.

Un jeu pour développer des vocabulaires d'annotation

De la même manière, dans la mesure où la construction de vocabulaires d'annotation est coûteuse et chronophage le recours à des internautes joueurs pourrait être testé afin de les construire. Ainsi, dans le cadre d'une démarche alliant *crowdsourcing* et *gamification*, nous pourrions proposer un jeu aux internautes. A partir d'une liste de mots déjà existante, on pourrait soumettre chacun de ces mots à des internautes et leur demander des synonymes, puis des termes génériques, puis des termes spécifiques. Par double saisie, les mots proposés par les internautes seraient validés puis considérés comme des mots tabous, c'est à dire interdit à la proposition, afin de contraindre les internautes à chercher des mots moins évidents. Le score des joueurs serait calculé en fonction du nombre de mots proposés dans la limite d'un compte à rebours. Les meilleurs joueurs pourraient remporter des cadeaux.

Au delà de ces perspectives de nouvelles recherches autour du *crowdsourcing*, de notre point de vue, les organisations qui mettront en place les premières des démarches participatives pourront se démarquer et en tirer un avantage concurrentiel, mais lorsque cette pratique se sera généralisée, il est fort probable que la *gamification*, la récompense et la rétribution deviennent les seuls moyens de capter la participation des internautes. Mais, comme le suggère Rose Holley (Holley, 2010), les bibliothèques auraient grand intérêt à mutualiser le *crowdsourcing* au lieu d'entreprendre des démarches individuelles et concurrentielles chacune de son coté. Elles parviendraient ainsi à capter d'avantage de volontaires en offrant d'avantage de contenus à corriger.

Les bibliothèques ont déjà perdu le monopole d'intermédiaire obligatoire entre l'information et le public. Les conservateurs de bibliothèques pourraient donc mal vivre le fait de voir leurs compétences relativisées et mises sur le même plan que celles du grand public et avoir l'impression qu'on fait n'importe quoi avec

n’importe qui, que la hiérarchie producteur / consommateur, professionnel / amateur s’efface, et que le travail rémunéré de bibliothécaire est remplacé par du travail gratuit ou « ubérisé » de bénévoles. Néanmoins, cette “ubérisation” des bibliothèques semble inéluctable et pourrait aussi être très prometteuse.

Annexe 1 : Autres projets non évoqués dans le chapitre 2 (Panorama des projets de *crowdsourcing* appliqués à la numérisation des bibliothèques)

1- Mise en ligne et curation participatives : Internet Archive

Fondée dès 1996, Internet Archive est une organisation internationale non gouvernementale et à but non lucratif dont le siège se trouve à San Francisco. Internet Archive, qui propose la 2ème plus importante bibliothèque numérique en libre accès dans le monde, est financée, pour un budget annuel de 15 millions de dollars, par les dons des internautes et par les fondations Alexa Internet, Kahle/Austin Foundation, Alfred P. Sloan Foundation, William and Flora Hewlett Foundation. Des subventions fédérales de l'État américain sont également fournies pour la bande passante. Sur Internet Archive, on trouve, outre des livres numérisés, des enregistrements sonores, des vidéos, des jeux vidéos et des données de la recherche.

En France, la Bibliothèque Sainte-Geneviève, puis l'Institut National de la Recherche Agronomique et Sciences Po y participent. Il reste, en effet, impossible pour ces bibliothèques, comme pour d'autres institutions extérieures à la Bibliothèque nationale de France, d'héberger leurs documents numérisés directement sur Gallica. Cette impossibilité s'explique par le workflow trop rigide d'une bibliothèque numérique dont les métadonnées sont adossées sur le seul catalogue de la Bibliothèque Nationale, et peut être aussi d'une crainte de pollution par des contributions extérieures. Les bibliothèques peuvent toutefois avoir leurs contenus moissonnés par Gallica, mais ce moissonnage OAI-PMH est limité aux seules métadonnées, il ne comprend pas les fichiers numériques et il est conditionné par la nécessité, pour les bibliothèques, de développer leurs propres bibliothèques numériques avec leurs propres entrepôts OAI de métadonnées. Or, développer sa propre bibliothèque numérique est très coûteux en moyens humains, matériels et financiers pour des résultats généralement médiocres en termes de fonctionnalités, de pérennité et de visibilité. Pour toutes ces raisons, la

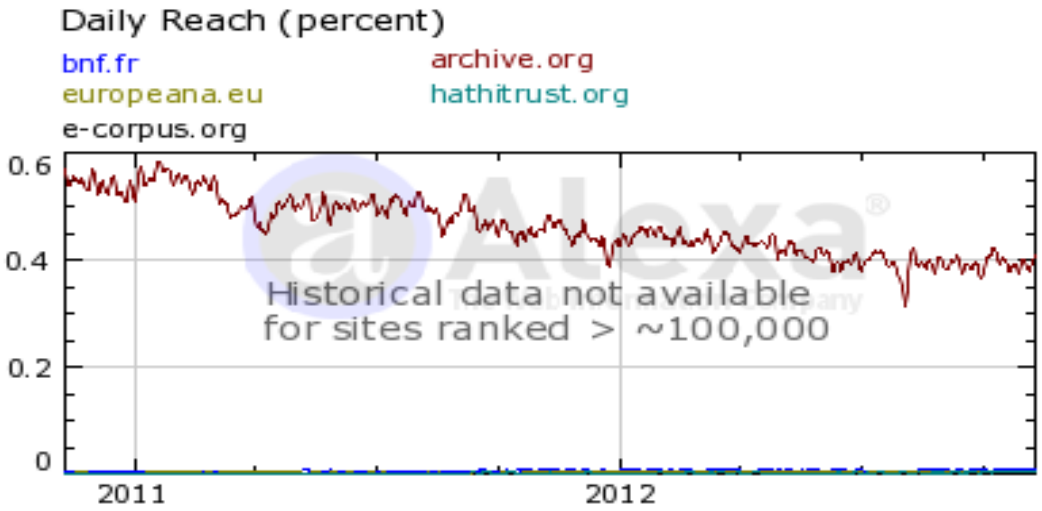
Bibliothèque Sainte-Geneviève a fait le choix, dès 2010, de verser le produit de sa numérisation dans Internet Archive.

Les particuliers ont également cette possibilité. On peut ainsi parler de “mise en ligne participative” et de *crowdsourcing*. Ainsi, plus d’un million de livres numérisés par Google dans le cadre de son programme de numérisation Google Books a été versé sur Internet Archive sous une licence Public Domain Mark et ce en toute légalité puisqu’il s’agit de livres libres de droits et que les auteurs des photographie des pages n’y ont pas laissé de marques de leurs personnalités artistiques. D’autres particuliers ont aussi numérisé avec leur scanner leurs propres documents tombés dans le domaine public.

La mise en ligne d’un livre numérisé est assez simple. On commence par transférer les fichiers numérisés puis on en saisit les métadonnées dans un formulaire web. La plateforme engendre automatiquement l’OCR de ces fichiers et aussi des fichiers EPUB et MOBI pour lectures sur des liseuses. Internet Archive présente ainsi l’avantage d’offrir systématiquement des livres téléchargeables sous forme de fichiers mobi pour la liseuse Kindle ou sous forme de EPUB pour toutes les autres marques de liseuses. Une offre quasiment inexistante dans les bibliothèques numériques françaises.

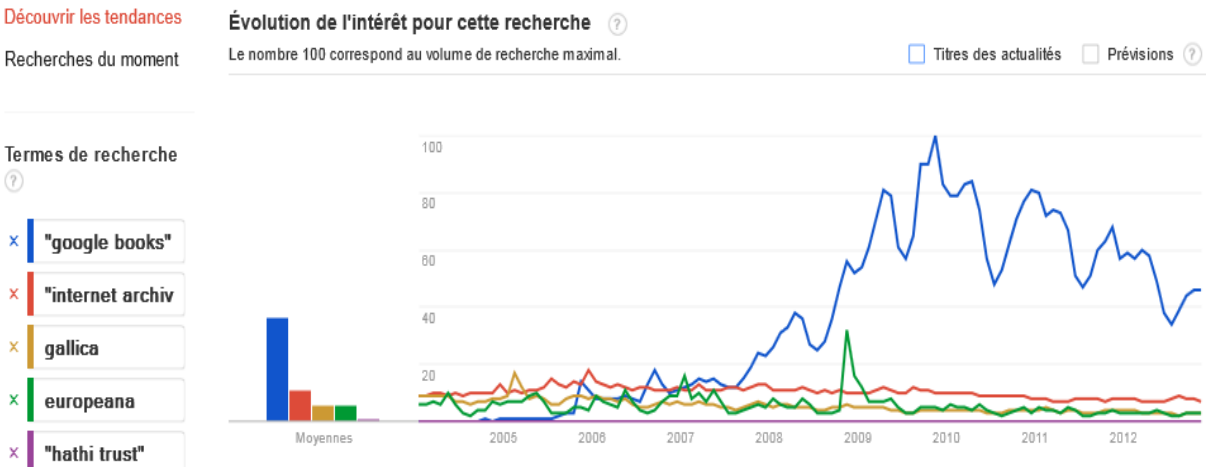
Au 25 octobre 2012, il y avait 3 678 804 livres sur Internet Archive. Sur la totalité de l’année 2011, il y a eu 227 244 392 téléchargements soit une moyenne de 18 937 033 téléchargements par mois, soit une moyenne de 5,15 téléchargements par livre et par mois minimum.

D’après les données du site alexa.com, ces statistiques sont bien supérieures à celles des autres bibliothèques numériques, en dehors de Google Books dont les statistiques ne sont pas accessibles :



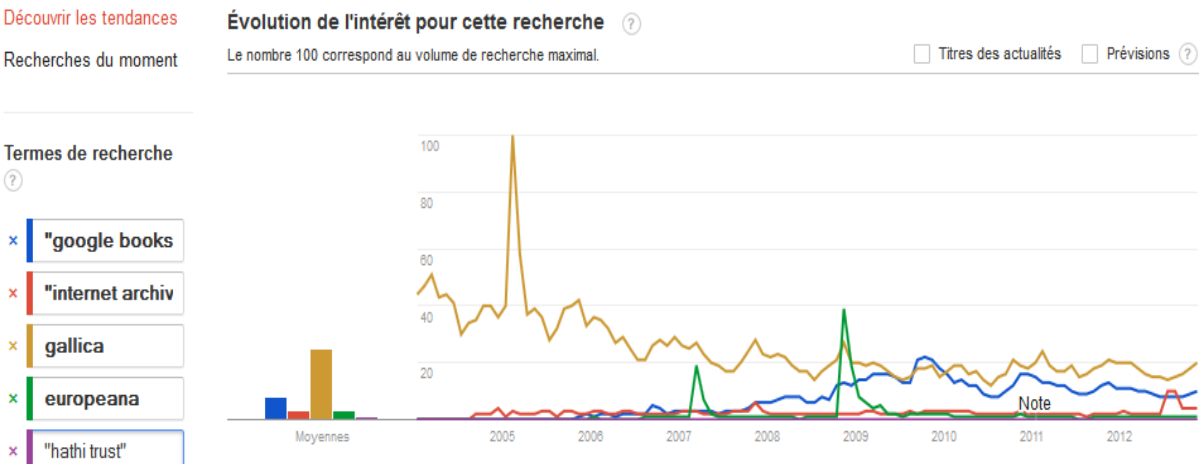
Comparaison du trafic web sur archive.org, bnf.fr, europeana.eu, hathitrust.or et ecorpus.org d'après le site alexa.com

Néanmoins, d'après les statistiques mondiales du nombre d'internautes ayant saisi Google Books dans le moteur de recherche Google, il semble qu'ils soient bien plus nombreux que ceux qui ont saisi Internet Archive, d'après Google Trends :



Statistiques du nombre d'internautes dans le monde ayant saisi "Google Books" "Internet Archive" Gallica, Europeana ou "Hathi Trust" dans Google d'après Google trends

Néanmoins, concernant la France, Gallica resterait en position dominante :



Statistiques du nombre d'internautes en France ayant saisi "Google Books" "Internet Archive" Gallica, Europeana ou "Hathi Trust" dans Google d'après Google trends

Internet Archive comprend également le projet OpenLibrary, un projet de catalogage participatif et fait appel, pour le projet Gutenberg, à l'effort de correction de l'OCR par les bénévoles de Distributed Proofreaders. D'après un article de Actualitté publié en octobre 2013, la Quadrature du Net, une "association de défense des droits et libertés des citoyens sur Internet", serait en train de développer, un logiciel reCaptcha-like pour Internet Archive.

Internet Archive est la première organisation à payer le salaire de ses collaborateurs sous la forme de BitCoins. Le BitCoin est une monnaie virtuelle. Chaque porte-monnaie est identifié par une adresse URL. Les transactions sont des transferts de valeurs entre des adresses Bitcoin, à l'image de l'échange de courriels. Une banque collaborative et à but non lucratif a été édifiée, l'Internet Archive Federal Credit Union (IAFCU) basée à New Brunswick dans le New Jersey : https://iafcu.org (consulté le 23 juin 2016 mais le site n'est plus accessible)

Pour un panorama exhaustif de collections numériques alimentées par les internautes, nous pourrions également évoquer Commons Wikimedia, la médiathèque de Wikipédia qui permet à des institutions d'ouvrir leurs portes à des wikipédiens pour qu'ils puissent effectuer des prises de photos d'objets

anthropologiques (Muséum de Toulouse) ou de salles de lecture (Bibliothèque Sainte-Geneviève, par exemple) ou encore le projet Picture Australia qui a réunit 2641 bénévoles qui ont versé leurs archives photos sur Flickr, le projet Wir waren so frei qui a récolté des photos de la chute du mur de Berlin, le projet Open Call - Brooklyn Museum qui a consisté à organiser une exposition à partir des photos proposées par des internautes, le projet anglais “Pin-a-tale”, le projet américain “Make history” de récolte de témoignages, photos, vidéos et autres documents sur le 11 septembre 2001, ou encore “Click! A Crowd-Curated Exhibition”, un projet du même Brooklyn Muséum. Sur le thème “The Changing Faces of Brooklyn”, des photographes proposaient leurs œuvres aux internautes qui les évaluaient et créaient ainsi leur propre exposition. 3 344 personnes y ont participé et ont générés 410 089 évaluations. Leur évaluation coïncidait souvent avec celle des experts. En France, des initiatives assez similaires ont également fait appel aux particuliers, en particulier aux archives du Débarquement en Provence, à celles de la Grande guerre, sur la Shoah, et aux Archives départementales de Seine-Maritime. Des événements de numérisations collectives in situ qualifiées de “ExtravaSCANza” ont été proposés par les Archives Nationales des USA en partenariat avec Wikimedia. La sélection de documents (peintures, livres, photos…) peut aussi utilement être organisée par des internautes, sur le modèle du Museum of Online Museums, sous une forme de “curation” participative comme le relate (Terras, 2010). Enfin, nous pouvons évoquer la collecte de documents d’archives autour de l’attentat contre Charlie Hebdo organisée par la Bibliothèque de Harvard[43] (Breton, 2015).

2- la numérisation à la demande sous forme de *crowdfunding* appliquée aux bibliothèques numériques

2.1- Le livre à la carte, Phénix Éditions

Le livre à la carte (Libris Éditions) a été lancé, en tant qu’expérimentation, en mai 1997 à la Bibliothèque nationale de France afin de permettre la réimpression

[43] http://cahl.webfactional.com (consulté le 23 juin 2016)

d’ouvrages sous la forme de fac-similés. De nombreuses limites ont été données au projet dès son commencement, ce qui témoigne probablement de certaines réticences. Par exemple, seuls les livres du département des imprimés ont été concernés, les documents de la Réserve des ouvrages rares de la BnF étaient donc d’emblée exclus du projet. La réimpression se faisait sous la forme d’imprimés brochés ou avec une reliure cartonnée. 140 commandes ont été passées entre le 25 mai et le 25 août 1997.

Sur le même modèle, et comme relaté dans (Delcourt, 2001), la société Librissimo (devenue Phénix Éditions lors de son intégration par Alapage, France Telecom) fondée par Henri Le More, a sollicité, quelques années plus tard, dès 1999 la bibliothèque municipale de Troyes afin de lui proposer la numérisation à la demande de ses collections et d’en produire des fac-similés. Il s’agissait, à nouveau, d’un service de production de fac-similés à la demande au moyen de la numérisation, mais la production d’exemplaires numériques seuls ne semble pas avoir été envisagée. L’impossibilité de séparer le service de numérisation à la demande de celui d’impression à la demande, y compris en terme de coûts, explique peut-être en partie que le projet ait rencontré un succès limité. Le coût était ainsi de 3 à 4 F par page, soit 600 F à 800 F pour un livre de 200 pages. Toutes les demandes de devis n’aboutissaient pas à des commandes.

Néanmoins, le volume de commandes pour la Bibliothèque de Troyes semble avoir été, début 2001, d’une centaine de commandes par mois, un chiffre non négligeable qui a permis à la Bibliothèque de développer sa bibliothèque numérique sans coûts en dehors de ceux liés aux opérations pour sortir et rentrer les ouvrages. Un atelier de numérisation sur place était installé à la Bibliothèque municipale de Lyon et envisagé pour Troyes. Phénix Éditions avait conclu également un partenariat avec l’Ecole polytechnique, Le Saulchoir et Poitiers. Henri Le More, le fondateur de Phénix Éditions, qui a été rencontré en 2006 puis en 2009 aurait depuis cette expérience rejoint les amis de la BnF.

2.2- Juan Pirlot de Corbion : de Chapitre.com à YouScribe

Juan Pirlot de Corbion, fondateur de Chapitre.com puis de Youscribe.com a été rencontré à 3 reprises en 2009. Il a eu une expérience avec la Bibliothèque nationale de France qui avait fournit une partie non négligeable de son catalogue. 400 000 notices bibliographiques avaient ainsi été exposées sur un portail commercial afin de permettre aux internautes de commander la numérisation à la demande de documents. D'après des informations informelles recueillies oralement lors d'entretiens avec Juan Pirlot de Corbion, le projet aurait généré une demande comprise entre 50 et 100 demandes de devis par jour. Néanmoins, malgré ce succès, le projet a été contraint d'avorter. Les demandes étaient trop nombreuses pour les effectifs humains de la Bibliothèque nationale de France. Et la bnF devait, pour toutes ces demandes, vérifier la présence des documents et la possibilité matérielle de les numériser, les extraire de ses collections, et les décrire matériellement (nombre précis de pages, angles d'ouvertures des reliures, formats des livres…). Ces détails étaient nécessaires pour déterminer le matériel de numérisation à utiliser, mesurer les volumétries et ainsi, évaluer le coût de numérisation. Dans ces conditions, le personnel a rapidement manifesté son hostilité pour un travail si contraignant, si peu valorisant et si coûteux en ressources humaines. De surcroît, le délai moyen entre la demande de devis de l'internaute et sa satisfaction était très long, puisqu'elle était suspendue à l'étape de demande de devis et, en particulier par la description matérielle que devait fournir la Bibliothèque Nationale. Lorsque le client recevait enfin un devis, il n'était pas rare que son envie d'achat soit passée ou qu'il soit surpris et découragé par le prix de la prestation. Dans ces conditions, un nombre important de demandes de devis n'aboutissaient pas à une commande, ce qui a eu pour effet de démotiver le personnel de la BnF et l'équipe du projet.

2.3- Adopter un livre sur Gallica

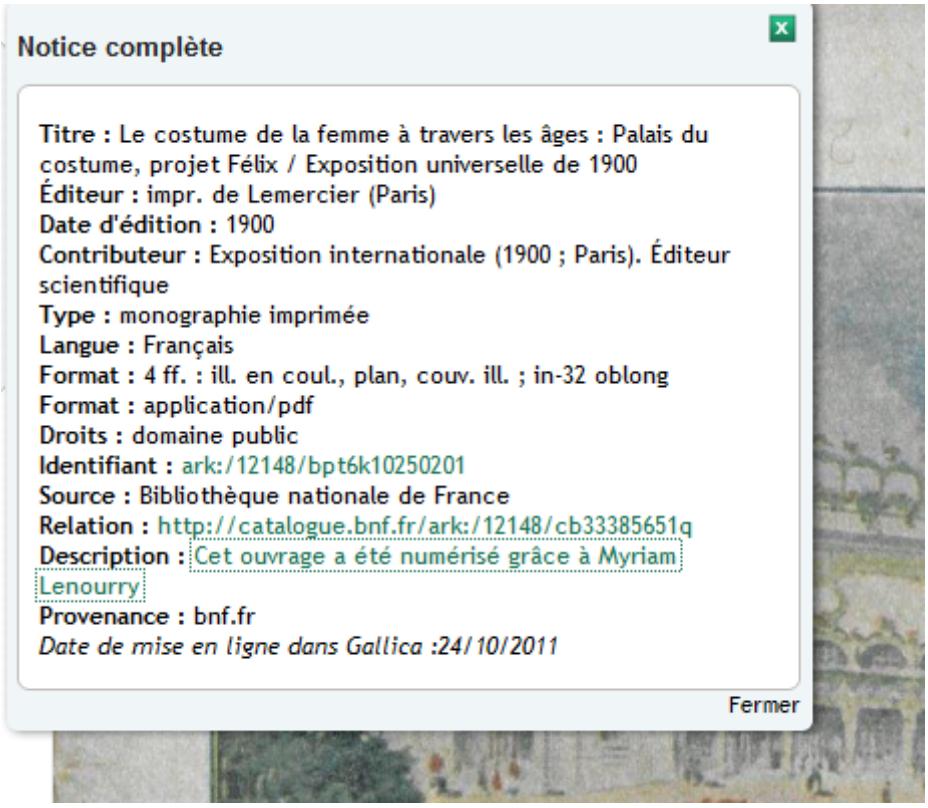
Le site web des Amis de la Bibliothèque nationale de France propose aux internautes de suggérer la numérisation de livres conservés à la BnF. Une fois signalé et après vérification que le document n'a pas déjà été numérisé et qu'il

peut l'être du point de vue juridique, chaque suggestion viendra compléter une liste des suggestions des internautes. Cette dernière liste reste séparée d'une autre liste qui comprend les suggestions émanant de la BnF. Le fait de ne pas mélanger les 2 listes reflète vraisemblablement la réticence des conservateurs de la BnF à partager la politique documentaire de la célèbre bibliothèque numérique avec le grand public.

Au 13 février 2013, on compte 66 livres proposés par les internautes et 537 proposés par la Bibliothèque nationale de France autour des thématiques suivantes : Les femmes (Féminisme ancien, Féminisme moderne, Condition féminine, Modes et costumes), Panorama du 19è siècle, Les Livres de sciences naturelles, Les grandes entreprises françaises, Le tour du monde en 80 livres, La vigne et le vin, La gastronomie (Paradis artificiels, Littérature, Produits de la bouche, Table). Ces quantités peuvent apparaître relativement modestes si l'on considère que ce service a été proposé depuis mars 2011.

Sur 22 mois, les amis de la BnF ont donc bénéficié de seulement 3 suggestions par mois en moyenne. Les documents qui sont proposés à l'adoption vont de 1556 à 1939, la moyenne tourne autour de 1836. 168 numérisations ont été financées par des internautes, soit une moyenne de 7,64 numérisations par mois. Aux journées des pôles associés de la BnF, le chiffre de 24 834 pages a été communiqué. Ces quantités plus importantes sont surtout le fruit des campagnes de communication de la BnF.

Sur la notice du document numérisé dans Gallica, apparaît, pendant 10 ans, le nom de la personne ayant financé la numérisation du document :



Notice bibliographique de la BnF affichant
le nom de la personne ayant financé la numérisation de l'ouvrage

Le délai entre le paiement et la mise en ligne semblent être d'environ 6 mois, mais un ouvrage dont la numérisation avait été commandée le 7 décembre 2011 n'était toujours pas en ligne le 13 février 2013, plus d'un an plus tard. D'autres, commandés le 4 mars, le 5 avril ou le 14 mai 2012 n'étaient toujours pas numérisés le 13 février 2013. L'intégration dans les trains de numérisation de masse de la BnF de ces numérisations à la demande doit donc probablement poser des difficultés d'organisation interne.

Le prix de la prestation oscille entre 5, 60 € et 16 406, 60 € pour une moyenne de 576,55 € au 13 février 2013. Le coût réel, après déduction fiscale oscille entre 5,71 et 5578,24 € pour une moyenne de 196,03 € par livre. Il est, en effet, possible de déduire de ses impôts 66 % de la somme versée pour les particuliers et 60 % pour les entreprises. Comme le rappelle la BnF sur son site, l'Instruction fiscale du 13 juin 2005 commentant l'article 127 de la loi de programmation pour la cohésion

sociale relatif à la réduction d'impôt au titre des dons aux œuvres versés par les particuliers (BOI 5 B-17-05, n°1014 du 13 juin 2005) stipule que « aux termes de l'article 200 du CGI, les dons et versements effectués au profit d'œuvres ou d'organismes d'intérêt général ayant un caractère philanthropique, éducatif, scientifique, social, humanitaire, sportif, familial, culturel, ou concourant à la mise en valeur du patrimoine artistique, à la défense de l'environnement naturel ou à la diffusion de la culture, de la langue et des connaissances scientifiques françaises ouvrent droit à une réduction d'impôt égale à 66 % du montant des sommes versées retenues dans la limite d'un plafond égal à 20 % du revenu imposable ».

Mais cela représente une somme encore importante. Si on considère les 168 livres dont la numérisation a été financée par les internautes, ce mode de financement par *crowdfunding* a rapporté à la BnF 168 x 576, 55 € = 96 860 € soit une moyenne de 4405 € par mois.

Les prix sont affichés a priori et nécessitent une description matérielle de chaque document (format, nombre de feuillets, angle d'ouverture, état...) afin d'en obtenir un devis auprès du délégataire. Ce mode de fonctionnement, par devis, limite probablement le nombre de documents proposés.

L'apparition d'un bouton de type ebooks on demand sur les notices de livres antérieurs à 1880 pour des raisons juridiques dans le catalogue de la BnF permettrait vraisemblablement d'augmenter de manière significative le nombre de commandes. En effet, la visibilité du service reste beaucoup plus faible sur le site des amis de la BnF que sur son catalogue informatisé de bibliothèque ou sur Gallica.

Sur le catalogue de la BnF, il est, en revanche possible de commander des reproductions intégrales ou partielles de documents pour 45 € par document de plus de 60 pages et 0,7 € la page concernant les livres de moins de 60 pages.

Il est à noter, pour finir, que la Bibliothèque nationale de France, a également fait appel aux dons et que, en 2012, cela lui a permis d'acquérir le livre d'heures de Jeanne de France de 1452, puis, en 2014, grâce à 2400 donateurs, de totaliser 300 000 € afin d'acquérir le manuscrit de François 1er Description des douze

césars avec leurs figures. Les généreux donateurs ont ensuite été invités au vernissage.

2.4- Le projet *crowdfunding* Numalire

Nous avons déjà largement évoqué le projet Numalire dans le chapitre dédié aux expérimentations. On trouvera néanmoins ici une présentation plus générale et synthétique du projet.

Numalire est un projet fondé par la société YABé fondée par Filippo Gropallo et Denis Maingreaud et hébergée, à ses débuts, au Labo de l'Édition à Paris, une pépinière d'entreprises innovantes et un lieu de rencontres où des événements et des conférences autour de l'édition innovante sont organisés. Le service propose de faire numériser et rééditer à la demande des documents anciens conservés au sein des bibliothèques sous la forme de numérisation à la demande, mais aussi de souscriptions participatives. Le site web a été ouvert le 7 octobre 2013 dans le cadre d'une expérimentation de 8 mois avec 8 bibliothèques : la Bibliothèque des Arts Décoratifs, la Bibliothèque Historique de la Ville de Paris, la Bibliothèque de l'Hôtel de Ville de Paris, la Bibliothèque Forney, la Bibliothèque de l'Académie nationale de médecine, la Bibliothèque Marguerite Durand, la Bibliothèque Sainte-Geneviève et les bibliothèques de l'Institut National de la Recherche Agronomique afin de tester ce modèle économique et, en fonction des résultats, de l'étendre à la France et à l'International.

Accédant aux métadonnées de l'ouvrage qui l'intéresse sur le site Numalire parmi 500 000 notices bibliographiques, via le catalogue d'une bibliothèque ou directement grâce à un moteur de recherche, chaque internaute peut ainsi proposer une souscription pour en financer la numérisation et est invité à mobiliser son réseau social pour y parvenir jusqu'à ce que la somme nécessaire à la numérisation de l'ouvrage soit réunie. Les souscripteurs peuvent bénéficier de e-book numérisés par la société Arkhênum basée dans la région de Bordeaux et de documents en *print on demand* produits via la société SoBook localisée à Roubaix. Les souscripteurs sont remerciés sur le site, sur la version papier et peuvent également l'être sur les bibliothèques numériques développées par les

bibliothèques. De “devisable” (possibilité de demander un devis et d’en suggérer ainsi la numérisation), l’ouvrage passe ainsi au statut “devis en cours” puis au statut “non numérisable” ou “souscription” en fonction de l’examen effectué sur le document par les bibliothèques qui en vérifient l’état, la complétude, la possibilité juridique de le numériser et l’inexistence de version déjà numérisée puis au statut final de “livre numérisé” si la souscription a été réunie. Le devis est produit sur la base des descriptions matérielles apportées par les bibliothèques (nombre de feuillets, formats, angles d’ouvertures…).

La bibliothèque bénéficie d’un intéressement. Le retour sur investissement pour la société Yabé, imaginé dans un premier temps, à partir de la vente de EPUB, se fera finalement via une marge sur la numérisation et la vente de *Print on Demand* (POD). Les documents numérisés ne seront donc pas vendus mais seront communiqués aux souscripteurs et diffusés par les bibliothèques sous licence Public Domain Mark.

En huit mois d’expérimentation, le site web du projet a attiré 70 000 utilisateurs qui ont consulté 115 000 pages. Parmi eux, 55 899 provenaient de France. Au total, un tiers venaient de région parisienne, un tiers de province et un tiers de pays étrangers comme la Belgique, le Canada, la Suisse, les USA, l’Italie, l’Algérie, l’Allemagne, le Maroc et l’Espagne. 91 % du trafic web provient des moteurs de recherche, seulement 5 % de liens directs, 3 % des bibliothèques et 1 % des réseaux sociaux. Si le nombre de consultations a connu son apogée en novembre 2013, il a connu une diminution sensible à partir de mi-décembre 2013. Un moins bon PageRank pourrait en être la cause. Néanmoins, l’impact sur le nombre de commandes est resté finalement limité puisque le taux de transformation nombre de visites / nombre de numérisations s’est même amélioré. Une enquête de satisfaction auprès des utilisateurs du service conduite en juin 2014 auprès de 380 personnes et ayant généré 118 réponses a permis de révéler que 70,59 % des répondants étaient des hommes, qu’ils appartenaient à des classes sociales élevées (universitaires, étudiants, cadres supérieurs). 51,43 % d’entre eux ont effectué leur démarche pour leur travail.

Au 27 novembre 2013, il y avait eu 150 demandes de devis, 80 souscriptions en cours et 5 souscriptions abouties. A la fin des 8 mois d'expérimentation, sur 414 demandes de devis, il n'y en avait eu que 36 s'étant concrétisés avec le financement d'une numérisation, soit 11 % seulement des devis demandés. En considérant qu'il a fallut une demi heure pour savoir si le livre demandé n'était pas déjà numérisé, si ses auteurs étaient morts il y a plus de 70 ans, pour aller chercher l'ouvrage, mesurer son format, son angle d'ouverture, et compter le nombre de ses feuillets, il aura donc fallut que les bibliothèques consacrent 207 heures d'un travail assez ingrat au projet pendant les 8 mois de l'expérimentation. Si on considère qu'un personnel de catégorie C coûte 30 € de l'heure à l'État, on peut considérer que le projet a coûté 6210 € à l'État, soit une moyenne de 6210 / 36 = 172 € par livre. L'argent non dépensé par les bibliothèques afin d'assurer ces numérisations, l'est donc sous la forme de temps de travail.

On constate également que la plupart des souscriptions ne concernent finalement qu'un seul souscripteur et que les souscriptions collectives restent relativement marginales avec 14 % seulement du total. Pour 20 % des répondants à l'enquête, le partage de la souscription est trop compliqué. Pour 48 % d'entre eux, le prix de la numérisation reste trop élevé, d'autant que pour 16,67 % d'entre eux, le prix est plus cher par rapport à celui pratiqué par les bouquinistes et qui permettent d'obtenir un original papier plutôt qu'une reproduction électronique.

Le projet Numalire a fait l'objet de nos expérimentations dans le cadre de cette thèse. Nous avons, en particulier, recherché des solutions pour diminuer le coût pour les bibliothèques et améliorer le rapport entre nombre de demandes de devis et nombre de commandes de numérisations.

2.5- revealdigital.com et Lyrasis

Fondé par Jeff Moyer (Ann Arbor, Michigan) et lancé en janvier 2013 à l'occasion du congrès de l'American Library Association, revealdigital.com est un projet de *crowdfunding* appliqué à la numérisation des bibliothèques basé à Saline dans le Michigan et soutenu par Lyrasis un projet qui propose de numériser les documents papier mis à disposition par les internautes et qui a travaillé en partenariat avec

Internet Archive (https://archive.org/details/lyrasis, consulté le 23 juin 2016). Les documents numérisés et mis en ligne sur archive.org portent la mention “Digitizing sponsor: Lyrasis Members and Sloan Foundation”. La Sloan Foundation est une fondation philanthropique à but non lucratif fondée par Alfred P. Sloan, un ancien dirigeant de la société General Motors.

L’initiative propose de « révéler les trésors cachés des bibliothèques ». Les bibliothèques peuvent proposer des documents à la numérisation et paramétrer les priorités. Lorsque la souscription est atteinte et après une période d’embargo de 2 ans, le document est mis en open access sur un modèle Gold Open Access proche de Unglue.it. Dans le cas contraire, cette période d’embargo sera déterminée par la société Reveal Steering Group. Dans tous les cas, les bibliothèques contributrices auront un accès réservé aux documents numérisés via la plateforme Reveal Digital développée à partir du logiciel iFactory. Au final, selon (Rathemacher, 2015), le coût de l’opération pour les bibliothèques ne représenterait plus que 20 % des coûts de projets de numérisation menés classiquement, en dehors de revealdigital.

A la différence d’autres projets, la sélection des documents dont la numérisation est proposée a été faite en amont avec production de devis. La politique de sélection des documents méritant d’être numérisés n’est donc pas partagée. revealdigital.com ne réclame pas de droits de propriété sur les documents numérisés.

Ce modèle économique est inspiré de l’Open Access qui vise à favoriser la diffusion libre et gratuite la plus large possible des publications de la recherche scientifique. L’Open Access est divisé en deux grands modèles, le Green et le Gold Open Access. Le modèle Green Open Access consiste à ce que, au delà de la publication chez un éditeur, les auteurs déposent aussi leurs publications sur des archives ouvertes ou institutionnelles, généralement après une période d’embargo demandée par l’éditeur et dans des versions ne conservant pas la mise en forme qui demeure la propriété de l’éditeur à part si ce dernier autorise de déposer sa version finale. Cette diffusion permet une diffusion libre et gratuite des résultats de la recherche scientifique, mais aussi une meilleure visibilité et un plus

fort taux de citation des articles. Avec le modèle Gold Open Access, c'est l'éditeur qui propose à l'auteur de payer pour que son article soit accessible à tous librement et gratuitement. Ce ne sont donc plus les lecteurs qui achètent les accès aux articles mais les auteurs et leurs institutions qui financent la diffusion en Open Access des articles et ce, directement sur le site de l'éditeur, sans embargos ni pertes de mise en page. Le projet revealdigital s'inspire de ce dernier modèle. Le projet numérise les collections des bibliothèques, leur donnent accès aux documents numérisés et ouvrent ensuite une souscription pour financer les frais.

2.6- Le projet FeniXX

D'après le site du projet, il a été créé "en juillet 2014 par le Cercle de la Librairie, sous l'égide du syndicat national de l'édition et en partenariat avec le Ministère de la Culture". La société Jouve y participe également.

Une société de gestion collective du nom de FeniXX et représentant à la fois les éditeurs et les auteurs va pouvoir commercialiser des numérisations parmi les près de 500 000 ouvrages indisponibles du 20è siècle inscrits sur la base ReLIRE de la Bibliothèque nationale de France (https://relire.bnf.fr, consulté le 23 juin 2016). Le Registre des Livres Indisponibles en réédition électronique est une disposition qui provient de la loi du 1er mars 2012 relative à l'exploitation numérique des Livres Indisponibles du 20è siècle. Son objectif est de permettre la numérisation les ouvrages français non encore tombés dans le domaine public (publiés entre le 1er janvier 1901 et le 31 décembre 2000) tout en n'étant plus exploités commercialement

La société FeniXX sera chargée de percevoir les droits patrimoniaux des auteurs et de les répartir. Les auteurs et les éditeurs peuvent s'y opposer éventuellement afin d'exploiter ces œuvres en dehors de la société FeniXX. Dans le cas contraire, s'ils acceptent les conditions de la société, ces œuvres indisponibles pourront alors être numérisées par et exploitées commercialement par de tierces personnes sous licences d'exploitation non exclusives. Les éditeurs et les auteurs percevront au moins la moitié des droits. Des ventes d'impression à la demande sont également prévues dans le dispositif.

Le catalogue sera distribué via la plateforme Eden Livres et les ouvrages numérisés seront diffusés sous forme d'extraits via Gallica et l'éditeur CAIRN. Un volet Prêt Numérique en Bibliothèque fera également partie du projet.

Pour un panorama plus complet, nous pourrions évoquer également le site unglue.it qui propose aux internautes des souscriptions *crowdfunding* afin de permettre la publication en Open Access de livres, le projet Maine Shared Collections Strategy (MSCS - maineinfonet.org) lancé en 2010 sur le modèle de Ebooks on Demand, adossé sur Internet Archive et sur Hathi Trust et qui propose également un service de *Print on Demand* (Revitt, 2013) ou l'initiative de février 2013 de la Duke University Library, l'International Amateur Scanning League (http://www.scanningleague.org, consulté le 23 juin 2016) fondée par Carl Malamud, qui propose de mobiliser des bénévoles afin de numériser le patrimoine et qui leur offre des récompenses : une médaille Katharine Graham pour 10 DVD gravés, une médaille Thomas Edison pour 75 et une médaille Sonia Sotomayor pour 100 DVD. En France, nous aurions pu évoquer le service de numérisation à la demande des Archives Départementales des Hautes Alpes et, dans un domaine un peu plus éloigné, celui du *crowdfunding* appliqué à la restauration de reliures, nous aurions pu évoquer le projet de l'Ecole Nationale des Chartes "Sauvez nos reliures" qui, au 14 avril 2015, a permis de mobiliser 65 mécènes et de restaurer 50 ouvrages.

3- L'impression à la demande (*Print on Demand*, POD)

3.1- Electronic Library (eLib) et Higher Education Resources ON Demand (HERON)

Electronic Library est un projet écossais doté d'un budget de 20 millions de livres sterling développé depuis 1994 à l'initiative de l'Université de Stirling et en partenariat avec Napier University, South Bank University, et Blackwell's Bookshops et Blackwell's Information Services.

Dans le programme, était mentionné un volet "on demand publishing in the humanities" portée par la Liverpool John Moores University. (Rusbridge, 1995).

3.2- Amazon BookSurge (CreateSpace)

Les premiers projets de *print on demand* appliqués aux projets de bibliothèques numériques ont été portés par Amazon BookSurge devenu depuis CreateSpace.

Le 23 mars 2009, un partenariat était conclu entre Amazon BookSurge et la bibliothèque de l'Université de Cornell (USA). Ce partenariat ouvrait à Amazon la possibilité d'utiliser les livres numérisés par la bibliothèque pour produire, à partir des documents numériques, des livres imprimés à la demande et pour les vendre. En contrepartie, une part des bénéfices réalisés par la commercialisation de ces livres imprimés était reversée à la bibliothèque.

En février 2010, la célèbre British Library annonçait que 65 000 livres numérisés et libres de droits seraient commercialisés sous forme d'imprimés à la demande via Amazon BookSurge. En octobre 2010, la non moins célèbre Bibliothèque du Congrès signalait, à son tour, proposer 50 000 livres disponibles en *print on demand*.

3.3- Gallica et *Print on Demand*

Pour sa part, la Bibliothèque nationale de France a noué un partenariat avec Hachette depuis le 31 juillet 2013. La réimpression, en noir et blanc, de 180 000 livres publiés avant 1900 et de 10 000 partitions de musique peut être commandée par l'intermédiaire de l'entreprise Lightning Source. Entre janvier et octobre 2013, 20 000 impressions à la demande auraient ainsi été commandées comme l'explique BnF-Partenariats dont les propos ont été repris par (Klopp, 2014).

Auparavant, la Bibliothèque Nationale est également devenue partenaire de la société Chapitre.com depuis mars 2011. Les impressions commandées depuis Chapitre.com sont ensuite produites par les Éditions du Net. Les délais de livraison n'excéderaient pas 4 jours. Le prix final moyen serait situé entre 15 et 20 euros mais il pourrait aller de 6 € minimum à 40 € maximum. Chapitre.com toucherait 25 % des revenus tirés de ce service. La part reversée à la Bibliothèque nationale de

France n'a pas pu être connue. Quoi qu'il en soit, le nombre de vente serait de 5 à 6 commandes par jour d'après les chiffres qui ont été communiqués dans la presse au cours des premiers mois de lancement du service.

Enfin, les sociétés Edilivre et The Book Edition, sont également partenaires de la BnF pour l'impression à la demande. Ces sociétés ont recours aux services de So Book pour l'impression.

L'un des nombreux ouvrages dont l'impression à la demande peut être commandée sur Gallica

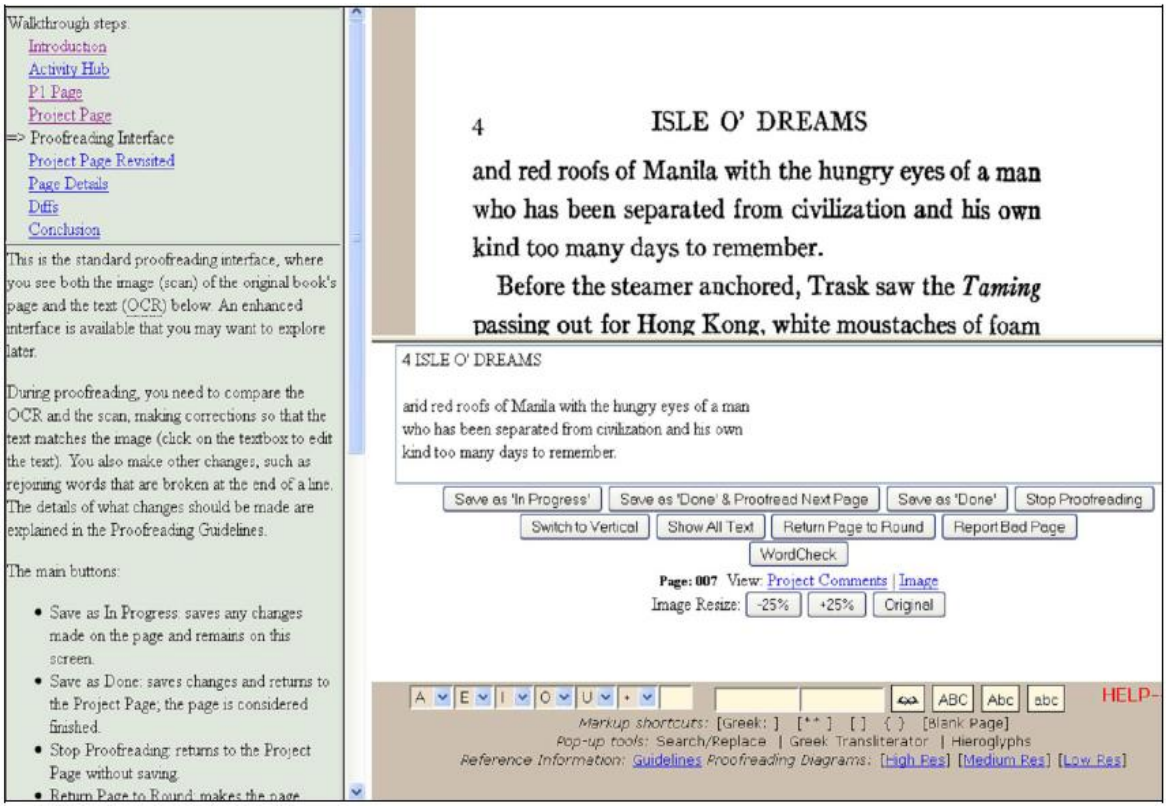
Pour un panorama plus complet, il resterait à évoquer les possibilités offertes par la société française Jouve, l'un des leaders mondiaux dans l'impression à la demande, les sociétés Lulu publishing (lulu.com), Lightning source, Virtual Bookworm, Wingspan press, iUniverse et Xlibris.

4- La correction participative de l'OCR et la transcription participative de manuscrits

4.1- La correction participative de l'OCR

4.1.1- Distributed Proofreaders (DP ou PGDP)

Porté par une organisation à but non lucratif, ce projet, fondé en 2000 par Charles Francks, soutient la correction participative par des bénévoles des livres numérisés de Internet Archive et tout particulièrement du projet Gutenberg auquel il a été officiellement rattaché en 2002. A l'origine, les internautes devaient identifier des livres correspondant à des critères notamment juridiques, les numériser et les océriser. Désormais, les documents numérisés proviennent de Internet Archive, mais il reste possible de scanner soi-même des livres. Ce projet est l'un des plus anciens projets de *crowdsourcing* dans le domaine de la numérisation. Son objectif est la mise à disposition de ebooks gratuits. En janvier 2004, Distributed Proofreaders Europe (http://dp.rastko.net, consulté le 23 juin 2016) a été créé. En octobre 2013, ce sous projet européen avait déjà produit 787 ebooks. En décembre 2007, Distributed Proofreaders Canada (http://www.pgdpcanada.net, consulté le 23 juin 2016) a été créé.



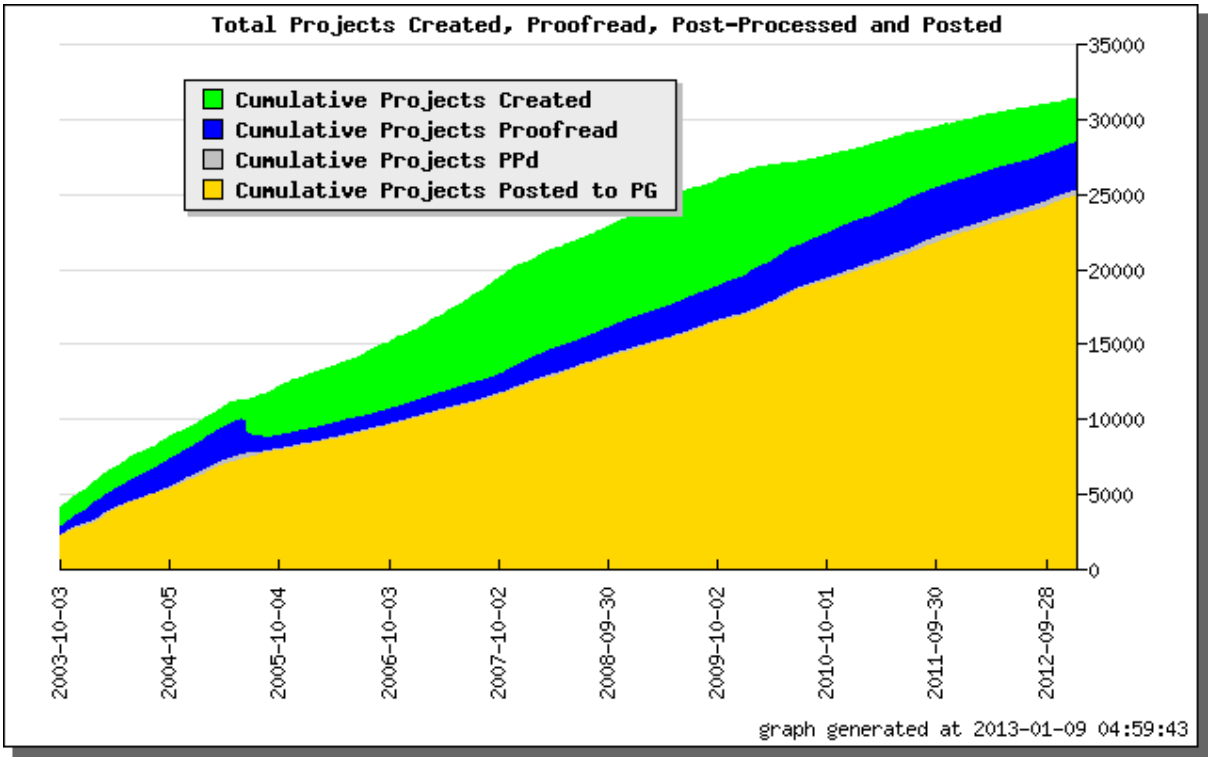
Interface de saisie de Distributed Proofreaders d'après (Kowalska, 2013)

Voici les données statistiques récoltées sur http://www.pgdp.net/c/stats/stats_central.php (consulté le 23 juin 2016) et dans la littérature d'après (Newby, 2003), (Holley, 2009), et (Kowalska, 2013) :

Date	Nombre de contributeurs	Nombre de ebooks
Après une semaine d'activités	Un millier de volontaires	Une quarantaine
En 2002	Plus de 6000 volontaires	Plus de 250 000 pages dont 110 000 pages en décembre 2002
En 2003		10 000 ebooks (75 ebooks par mois en moyenne)
En 2009	89 979 volontaires cumulés (dont 3000 actifs)	environ 16 000 ebooks (en moyenne 2000 ebooks par an)

Avril 2011		20 000 ebooks
4 mai 2012	112 825 comptes	
Le 9 janvier 2013	1890 membres actifs (sur les 30 derniers jours) dans le monde	24 893 textes ont été entièrement corrigés
En 2013		30 794 projets créés sur la plateforme dont 24 133 corrigés et 23 782 encore en cours de correction

Les statistiques de production sont en croissance depuis 2003 comme l'indique le diagramme suivant tiré du site :



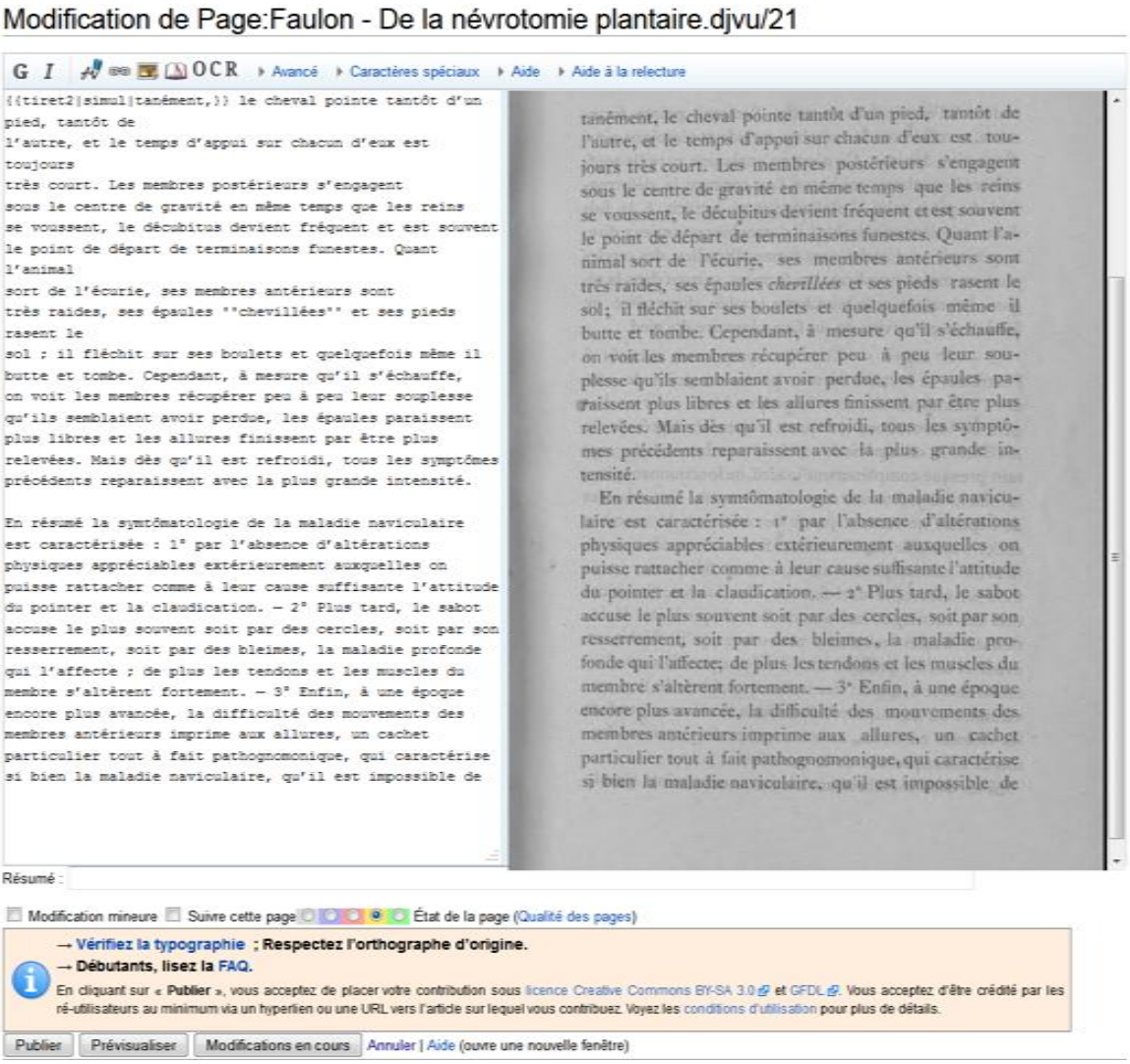
Statistiques de production tirées du site Distributed Proofreaders

Chaque page de texte est examinée par 2 correcteurs différents. Le premier correcteur voit l'image de la page numérisée et travaille à corriger son OCR brute.

Le second correcteur voit également l'image d'origine qui lui sert également de référence et vérifie l'OCR déjà corrigée par le premier correcteur.
L'encodage Text Encoding Initiative (TEI) était annoncé, en 2003, comme allant être de plus en plus développé (Newby, 2003)

4.1.2- Wikisource

Né le 24 novembre 2003, le projet Wikisource (initialement présenté sous le nom de projet "sourceberg") est un projet Wikimedia développé à partir de la plateforme MediaWiki et de son extension ProofreadPage.
Bien qu'il ne soit pas obligatoire d'être authentifié pour contribuer, le système de double vérification mis en place par Wikisource donnerait d'assez bons résultats en termes de qualité. A l'instar de Wikipedia, un historique permet également de restaurer des versions anciennes en cas de malveillance, et de surveiller les modifications de telle ou telle page.
Les documents obtenus sont librement téléchargeables au format ePub afin de pouvoir être lus sur des tablettes et des liseuses.



Page d'une thèse de médecine vétérinaire du début du 20è siècle conservée à l'Ecole Nationale Vétérinaire de Toulouse avec le texte OCR corrigé à gauche et sa photographie d'origine dans la partie de droite.

Au 4 janvier 2004, Wikisource comptait seulement 100 participants. En juillet 2004, on en trouvait plus de 500 pour plus de 1400 articles. Au 30 avril 2005, ils étaient 2667 participants pour 19000 articles. Au 27 novembre 2005, la version anglaise comptait 20 000 textes. Le 14 février 2008, elle comptait 100 000 et 250 000 en novembre 2011.

Le 11 février 2013, on trouve 122 942 pages de textes dont 95 275 (77,5 %) en mode image. 10 042 textes sont disponibles à la relecture.

Les pages peuvent avoir différents statuts :

- 624 517 pages non corrigées (et restant à corriger)
- 65 637 pages sans texte (pages de garde, pages blanches, plats, contre-plats, planches sans textes...)
- 351 395 pages corrigées (chacune par un ou plusieurs contributeurs, mais déclarées conformes aux fac-similés par au moins un contributeur)
- 112 396 pages validées (Une page doit être validée par 2 contributeurs pour avoir le statut de validé)
- 607 pages à problèmes (impossibles à corriger en raison d'un document original en mauvais état ou d'une prise de vue de mauvaise qualité, par exemples).

On compte à cette même date 32 110 utilisateurs enregistrés dont :

- 215 contributeurs actifs
- 21 administrateurs (des contributeurs qui peuvent protéger des pages contre le vandalisme, bloquer les vandales, réparer des erreurs, marquer comme vérifier une modification, gérer les filtres anti-abus)
- 3 bureaucrates (administrateurs nommés par la communauté Wikisource après candidature, afin de gérer les statuts des contributeurs. Ils peuvent ainsi donner le statut d'administrateur à un contributeur ou donner le statut de bureaucrate à un administrateur)
- 16 patrouilleurs (chargés de surveiller et de valider les pages modifiées contre le vandalisme)
- 16 robots (qui effectuent des tâches répétitives pour téléverser des pages, corriger les textes avec l'aide de fonctions de type rechercher/remplacer, créer ou corriger des catégories)

En 2008, une centaine de thèses de l'Ecole Nationale Vétérinaire de Toulouse ont été numérisées, dans le cadre d'un mécénat, avec l'aide de la fondation Wikimedia, puis diffusées sur Commons Wikimedia puis sur Wikisource et le texte océrisé

avait ainsi pu être, en grande partie, corrigé par des internautes de manière participative.

Le 7 avril 2010, la Bibliothèque nationale de France et Wikimédia France ont signé un partenariat. Rémi Mathis, qui était à la fois conservateur à la Bibliothèque nationale de France et membre du conseil d'administration de Wikimedia France depuis 2009 et qui en est le président depuis 2011 a probablement beaucoup œuvré pour qu'un tel partenariat voit le jour. La BnF a publié un communiqué de presse, le jour de cette signature, pour annoncer qu'elle avait livré 1416 livres sur Wikisource afin que leurs OCR soient corrigés de manière participative. Le corpus sélectionné comporte des textes avec des qualités d'OCR diverses, des livres de toutes époques et de tous formats, étant entendu qu'il s'agissait d'une simple expérimentation. Ainsi, on trouve 359 documents en mode image (sans OCR), et 1057 documents océrisés avec des qualités diverses. Au total, ce sont 573 310 pages dont la correction a été proposée aux internautes. Hélas, la participation n'a pas été aussi importante que prévue, seuls quelques wikipédiens habitués ayant apporté leurs contributions si bien que seulement 38 livres avaient pu être intégralement corrigés, malgré les divers communiqués de presse de la BnF et de l'association Wikimédia et malgré la tenue de rencontres Wikimedia GLAM en 2010 à l'Assemblée Nationale au cours desquelles, le Directeur Général adjoint de la Bibliothèque nationale de France intervint afin de présenter la collaboration entre son institution et Wikisource. Par ailleurs, la réintégration des données dans Gallica semble problématique. La synchronisation entre les index de recherche, le fichier ALTO de géoréférencement des mots dans le texte reste à faire et semble assez complexe à concrétiser. Comme la BnF en a fait l'amère expérience, le principal inconvénient de Wikisource réside probablement dans le fait qu'il ne contient aucune information de structure (de type ALTO) et ne permet pas de fonctionnalités de surlignage, ni de réintégration facile des données au sein de bibliothèques numériques participantes. Au delà de la communication institutionnelle, cette expérience rappelle que la réintégration des données produites par les internautes doit être une question centrale de tout projet de *crowdsourcing*.

En France, les Archives Nationales, les Archives départementales du Cantal et les Archives départementales des Alpes Maritimes (depuis janvier 2012) collaborent également avec Wikisource.

4.1.3- California Digital Newspaper Collection (CDNC)

Le projet California Digital Newspaper Collection (CDNC) a été lancé fin 2011 par l'University of California, Riverside. La correction de l'OCR de 400 00 pages de journaux locaux publiés entre la fin du 19è siècle et le début du 20è siècle y est proposée. En mai 2014, la collection comportait 61 412 numéros, 545 955 pages, et 6 364 529 articles (Zarndt, 2014). Les solutions logicielles de correction participative de l'OCR ("User Text Correction") ont été développées par DL Consulting à partir du logiciel Veridian 3.0. L'authentification des internautes est obligatoire s'ils souhaitent contribuer.

Voici les données statistiques sur le projet recueillie dans la littérature :

Date	**Nombre de contributeurs**	**Nombre de lignes corrigées**
9 premières semaines	96 bénévoles	50 000 lignes
1 an après le lancement	309 bénévoles	400 000 lignes
En 2012		578 000 lignes
Début 2013	450 bénévoles	750 000 lignes
Juin 2013	1340 internautes enregistrés dont 599 contributeurs actifs	1 160 465 lignes
Juin 2014	2246 internautes enregistrés dont 1266 contributeurs actifs	2 656 497 lignes, soit seulement 2 % de la collection totale

Sur la base de coûts moyens de correction de l'OCR auprès de prestataires de 0,50 $ pour 1000 caractères et d'une moyenne de 40 caractères par ligne, Brian Geiger évaluait, en 2012, à 11 560 $ le gain, ou plutôt l'argent non dépensé, pour

California Digital Newspaper Collection (578 000 lignes corrigées) (Geiger, 2012). En juin 2014, nous pourrions évaluer ce coût à 53 130 $ en reprenant ce mode de calcul : 2 656 497 / (1000 / 40) x 0,5. Ce calcul a été confirmé par (Zarndt, 2014)

A l'instar des projets précédemment évoqué, on observe que la plupart du travail effectué est l'œuvre de quelques contributeurs très actifs :

User rank	Lines corrected Jun 2014	Lines corrected Oct 2012
1	717,855	242,965
2	271,972	87,515
3	120,220	31,318
4	113,787	24,144
5	109,999	23,184
6	99,999	19,240
7	94,742	18,898
8	65,637	16,875
9	63,786	11,784
10	59,724	9,762

Top 10 des meilleurs contributeurs en juin 2014, comparé à octobre 2012 d'après (Zarndt, 2014)

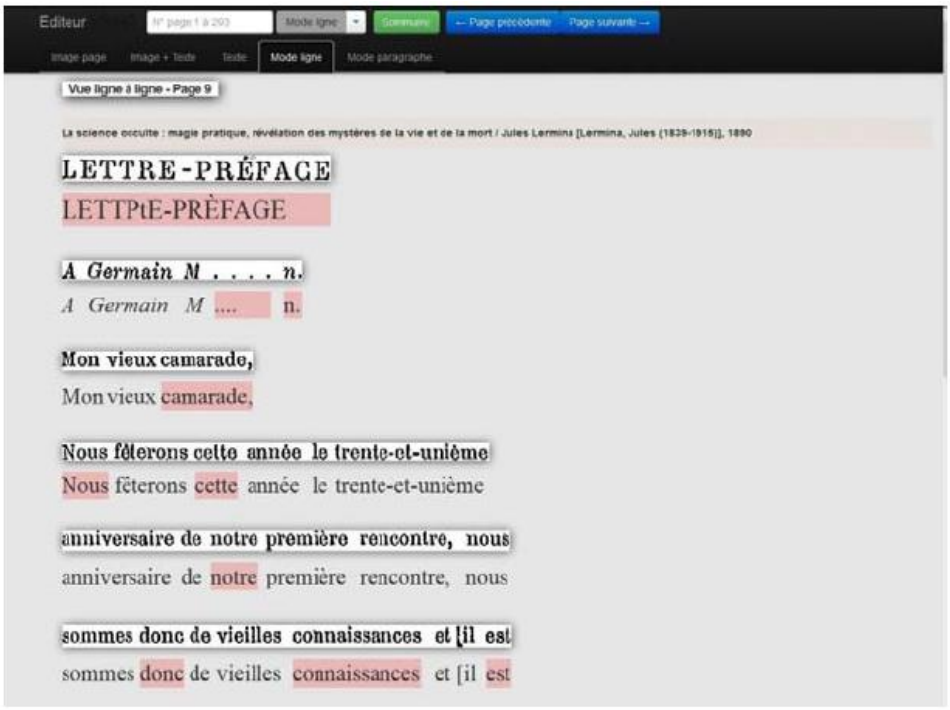
4.1.4- La Bibliothèque nationale de France et la plateforme Correct (projet FUI12 Ozalid)

En partenariat avec Orangelab, la Bibliothèque nationale de France a lancé un projet de recherche et développement pour l'expérimentation du *crowdsourcing* dans le cadre d'une plateforme baptisée "Correct". L'objectif pour la société Orange serait de commercialiser un logiciel puis de l'ouvrir à d'autres bibliothèques.

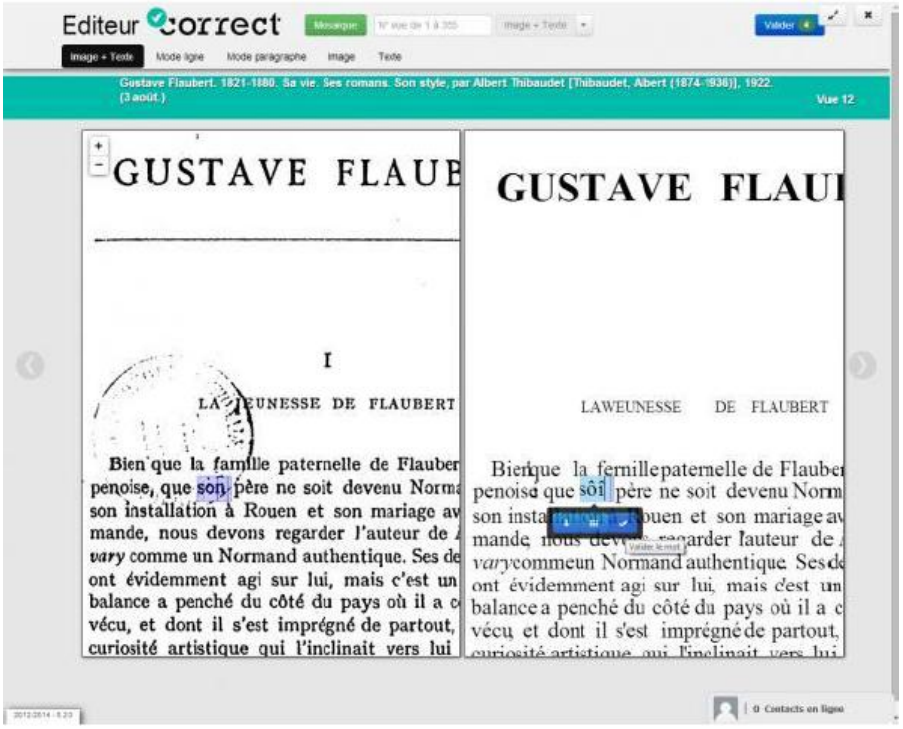
Les autres partenaires sont Jamespot, Urbilog, I2S, ISEP, INSA Lyon, Université Lyon 1 – LIRIS et l'Université Paris 8. Le programme d'origine prévoyait les étapes suivantes :

- En 2012-2013, une interface devrait permettre la correction participative de textes numérisés.
- En 2013-2014, le développement de fonctionnalités autour de la mise en forme du texte et des tables de matières
- En 2014-2015, le développement de fonctionnalités d'enrichissement éditorial (indexation, vocalisation, annotation...)

Finalement, la plateforme de correction participative de l'OCR n'a été ouverte que le 24 novembre 2014.



Copie d'écran de la plateforme Correct (mode ligne à ligne)



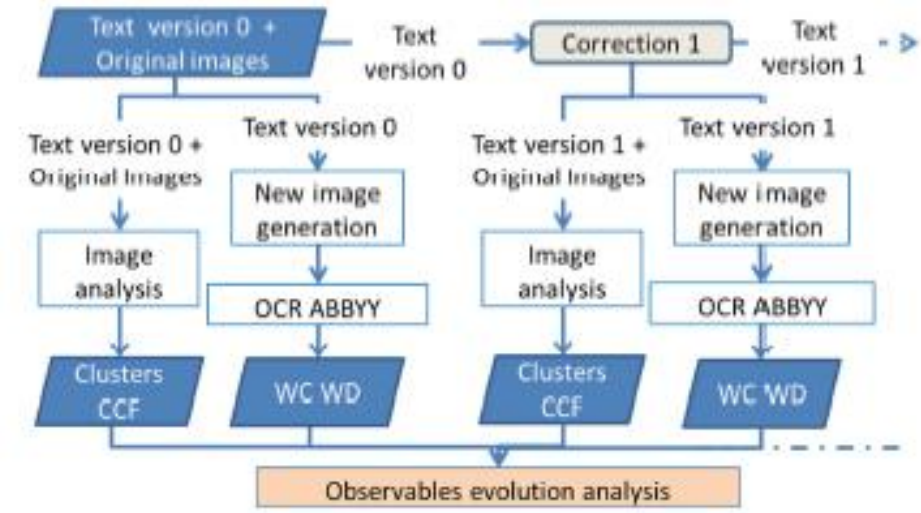
Copie d'écran de la plateforme Correct (mode vis-à-vis : texte + image)

Les corrections des internautes seraient confrontées par le système afin de pouvoir valider ou identifier les contradictions. D'après (Lagarrigue, 2014), la qualité du texte est jugée suffisante lorsque la majorité des internautes n'a plus de nouvelles corrections à proposer.

Deux méthodes ont été expérimentées :

- Corrections simultanées par plusieurs correcteurs indépendamment puis fusion des corrections. Puis correction de la nouvelle version obtenue de la même manière.
- Corrections successives par chaque correcteur, un par un.

  En cas de conflits de corrections, un système de vote a été mis en place.



Protocole d'évaluation de la qualité d'après (Lagarrigue, 2014)

D'après un article publié sur Actualitté le 1er décembre 2014 ("la plateforme correct de la BnF"), il n'y avait à cette date que 214 inscrits et 6 groupes thématiques de corrections. Au 23 janvier 2015, il y avait 425 inscrits (soit 6 nouveaux inscrits par jour) ayant corrigé 782 199 mots (soit 3 631 pages de textes) (Bureau Van Dijk, 2015). Selon cette même étude des consultants du Bureau Van Dijk, au 25 mars 2015, il y avait 528 inscrits ayant corrigé 1 322 674 mots (soit 5 746 pages de textes). De manière assez similaire aux autres projets, une minorité de contributeurs ont effectué la majeure partie des corrections. Ainsi, les 10 premiers contributeurs ont effectué près de la moitié des corrections et les 37 correcteurs les plus actifs (soit 8 % des contributeurs) ont corrigé 63 % du total. Les inscriptions sont fortement corrélées avec la communication sur les réseaux sociaux (comptes Twitter, comptes Facebook, blogs de la BnF). Malgré cela, 47 % des inscrits auraient abandonné leur participation sous 48 heures.

Une enquête approfondie menée par le Bureau van Dijk et portant sur 159 questionnaires renseignés par 100 utilisateurs de Correct et 59 non utilisateurs, montre notamment que 42 % des utilisateurs de la plateforme sont des professionnels des bibliothèques et de la culture, que 82 % des connexions ont lieu en semaine, et que 75 % choisissent des documents à corriger selon leurs centres d'intérêts. L'aspect réseau social semble avoir insuffisamment été

approprié (à cause du nombre insuffisant de contributeurs et car l'activité de correction qui peut être considérée comme solitaire). La note moyenne donnée au projet par les répondants est de 6,5 / 10.

4.1.5- Franscriptor

Franscriptor est une initiative française visant à transcrire, corriger, indexer et moderniser des textes anciens. Le travail peut être bénévole ou rémunéré. C'est celui qui dépose un document à transcrire qui le décide et, le cas échéant, fixe le prix de la transcription. Une double saisie peut être demandée pour un coût double.

4.2- La transcription participative de manuscrits

4.2.1- What's on the menu ? (WOTM)

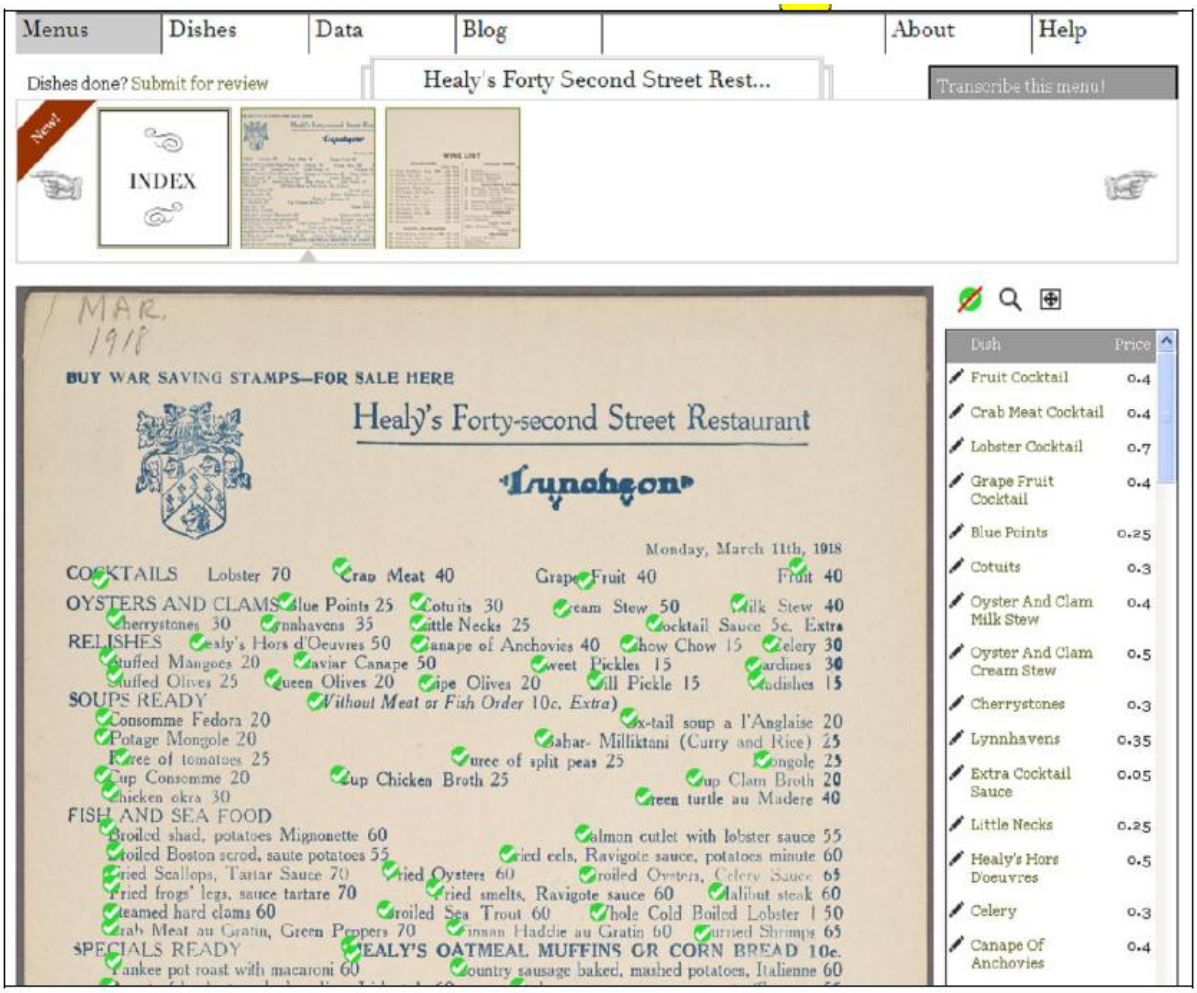
Développé par la New York Public Library et lancé en avril 2011, "What's on the menu ?" est un projet qui demande aux internautes de transcrire 45 000 menus de restaurants dont les plus anciens remontent à 1840.

L'authentification n'est pas obligatoire pour participer. Les transcripteurs volontaires sont d'ailleurs invités à ne pas se soucier des accents. Dans tous les cas, les transcriptions sont ensuite validées par des experts. Les données finales sont librement réutilisables via des exports de type tableur et une API.

Des récompenses ont également été proposées aux contributeurs sous la forme d'événements et même de repas (Spindler, 2014).

Date	**Nombre de transcriptions**	**Nombre de visiteurs**
En 3 mois (juin 2011)	3000 menus	
En 4,5 mois (Août 2011)	10 000 menus	58 000 visiteurs uniques
En 10 mois		3 millions de

		pages vues
En 17 mois (Septembre 2012)	1 083 509 transcriptions de 15 630 menus (environ 800 000 plats)	
En mars 2015	17 541 menus (1 328 904 plats)	



Capture d'écran de What's on the Menu ? d'après (Kowalska, 2013)

4.2.2- Ancient Lives

Lancé en juillet 2011 par l'Université d'Oxford, et porté par la Citizen Science Alliance, ce projet parallèle à Old Weather, et à Galaxy Zoo est intégré dans le réseau Zooniverse. Il consiste en une transcription de papyrus égyptiens par

microtâches (reconnaissance de caractères de grec ancien). Les contributeurs peuvent interagir, via les forums et les espaces de commentaires pour chaque document. L'authentification n'est obligatoire que pour agir dans la partie sociale du site, pas pour contribuer.

L'affichage simultané du texte océrisé et de l'image du texte se fait par superposition des caractères corrigés et de l'image de ces caractères, contrairement à la plupart des projets de correction participative de l'OCR (affichage à la page).

Une campagne de communication dans la presse, sur la BBC et via les réseaux sociaux a permis de mieux faire connaître le projet auprès de public.

De juillet 2011 à décembre 2012, au delà de 1,5 million de transcriptions ont déjà été obtenues.

Capture d'écran de l'interface de transcription de Ancient Lives, d'après (Carletti, 2013)

4.2.3- ArcHIVE

Proposé par les archives nationales d'Australie, depuis avril 2011, ArcHIVE propose la transcription participative de catalogues d'archives, sans authentification obligatoire des internautes. Le projet a développé, en particulier, des systèmes de valorisation afin de susciter la contribution des internautes, par l'affichage du classement des meilleurs contributeurs, par une rétribution symbolique sous la forme de fac-similés (*print on demand*), d'objets, de marque-pages, de posters que l'on peut échanger lorsqu'un certain nombre de points ont été gagnés.

Les documents eux-mêmes sont classés selon 3 niveaux de difficultés. Un forum permet aux contributeurs d'échanger et de s'entraider.

300 documents ont été intégralement complétés en moins de 15 jours après le lancement du site. Sur les 3500 documents proposés, 1500 ont été transcrits début 2013.

4.2.4- What's the score (WTS)

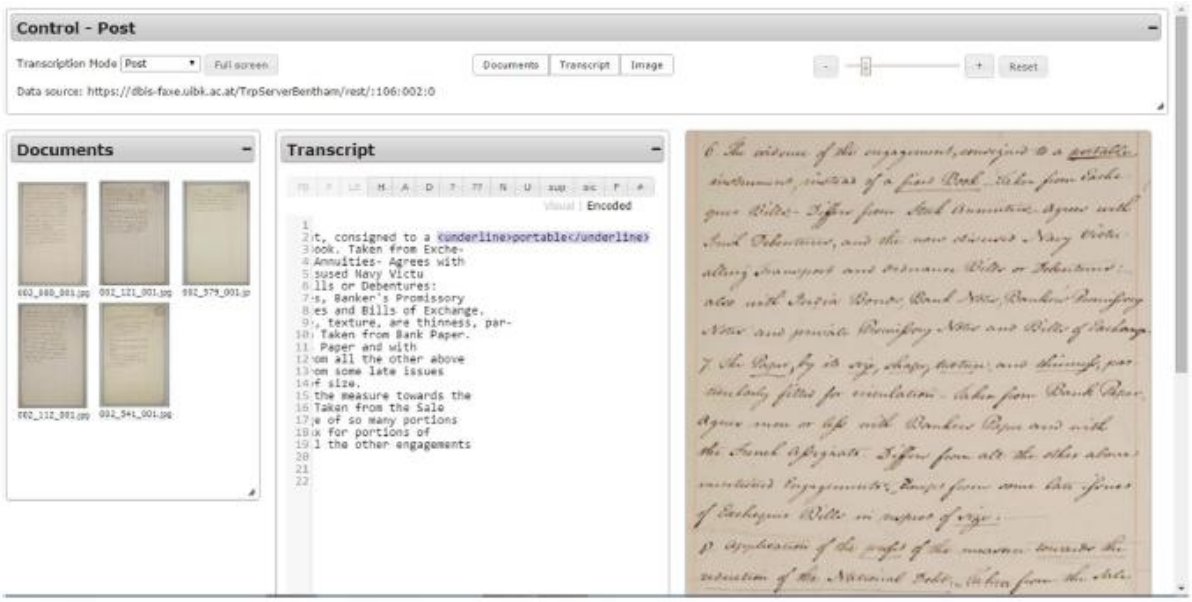
Développé en mai 2012 pour les collections musicales de la Bodleian Library de l'Université d'Oxford, le projet What's the score est un projet de correction et de description participatives de partitions de piano numérisées. Sponsorisé par Google, le projet utilise des logiciels développé dans le cadre du projet Zooniverse et se distingue par l'originalité de son ergonomie.

Les internautes sont, en effet, invités à définir les zones de l'image qu'ils vont décrire ou transcrire à la manière d'une segmentation utilisée par les logiciels d'OCR.

Capture d'écran de What's the score (d'après McKinley, 2012)

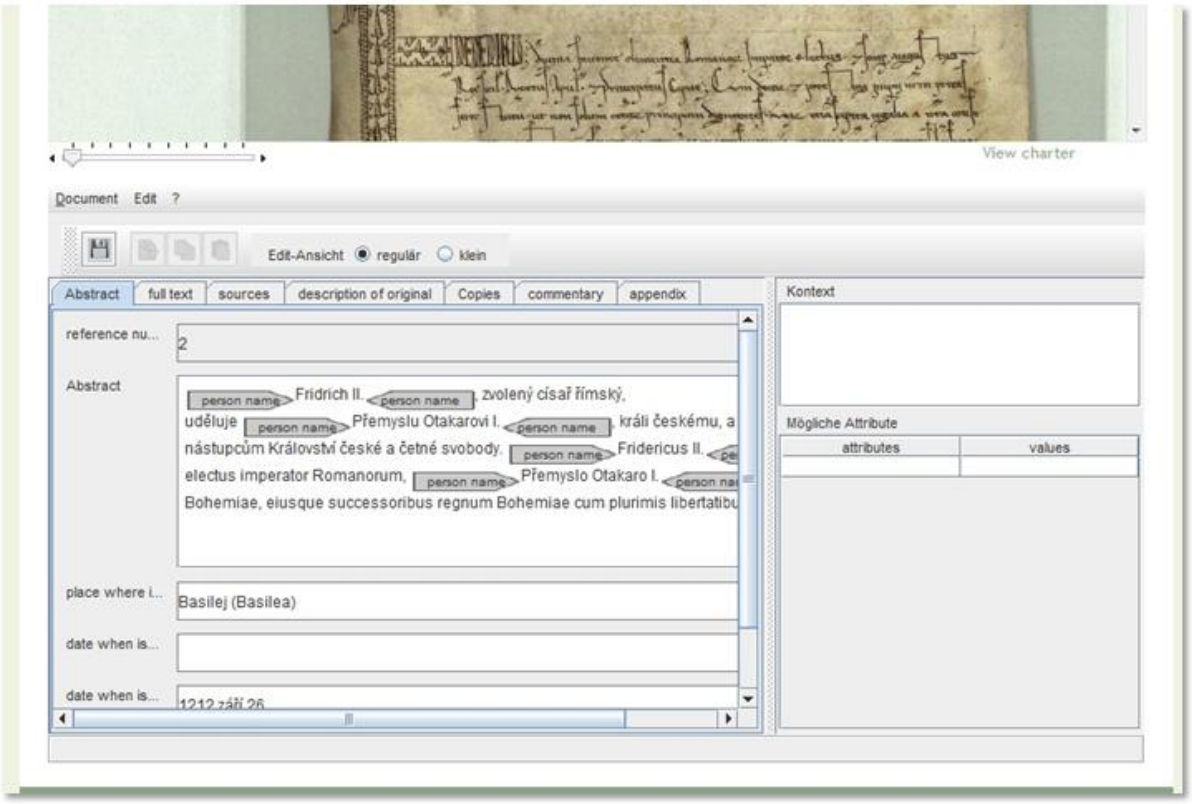
4.2.5- Transkribus

Transkribus est une plate-forme ouverte et gratuite pour la transcription des documents manuscrits et imprimés proposant des fonctionnalités de *crowdsourcing* via l'outil TSX développé en PERL et HTML 5.0. Le projet associe l'University College London (à l'origine du projet Transcribe Bentham), l'University of Innsbruck (à l'origine du projet européen de *crowdfunding* Ebooks on Demand), l'University of London Computing Centre, le National Research Centre Athens et le Dutch Institute for Lexicography dans le cadre du projet européen tranScriptorium.



Capture d'écran de l'outil TSX d'après Günter Mühlberger, Sebastian Colutto, Philip Kahle (21 mai 2015). TRANSKRIBUS : An Open Platform for the Transcription and Recognition of Handwritten and Printed Documents

Afin de dresser un panorama plus complet des projets de transcription collaborative, nous aurions également pu évoquer le projet Do it yourself History (DIY History), développé par l'Université de l'Iowa au printemps 2011, qui a pour originalité d'investir fortement dans le *community management* et d'autoriser des contributions sans authentification mais soumises à validation par des experts. Le projet Monasterium Collaborative Archive (MOM-CA), porté par l'International Centre for Archival Research (ICARUS), et lancé en 2002, s'adresse, quant à lui aux médiévistes à qui il propose, après authentification obligatoire, la transcription et la valorisation éditoriale de 250 000 manuscrits médiévaux provenant d'une cinquantaine d'institutions européennes. Environ 150 historiens et érudits sont déjà inscrits sur le site. 14 experts interviennent dans la communauté des contributeurs, notamment pour valider les contributions.



Capture d'écran du site Monasterium - Collaborative Archive (MOM-CA)

Le site Citizen Archivist Dashboard, développé par les archives nationales des USA, permet, quant à lui, aux internautes de taguer, transcrire des archives de manière collaborative, mais aussi d'y ajouter des articles de présentation et de commentaires et même d'uploader des documents scannés. Enfin, les archives nationales des USA avec le National Archives Transcription Pilot Project proposent différents niveaux de difficulté de correction (débutant, intermédiaire, difficile). Lié à Citizen Archivist Dashboard ce projet permet aux internautes d'ajouter également des archives et d'écrire des présentations de documents. Nous pourrions également évoquer les projets de transcriptions néo-zélandais Wanganui Library, le projet Field Notes of Laurence M. Klauber du San Diego Natural History Museum, le projet Notes from Nature, le projet Transcribe Bushman et le Smithsonian Digital Volunteers Transcription Center. En France, le consortium des archives des ethnologues a lancé le projet Transcrire afin de permettre au plus

grand nombre de participer à la transcription des carnets de terrain des ethnologues.

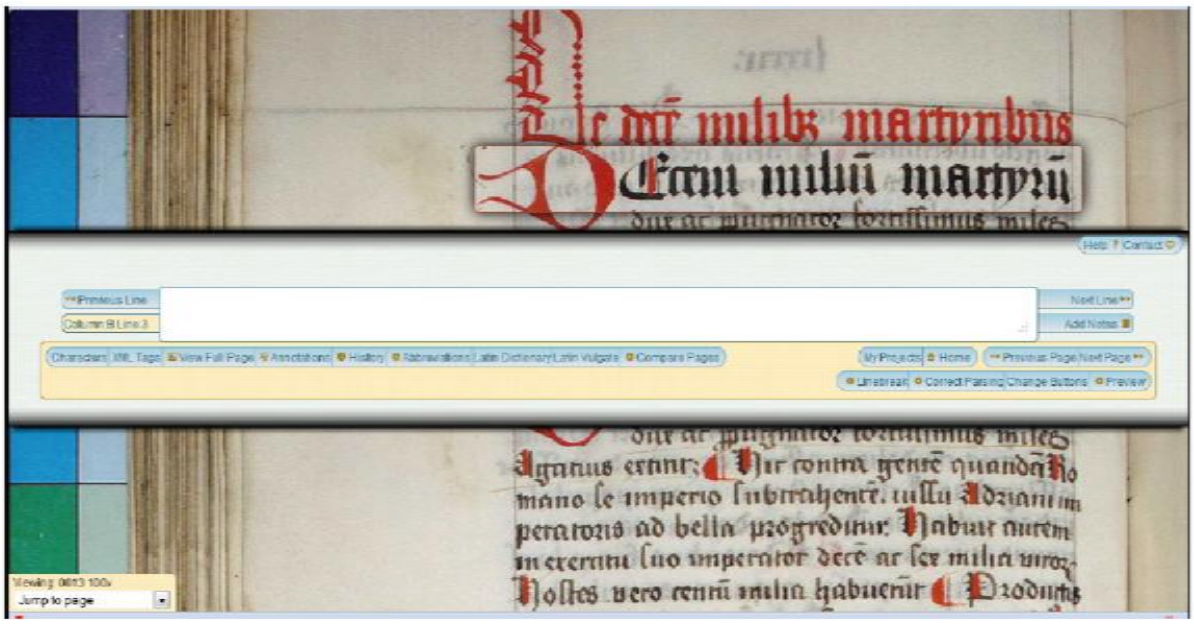
4.3- Les logiciels de correction / transcription participatives

4.3.1- T-Pen

T-Pen, un outil libre d'aide à la transcription des manuscrits offrant des possibilités de transcriptions collaboratives. Le projet est développé par le Center for Digital Theology at Saint Louis University (SLU) et financé par la Andrew W. Mellon Foundation et le National Endowment for the Humanities (NEH).

Il est principalement utilisé pour des projets de transcriptions de manuscrits médiévaux ou de manuscrits de la Renaissance.

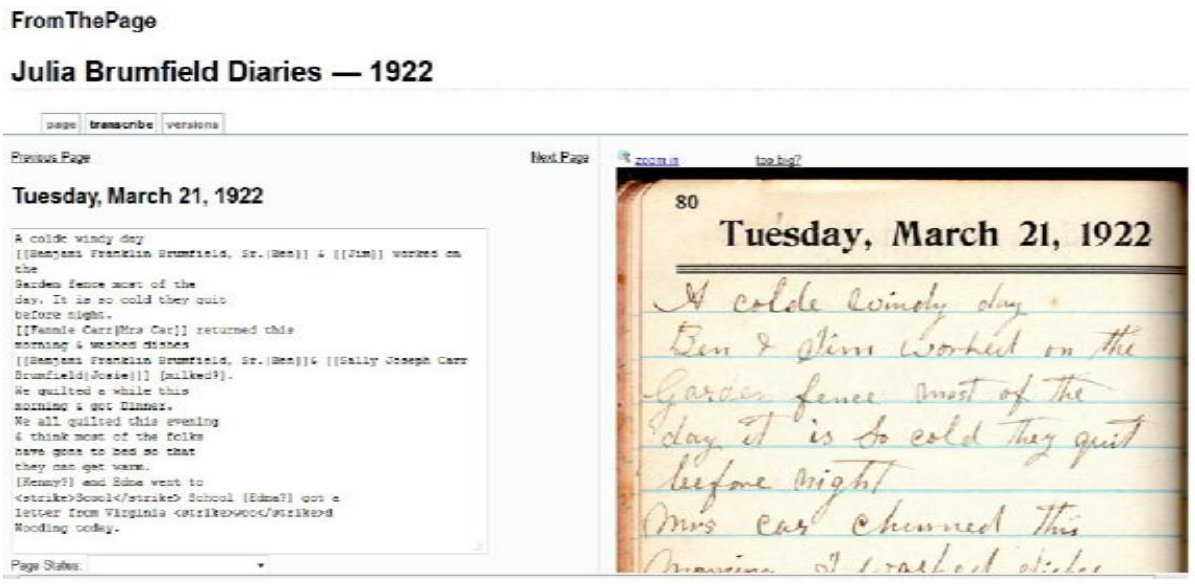
Au lieu de demander aux transcripteurs la connaissance des normes TEI, de simples boutons sont utilisés.



Interface de transcription de T-Pen d'après (Brokfeld, 2012)

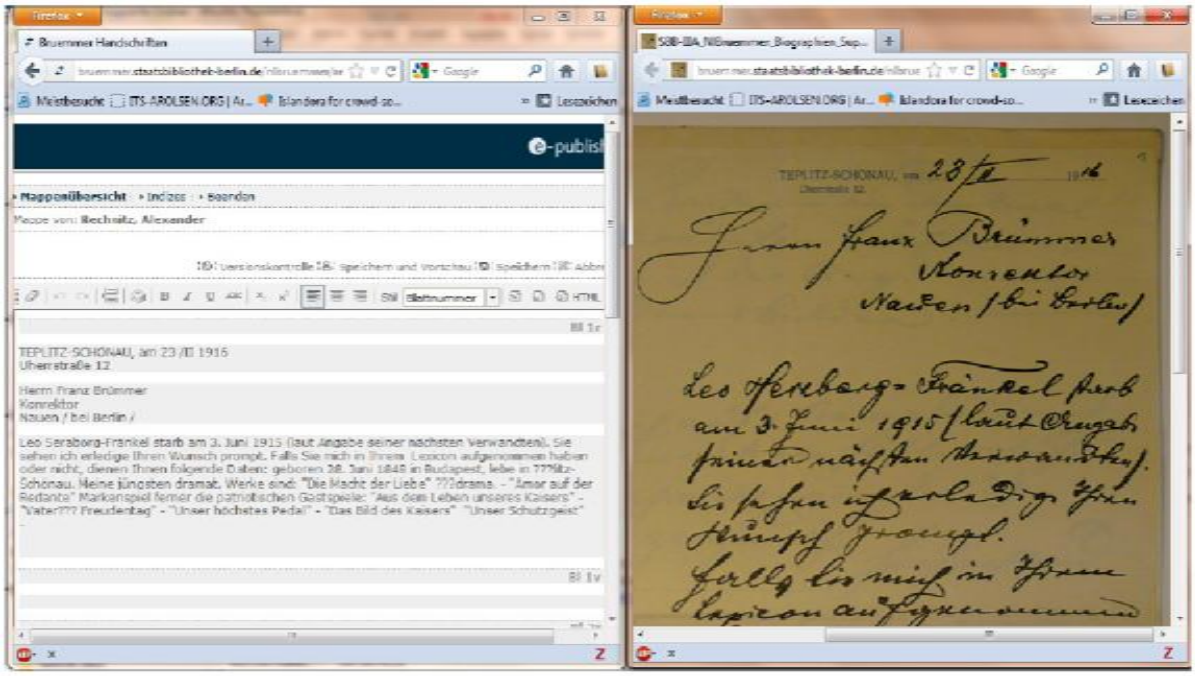
4.3.2- FromThePage

FromThePage est un logiciel de transcription participative de manuscrits développé par Ben Brumfield, un développeur informatique.



Interface de transcription de FromThePage d'après (Brokfeld, 2012)

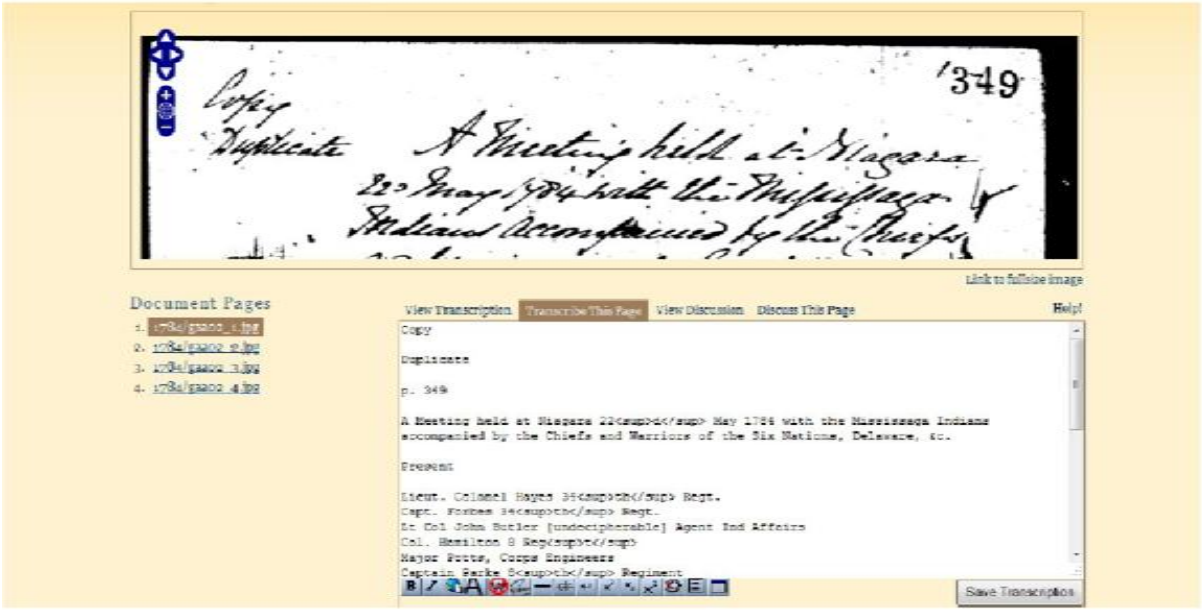
4.3.3- Refine!



Interface de transcription de Refine! d'après (Brokfeld, 2012)

4.3.4- Scripto

Développé par le Center for History and New Media de la George Mason University est un logiciel Open Source utilisé, en particulier pour la transcription de l'American War Department.



Interface de transcription de Scripto d'après (Brokfeld, 2012)

A la lumière du mémoire (Brokfeld, 2012) on peut proposer le tableau comparatif, en fonction des domaines, des 5 logiciels :

	Bibliothèque	Archives	Généalogie	Science	Muséum
Wikisource	++	++	-		++
Bentham Transcription Desk	++	++			++
T-PEN				++	
FromThePage	++		++	-	
Refine!	++	++	-		
Scripto	+	+	-	-	++

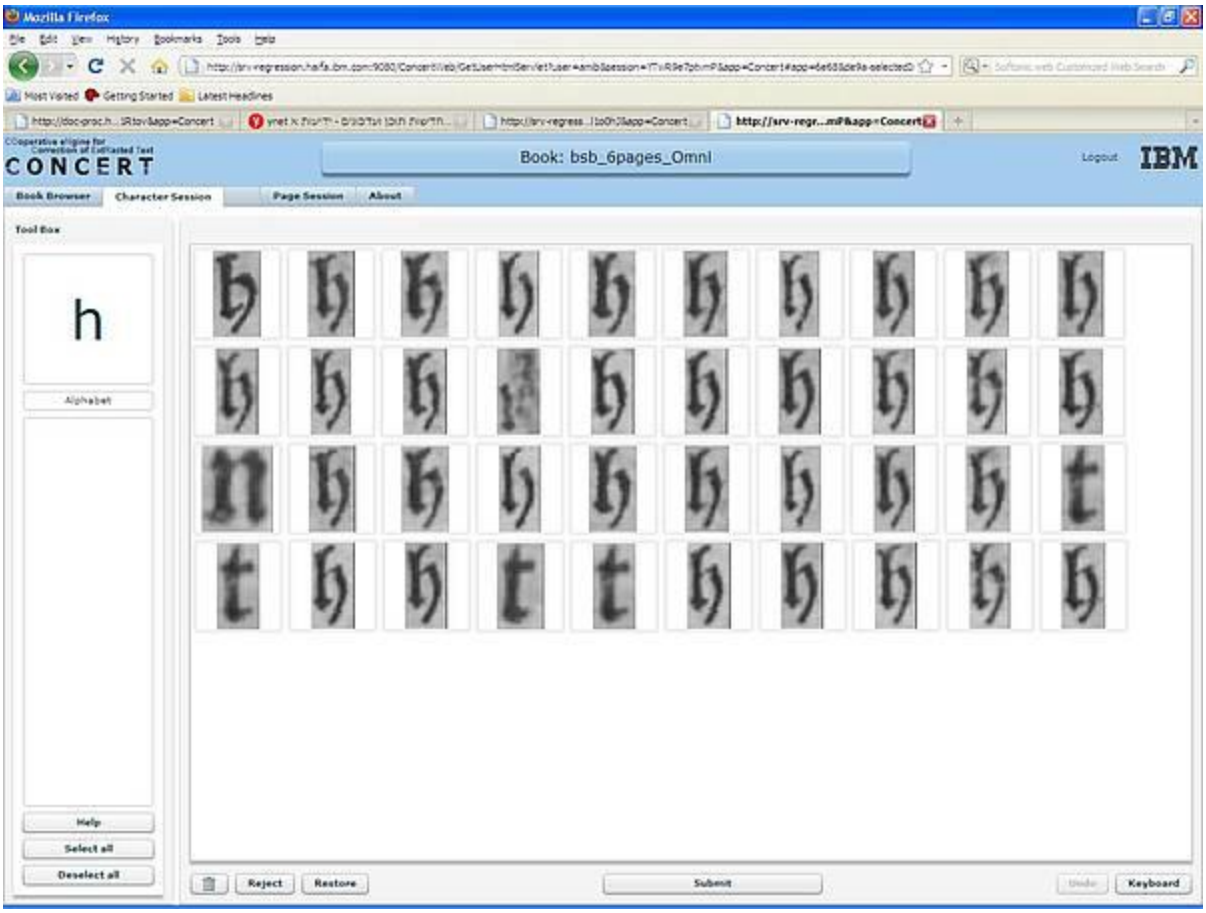
4.4- La *gamification*, la correction de l’OCR en jouant

4.4.1- COoperative eNgine for Correction of ExtRacted Text (CONCERT)

CONCERT a été développé depuis 2009 dans le cadre du projet européen IMPACT par IBM Israël en partenariat avec Haifa, Koninklijke Bibliotheek, The British Library, Osterreichische Nationalbibliothek, Universitat Innsbruck, Deutsche Nationalbibliothek, Bayerische Staatsbibliothek, Staats- und Universitatsbibliothek Gottingen, ABBYY Production, Instituut voor Nederlandse Lexicologie, National Centre for Scientific Research "Demokritos" Centrum fur Informations- und Sprachverarbeitung, University of Munich, University of Bath, University of Salford, Bibliotheque Nationale de France, Biblioteca Nacional de Espana and Poznan, Supercomputing and Networking Center in Poland”.

Le projet CONCERT offre des fonctionnalités originales et innovantes pour la correction participative de l’OCR. En effet contrairement à la plupart des autres projets qui affichent sur une même page l’image du texte et le texte océrisé, avec l’interface CONCERT, celle-ci se fait ligne à ligne. La démarche est donc assez différente des autres modèles.

Après une authentification obligatoire, les internautes se voient, en effet, proposer un échantillon de caractères à identifier (étape du “tapis” ou “carpet” en anglais). Un “tapis” se compose d'un écran sur lequel sont affichées toutes les occurrences du même caractère d'imprimerie numérisé sur toutes les pages d'un même document et en particulier celles que le logiciel de reconnaissance de caractères suspecte d'être mal identifiées. De cette manière, l'humain chargé de corriger l'OCR du document numérisé pourra identifier bien plus rapidement les erreurs générées par le logiciel. Dans la capture d'écran ci-dessous, l’œil humain peut ainsi rapidement identifier la présence d'un caractère « n », de 4 caractères « t » et d'un caractère plus difficile à déterminer. Le reste des caractères affichés est bien constitué de caractère « h ». Avec l'aide de l'humain et par cette méthode d'apprentissage, le logiciel va pouvoir améliorer la qualité de son OCR.



Copie d'écran du « tapis » tel qu'il apparaît avec le logiciel CONCERT d'après https://www.digitisation.eu (consulté le 23 juin 2016, mais la ressource n'était plus en ligne)

Les caractères qui ne sont pas des caractères « h » sont rapidement reconnus par l'œil humain

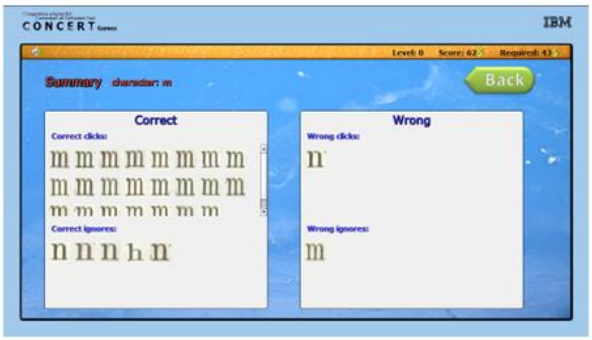
Dans une deuxième étape, le système demandera ensuite aux internautes de corriger non plus des lettres mais des mots considérés comme suspects. Au fil des corrections, de nouveaux mots sont susceptibles d'être ajoutés dans le dictionnaire. Enfin, une troisième et dernière étape consiste à proposer les mots dans leurs contextes (Neudecker, 2010). Ces 3 étapes permettent de diversifier les activités et de les adapter au niveau des contributeurs. Ainsi, les tâches leurs sont attribuées selon leur expérience et leurs compétences.

Au départ de chaque nouvelle session, des mots tests et des erreurs aléatoires sur “tapis” sont soumis aux contributeurs novices afin de vérifier qu’ils ne sont pas malveillants et mesurer leur taux de réussite (Karnin, 2010). Pour leur part, les meilleurs contributeurs peuvent même être rétribués.

Les caractères issus de l’OCR sont divisés en 3 groupes : les caractères sûrs qui sont considérés comme valides, les caractères non sûrs qui seront corrigés manuellement et les caractères moyennement sûrs qui seront corrigés via l’étape du tapis. (Conteh, 2009)

Ces technologies permettraient d’améliorer considérablement la productivité de correction de l’OCR. Un petit livre qui prenait auparavant 4 heures à être corrigé, n’en prendrait ainsi plus qu’une.

Les équipes travaillant sur le projet CONCERT ont également développé des jeux sur réseau et sur smartphone Android :



Captures d'écrans des jeux développés par CONCERT

4.4.2- TypeAttack

Sur le modèle de DigitalKoot, TypeAttack, contraction de "typing and attacking" est un jeu développé sous Facebook et qui invite à corriger l'OCR de documents numérisés par des internautes en compétition pour saisir le plus de mots avec le plus de précision possible et en un temps limité (Jovian, 2011). Il est le résultat de la collaboration entre la National University of Singapore et la Bibliothèque Nationale de Singapour. Le jeu autorise, de surcroît, une compétition, sur Facebook, entre relations du même réseau social.

Les joueurs ont accès à des indicateurs de leur propre vitesse de saisie et de celle de leurs compétiteurs afin de provoquer une émulation entre eux. Les saisies sont confrontées afin d'obtenir un texte corrigé de bonne qualité. Un classement des meilleurs joueurs est proposé. On peut proposer à un ou plusieurs membres de son réseau social Facebook d'entrer en compétition. Les scores peuvent être publiés sur le journal Facebook du joueur.

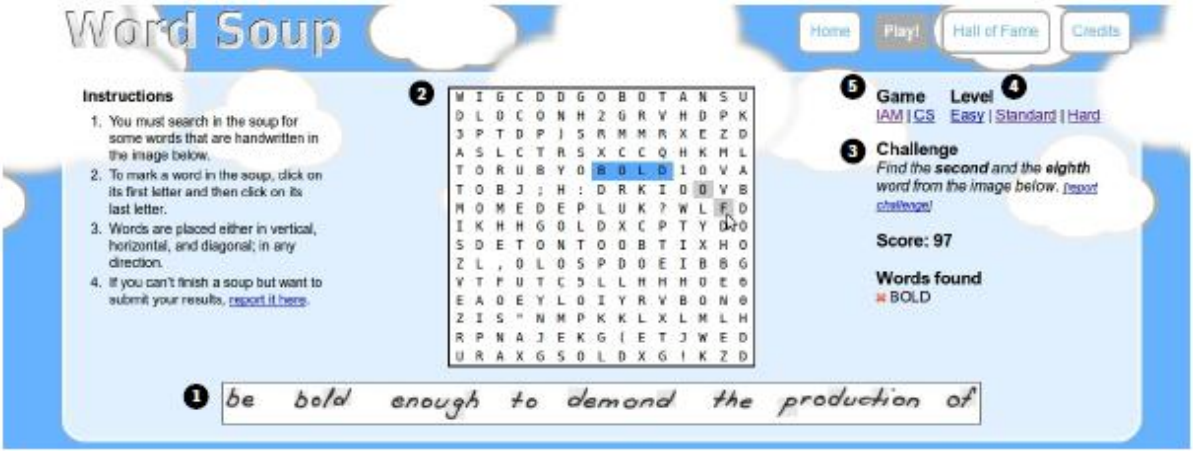
D’après (Jovian, 2011), après 5 semaines, 289 joueurs recrutés à partir de seulement 20 comptes Facebook avaient contribué. D’après ce même auteur, ce site a permis de traiter 505 extraits d’environ 5 lignes chacun en 5 semaines avec un taux de reconnaissance de 99,1 %. Il a rassemblé 3980 joueurs qui s’y sont consacré en moyenne 10 minutes avec un maximum de 5 heures.

Capture d’écran de TypeAttack

Capture d'écran de "Type-and-Earn" d'après (Jovian, 2011)

4.4.3- Word Soup Game

Inspiré par ReCaptcha, le Word Soup Game est un jeu dans lequel il faut identifier des mots au sein d'une grille de lettres désordonnées appelée soupe de mots. Les mots sont issus d'une OCR à transcrire.



Capture d'écran de Word Soup d'après (Alabau, 2012)

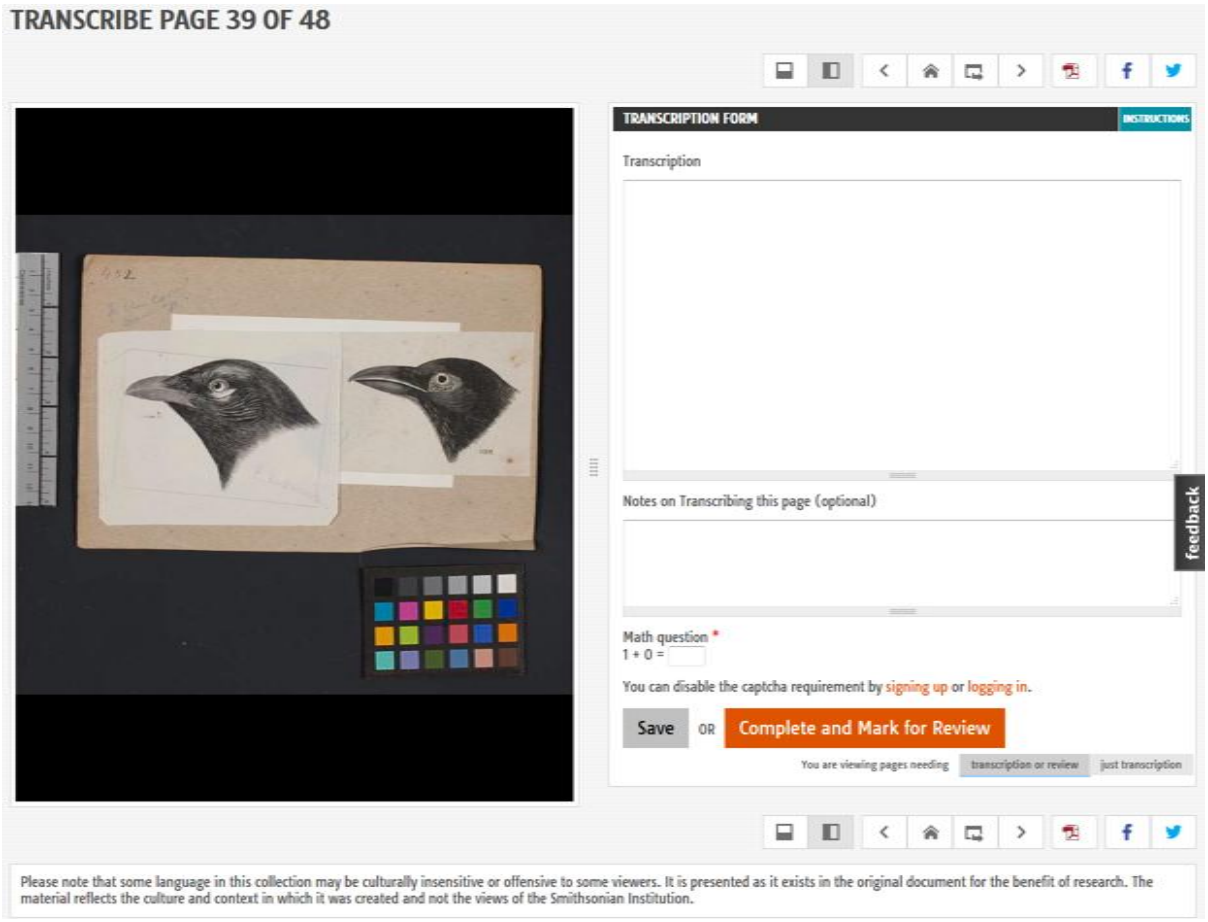
4.4.4- Un jeu pour corriger l'OCR en arabe

Le jeu développé à titre expérimental, consiste à montrer un mot au joueur pendant une courte durée, puis à demander au joueur de s'en souvenir et de le saisir en un temps limité et déterminant son score. Le joueur peut revoir le mot mais perdra des points à chaque fois.

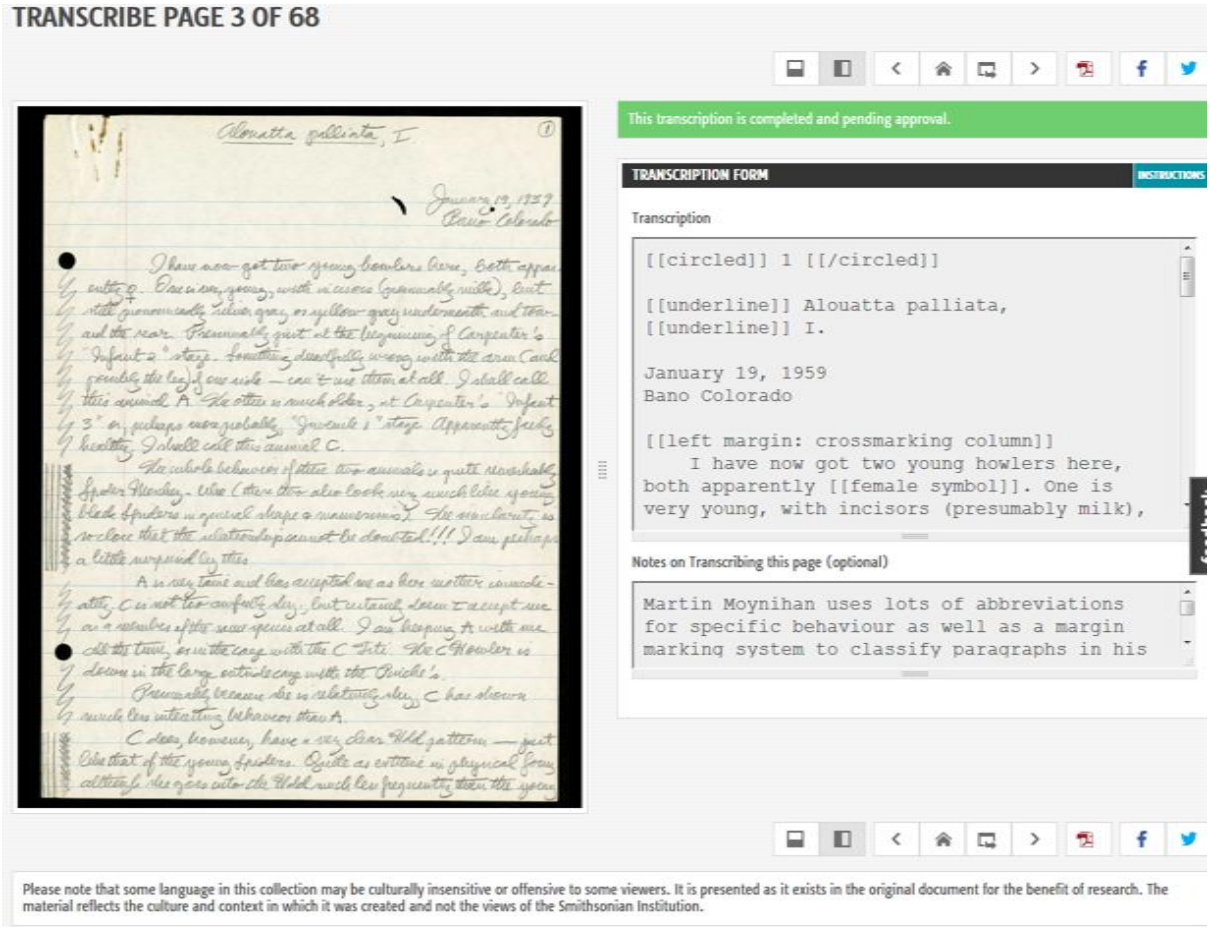
Interface du jeu d'après (AlRouqi, H., 2014)

4.4.5- Biodiversity Heritage Library (BHL) : Smorball et Beanstalk

La Smithsonian Institution a développé une plateforme de transcriptions de manuscrits et d'indexation d'images : https://transcription.si.edu (consulté le 23 juin 2016)



Capture d'écran d'une transcription d'image sur https://transcription.si.edu (consulté le 23 juin 2016)



Capture d'écran d'une transcription de manuscrit sur

https://transcription.si.edu (consulté le 23 juin 2016)

Mais, au delà de ce *crowdsourcing* classique, des jeux de transcription de textes océrisés sont également proposés. Ils ont été développés par Tiltfactor, un centre de recherche sur les *serious games* fondé par Mary Flanagan qui est notamment à l'origine du projet Metadata Games déjà évoqué.

Le premier jeu Beanstalk consiste à saisir un extrait de texte numérisé afin de faire grandir un haricot.

Capture d'écran de http://beanstalkgame.org (consulté le 23 juin 2016)

Le deuxième jeu, Smorball, est plus compétitif et consiste à faire des transcriptions le plus rapidement et le plus exactement possible afin de vaincre les adversaires.

Copie d'écran de http://smorballgame.org (consulté le 23 juin 2016)

Afin d'assurer la meilleure qualité de la transcription finale, les transcriptions des joueurs sont confrontées entre elles.

5- Folksonomie, catalogage et indexation participatives

5.1- Le *crowdsourcing* explicite : le tagging volontaire

5.1.1- le steve.museum

Ce projet d'indexation participative fut l'un des premiers projets culturels à faire du *crowdsourcing*, dès 2005. Il regroupe le Guggenheim Museum, le Cleveland Museum of Art, le Metropolitan Museum of Art, le San Francisco Museum of Modern Art, et Archives & Museum Informatics et a reçu 1 million de dollars de l'US Institute of Museum and Library Services.

Une première expérimentation informelle a été conduite en 2005 à partir de seulement 30 images. 80 % des mots clés obtenus par les volontaires ne faisaient nullement référence à des termes présents dans les index et thesaurii professionnels (AAT et ULAN). (Chun, 2006)

Les statistiques suivantes ont été récoltées dans la littérature (Holley, 2010 ; Paraschakis, 2013 ; Spindler, 2014) :

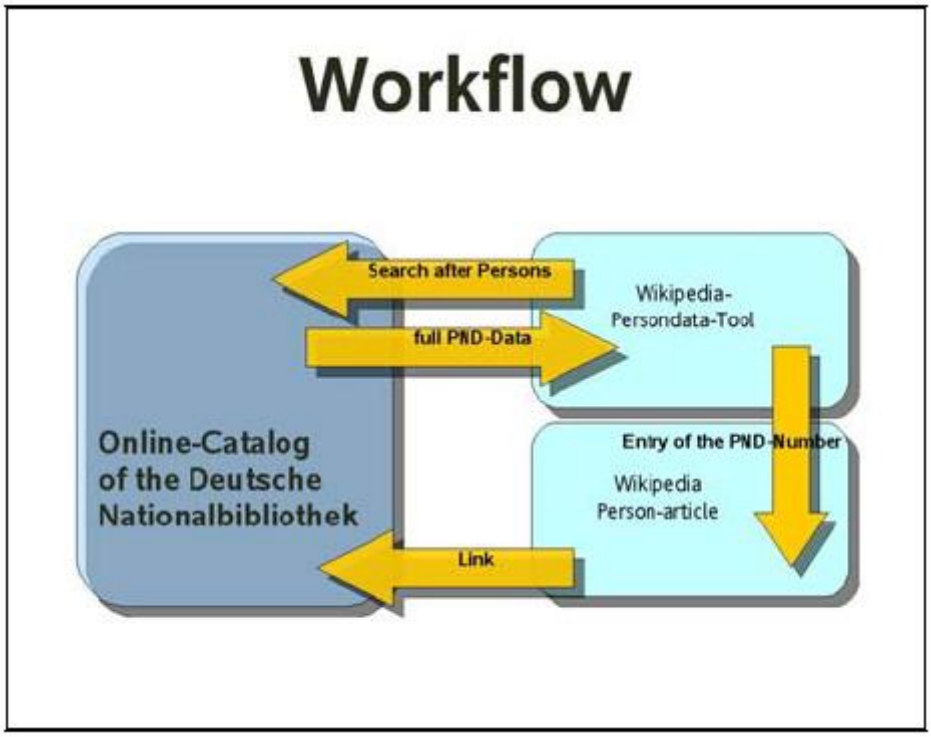
Date	**Nombre d'images à indexer**	**Nombre de tags**	**Nombre d'internautes participants**
Entre mars 2007 et mars 2008	1784 images	36 981 tags	1621 internautes
Fin 2010		468 120 tags	
Entre 2005 et 2013	96 896 objets	551 947 termes	

Fin 2010, 86 % des mots clés proposés par les internautes ne figuraient pas dans les langages documentaires utilisés par les professionnels du projet. La plateforme permet aussi bien d'ajouter que d'enlever et de corriger des termes.

D'après (Smith-Yoshimura, 2011), le site n'est pas modéré mais dispose d'une liste de "profanes" qui sont interdits d'accès.

5.1.2- GLAM Wikimedia

Un accord entre la Bibliothèque nationale d'Allemagne et Wikimedia permet d'enrichir automatiquement Wikipedia à partir du fichier des autorités de la Bibliothèque nationale d'Allemagne qui contient plus de 2 millions de noms. En échange, un lien vers le site de la Bibliothèque est systématiquement ajouté depuis les notices Wikipédia, ce qui améliore considérablement sa visibilité.



D'après (Danowski, 2007)

L'outil Wikipedia sur les données de personnes recherche via une URL spéciale créée dans l'OPAC de la Bibliothèque nationale d'Allemagne où les fichiers d'autorité sont intégrés. Depuis l'OPAC les métadonnées complètes pour une ou plusieurs personnes qui correspondent au nom recherché est retourné. L'outil présente les métadonnées du Wikipedia comparables avec les entrées de la PND à l'utilisateur. L'utilisateur décide s'il s'agissait d'une correspondance et si l'identifiant, le numéro de PND doit être ajouté dans l'article. Au moyen d'un modèle, un lien a été créé à partir de l'article dans le catalogue de la Bibliothèque. Suivant ce lien vous pouvez trouver toute la littérature de et sur cette personne spéciale dans le catalogue de la Bibliothèque. Cela est possible parce que la clé de recherche a permis de récupérer uniquement les noms de personnes différenciées. Dans un très court laps de temps (environ 2 semaines) plus de 22.000 articles ont ainsi été liés.

(Hartman, 2014) relate enfin une expérimentation simple, modeste et pragmatique sur Flickr pour l'identification de manuscrits médiévaux. Le spécialistes des

manuscrits médiévaux Micah Erwin (Ransom Humanities Research Center de l'University of Texas at Austin) a ainsi publié sur Flickr des manuscrits qu'il ne parvenait pas à identifier, a demandé de l'aide sur les réseaux sociaux et a obtenu de l'aide via Flickr.

5.1.3- Les herbonautes

L'herbier national du Muséum national d'Histoire naturelle est le plus grand herbier du monde (10 millions de parts d'herbiers) et l'un des plus anciens (16è siècle). Il contient un grand nombre de spécimens types servant d'étalons référents pour chaque espèce et est continuellement étudié par les botanistes et alimenté dans le cadre de missions visant à découvrir de nouvelles espèces, à décrire la diversité du vivant, en connaître la répartition géographique et les dynamiques de populations. Son contenu a en grande partie été numérisé dans le cadre de grands programmes financés, en particulier, par la Melone Foundation. La saisie des métadonnées a également été considérablement accélérée dans le cadre de projets internationaux.

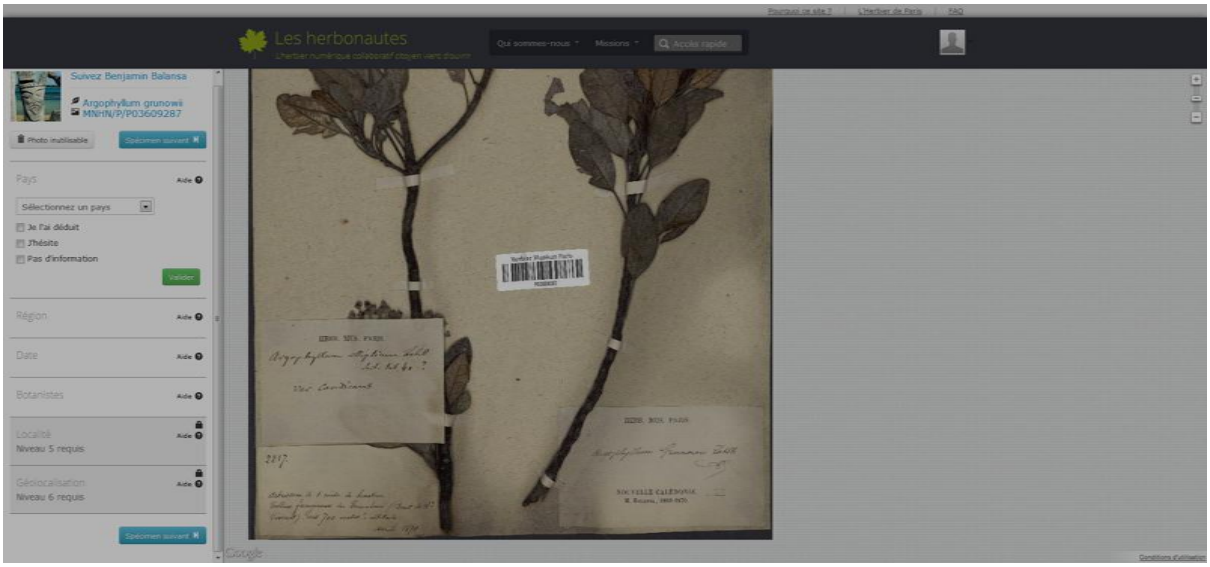
Le site de *crowdsourcing* les herbonautes a été développé par la société Bluestone et lancé en septembre 2012 afin de permettre aux internautes de participer à ce travail de saisies des métadonnées à partir des parts d'herbiers numérisées.

Afin de stimuler les internautes à participer, les ressorts de la *gamification* ont été utilisés. Chacun verra ainsi sa "carrière" sur le site évoluer en fonction du nombre de ses contributions. Pour pouvoir ainsi changer de niveau, il ne faut pas seulement avoir complété un certain nombre de notices, il faut aussi passer un quizz interactif sous forme de QCM sur la botanique.

A chaque niveau des droits en saisies sont octroyés :

- Niveau 1 : possibilité de saisir les pays
- Niveau 2 : possibilité de saisir les régions
- Niveau 3 : dates de récoltes
- Niveau 4 : botanistes récolteurs
- Niveau 5 : localités
- Niveau 6 : géolocalisation

La qualité des transcriptions des écritures manuscrites présentes sur les étiquettes est contrôlée classiquement par confrontation des saisies. L'internaute doit également pondérer la fiabilité de l'information saisie en indiquant si l'information a été déduite, s'il hésite ou s'il n'y avait pas d'information. En cas de divergences entre les saisies, un forum de discussion est proposé, sur le modèle de Wikipédia afin de résoudre la contradiction et trouver un compromis, l'expert ("chef de mission") conservant le dernier mot.



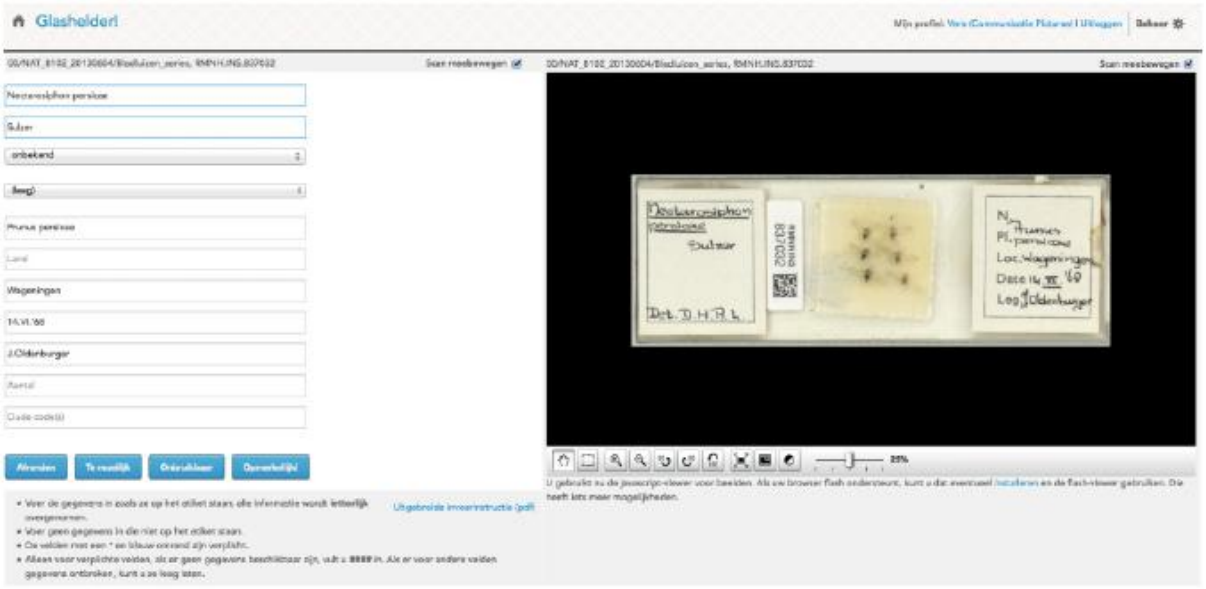
Capture d'écran des Herbonautes

Entre le 22 septembre 2012 et le 22 février 2013, 62 133 contenus de champs ont été complétés sur 2 456 spécimens d'herbier, soit une moyenne de 12 427 contributions par mois et 1677 planches d'herbier. Près de 16 % des objectifs étaient réalisés à cette date.

Ces herbiers numériques et participatifs ont parfois été qualifiés de "herbiers conversationnels" (Zacklad, 2015)

5.1.4- Les projets néerlandais Glashelder! et VeleHanden

Glashelder! est un projet assez similaire à celui des Herbonautes. Il a été lancé le 26 mars 2013 et a rassemblé 511 participants qui ont transcrits 100 000 spécimens de collections naturalistes 9 mois plus tard, à la fin du projet, le 19 janvier 2014. (Heerlien, 2015)



Capture d'écran de Glashelder! d'après (Heerlien, 2015)

VeleHanden est le résultat d'un partenariat public-privé entre les archives de la Ville d'Amsterdam et le prestataire de numérisation Pictura. Ce projet fonctionne selon un mode de double saisies des indexations ou des transcriptions. Un troisième contributeur vérifie les saisies des deux premiers contributeurs et, en cas de contradiction, valide la contribution qui lui semble la meilleure ou, s'il ne parvient pas à se décider, renvoi la décision vers le chef de projet. L'originalité de VeleHanden.nl vient du fait que les points accumulés par les contributeurs leur permettent d'obtenir des téléchargements gratuits de documents d'archives numérisés pour une valeur marchande inférieure à 0,50 € par heure de travail. Au total, en 2012, 1389 participants dont 691 hommes (dont 433 de plus de 50 ans, 145 de moins de 50 ans et 113 d'âges inconnus), 440 femmes (dont 84 de plus de 50 ans, 156 d'âges inconnus et 18 de moins de 50 ans) et 258 de genres inconnus. (Fleurbaay, 2012)

5.2- Le recours à la *gamification*

Si des projets comme les herbonautes font appel aux ressorts du jeu vidéo pour stimuler la participation des internautes, d'autres projets les invitent directement à jouer des jeux qui ont une finalité culturelle. On parle de "games with a purpose" (GWAP) ou de *gamification*. Certains de ces jeux sont apparus au périmètre du domaine de la numérisation du patrimoine. Nous les évoquons néanmoins brièvement car ils ont, par la suite, directement inspiré des jeux conçus plus spécifiquement afin de récolter des tags et des métadonnées autour de documents patrimoniaux numérisés.

5.2.1- Google Image Labeler

Développé par Luis Von Ahn et Laura Dabbish, lancé le 31 août 2006 et opérationnel jusqu'au 16 septembre 2011, date à laquelle Google a mis fin au projet, Google Image Labeler est l'un des projets de *gamification* les plus célèbres et les plus cité dans la littérature. Bien que non destiné aux seules images du patrimoine numérisé, ce jeu a directement inspiré les projets de bibliothèques comme Digitalkoot ou le projet reCAPTCHA pour Google Books développé également par Luis Von Ahn.

Le projet est né de la nécessité pour le célèbre moteur de recherche d'indexer les images du web qui jusqu'à présent, ne prend en compte que les mots qui avoisinent les images, non l'indexation des images elles-mêmes. Google Image Labeler a consisté à proposer aux internautes d'indexer ces images sous la forme d'un jeu.

Parmi de nombreux joueurs en ligne, deux d'entre eux étaient associés par le système, et chacun avait sous les yeux la même image et ils devaient chercher à trouver le même mot, le plus rapidement possible, pour la décrire. Aussitôt que les joueurs avaient proposé le même mot clé pour décrire l'image, une nouvelle image leur était proposée. L'objectif étant d'aller le plus vite possible et de décrire un maximum d'images pour marquer le plus de points possible. Ce faisant, grâce à ce jeu, Google obtenait, par confrontation des saisies, une indexation fiable des

images du web. En 2008, 200 000 joueurs ont ainsi ajouté plus de 50 millions de mots clés aux images.

5.2.2- ESP Game puis GWAP

Développé par Luis Von Ahn à la suite de sa première expérimentation avec Google Image Labeler et ouvert le 9 août 2003, l'objectif de ce jeu était relativement similaire : donner le maximum de mots clés communs avec son partenaire de jeu aux images présentées à la fois aux 2 joueurs. Néanmoins, la principale évolution apportée par Luis Von Ahn a été que les mots déjà validés au cours de précédentes parties pour l'indexation des images ont été transformés en mots tabous, c'est à dire en mots affichés et dont l'usage n'était plus autorisé pour l'indexation de ces mêmes images. Ceci permet de récolter des mots clés plus spécifiques et moins évidents tout en rendant le jeu plus difficile et plus intéressant. Lorsque les images ont suffisamment été indexées par les joueurs, ceux-ci ont alors tendance à passer n'ayant plus d'idées nouvelles de mots clés et le nombre de mots tabous étant devenu trop important. Les images sont alors retirées du jeu.
D'autres fonctionnalités nouvelles ont également été ajoutées. Il n'est plus nécessaire que 2 joueurs soient connectés simultanément, les données rentrées par un joueur précédent pouvant être préalablement enregistrées et proposées à un joueur qui lui, joue en temps réel. Des niveaux de difficultés, des barres de progression ont également été ajoutés. Et il est désormais possible de passer sur une image difficile. Afin d'éviter le vandalisme, les adresses IP des joueurs doivent nécessairement être différentes. Si un nombre anormalement élevé de personnes saisissent le même mot clé, elles sont également détectées et leurs saisies annulées. Cela permet d'éviter que plusieurs internautes s'amusent à saisir systématiquement le même mot pour éprouver ou mettre en difficulté le système.
En 2008, le jeu a été rebaptisé GWAP (Game with a purpose, ce qui signifie jeu avec un objectif)

Fonctionnement de ESP Games d'après (Von Ahn, 2004)

Captures d'écran de Google Image Labeler et ESP Game (d'après Vohn Ahn, 2004)

Capture d'écran de ESP Game d'après (Ipeirotis, 2011)

D'après (Von Ahn, 2004), du 9 août au 10 décembre 2003, en seulement 4 mois, 13 630 personnes ont joué au jeu, générant 1 271 451 mots clés pour 293 760 images différentes (sur les 350 000 images initialement sélectionnées).

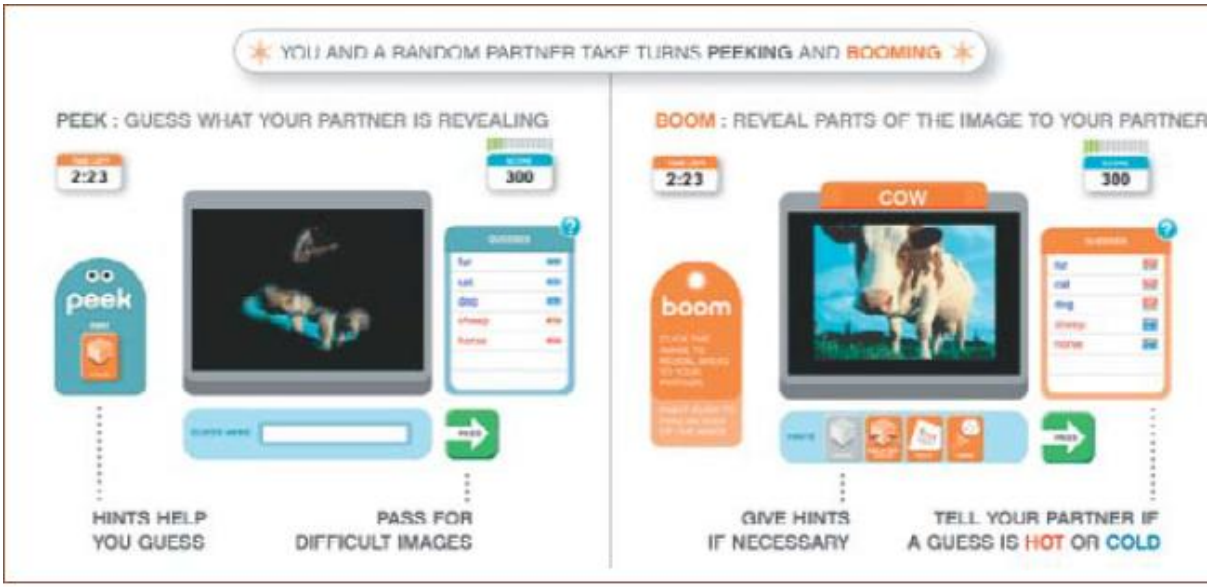
Plus de 80 % d'entre ces joueurs y ont joué plusieurs fois. Parmi eux, 33 personnes ont joué 1000 fois, ce qui représente plus de 50 heures de jeu. Plusieurs personnes s'y sont consacré plus de 40 heures par semaine.

Chaque joueur aurait joué en moyenne 91 minutes. Le nombre moyen de mots clés collectés par minute par un binôme de joueurs était de 3,89 soit 233 labels par heure. A ce rythme, d'après (Von Ahn, 2008) 5000 joueurs jouant 24 h / 24 au jeu pourraient indexer l'intégralité de Google Images (425 millions d'images) en 31 jours seulement.

5.2.3- Peekaboom

C'est un jeu, également développé par Luis Von Ahn, qui exploite les millions d'indexations collectées sur le précédent jeu, ESP Game et dont la finalité est, cette fois, d'aider les moteurs de recherche à localiser plus précisément des objets au sein des images sur le web. La tâche de distinguer et de localiser les objets, le premier plan, l'arrière plan… au sein d'une image est, en effet, très facile pour les humains mais est encore très difficile à réaliser par un algorithme.

Ce jeu se fait en binôme. Le premier joueur, “Boom” obtient une image avec un mot qui lui est associée (cf copie d’écran à droite), et doit révéler des parties de l'image, en cliquant dessus, pour le deuxième joueur “Peek” qui doit, lui, essayer de deviner le mot (cf copie d’écran à gauche). “Peek” peut entrer plusieurs suppositions que “Boom” peut voir et pour lesquelles il peut lui répondre s’il se “réchauffe” ou si, au contraire, il s’éloigne de la solution. Il est également possible de passer sur une image et un mot trop difficile à deviner.



Copie d’écran du jeu Peekaboom, d’après (Von Ahn, 2006)

Dans le cas d’un mot décrivant une partie d’un objet, “Boom” peut faire “ping” avec un clic droit sur elle et signifier ainsi à “Peek” que c’est cette partie seule qui doit être nommée. Par exemple, dans la capture d’écran suivante, il s’agit de faire deviner le mot trompe. “Peek” a donc fait un “ping” sur cette partie du corps de l’éléphant.

Copie d’écran du jeu Peekaboom, d’après (Von Ahn, 2006)

Ce faisant, “Boom” contribue à lever l'ambiguïté de la trompe avec le reste du corps de l'éléphant et permettra aux moteurs de recherche d’avoir accès à des objets georéférencés au sein d’objets plus grands.

En outre, “Boom” peut également aider “Peek” en lui signifiant quel est le type de mot à deviner (nom de personne, animal, objet, texte, verbe, associés ou non à un sujet principal).

Copie d’écran du jeu Peekaboom, d’après (Von Ahn, 2006)

Ces indications permettent d’obtenir des informations complémentaires sur la relation entre le mot et l'image. Le jeu ESP associait déjà les mots aux images, mais ne caractérisait pas comment le mot était lié à l'image. Avec Peekaboom, cette information est obtenue. Afin d’encourager les joueurs à utiliser ces indices, des points supplémentaires sont donnés pour tout mot deviné et ayant été ainsi

qualifié par un type, ce qui n'est pas le cas pour l'utilisation des indices "chaud" ou "froid", qui n'apportent pas de métadonnées et ne serviront pas en dehors du jeu.
De la même manière que le jeu permet de géoréférencer des parties d'un même objet, il est également possible de géoréférencer plusieurs objets au sein d'une même image. Une partie "bonus round" demandent aux 2 joueurs de cliquer sur l'objet spécifié dans l'image (dans la capture d'écran ci-dessous, il s'agit d'une voiture). Le nombre de points qu'ils obtiennent est proportionnel à la proximité entre chacun de leurs 2 clics.

Copie d'écran du jeu Peekaboom, d'après (Von Ahn, 2006)

Les 2 joueurs "Peek" et "Boom" bénéficient du même nombre de points à chaque mot deviné afin d'accroître la convergence de leurs intérêts et leur collaboration. Ce jeu est basé sur un modèle de collaboration et non sur la compétition, bien que le tableau des meilleurs joueurs soit affiché.
Peekaboom permet ainsi de géoréférencer des objets au sein d'images ou des parties d'objets au sein d'objets et de connaître leur type et la relation entre le mot et l'image afin de permettre à un moteur de recherche d'images de mettre en

évidence les zones de l'image de la même manière que les mots recherchés dans un texte sont mis en surbrillance par les moteurs de recherche.

D'après (Von Ahn, 2006), entre le 1er Août et le 1er septembre 2005, 14 153 joueurs ont expérimenté ce jeu générant 1 122 998 métadonnées. En moyenne chaque joueur a joué sur un peu moins de 159 images pendant près de 73 minutes. D'après (Von Ahn, 2008), en juillet 2008, ce jeu avait généré plus de 500 000 heures de jeu. Au total, près de 30 000 internautes différents ont généré, via ce jeu, environ 2 millions de données.

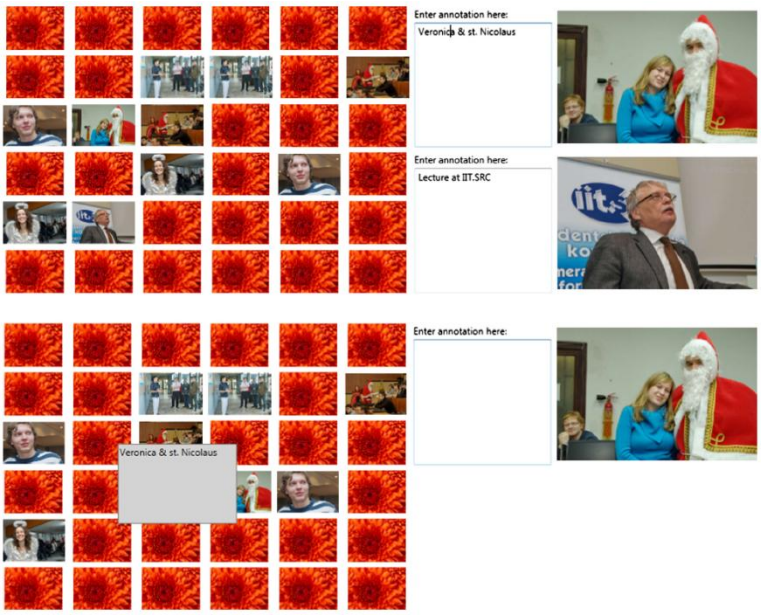
5.2.4- KissKissBan (KKB)

KissKissBan (KKB) est un jeu développé à partir de ESP game et qui ajoute un 3ème joueur, le "blocker". Ce joueur a pour mission d'entrer des mots clés bloquants et cachés afin d'entraver et de bloquer la possibilité des 2 premiers joueurs de trouver des mots communs. Il dispose de 7 secondes en début de partie pour saisir sa liste de mots bloquants. Plus il parvient à les bloquer grâce à ses mots, plus il marque de points. A la différence des mots tabous, les mots bloquants ne sont pas affichés pour le couple de joueurs et ils ne sont pas issus de mots clés déjà validés par des parties précédentes mais sont saisis par un joueur. Ce jeu fonctionne donc à la fois sur le mode collaboratif et compétitif (Ho, 2009). Il encourage les joueurs à utiliser des mots plus spécifiques et permet de recueillir un vocabulaire plus riche et évite également les possibilités de tricher. Ainsi, selon (Ho, 2009), sur 4994 étiquettes on obtiendrait en moyenne 6,56 étiquettes différentes avec ESP et 11,54 avec KKB

Capture d'écran de KissKissBan d'après (Ho, 2009)

5.2.5- PexAce

Ce jeu a été développé pour parer aux limites des Games With a Purpose (GWAP) qui nécessitent la présence simultanée de nombreux joueurs et pour éviter la malveillance de certains joueurs qui trichent ou cherchent à saboter ces jeux. PexAce fonctionne sur le modèle du jeu pour enfant memory et ne nécessite donc qu'un seul joueur.



Interface de PexAce d'après (Simko, 2013)

Dans la capture d'écran en haut, le joueur retourne deux cartes d'images sur le plateau de jeu (à gauche), il examine les 2 images affichées en grand (dans la partie de droite) et peut saisir ses annotations personnelles dans le formulaire situé au centre, entre le plateau de jeu de gauche et les grandes images à annoter. Plus tard, lorsque le joueur tombe sur une image qu'il a vu avant ailleurs (écran du bas), il peut revoir ses annotations en déplaçant le curseur de la souris sur les cartes cachées. Cela lui permet de découvrir plus facilement la paire du memory.
Ce jeu de memory permet donc, grâce à l'ajout de la fonctionnalité annotation, d'obtenir une indexation des images.

5.2.6- museumgam.es

Ce jeu a été construit avec l'aide de plugins du CMS WordPress et de développements PHP, MySQL, HTML et CSS. Les données qui font l'objet des jeux ont été récupérées via une API du Science Museum et une API du Powerhouse Museum.
Les jeux suivants ont été développés :

- Simple tagging : taguer les images
- Fact adding : ajouter des informations sur des objets de musée
- Donald's detective puzzle
- Dora's lost data : dans ce jeu, Dora est une jeune conservateur du patrimoine qui a accidentellement supprimé toutes les informations qu'elle allait ajouter à ses collections en ligne. On doit donc rapidement l'aider à ré-indexer tous les documents. A la fin de la partie, Dora félicite le joueur, lui donne son score, des indications sur sa performance (nombre moyen de tags par documents...) et l'invite à jouer à nouveau une partie.

Capture d'écran de la page d'accueil du jeu "Donald's detective puzzle", d'après (Ridge, 2011)

Pendant la période d'évaluation entre le 30 Novembre et le 1er 1 Mars, 196 sessions de jeu ont été jouées par 47 joueurs enregistrés (avec une légère domination sociologique des femmes d'une trentaine d'années). Au total 1079 parties ont été jouées (en moyenne 5,51 parties par session) qui ont permis de récolter 6039 tags (18 par document en moyenne), 2232 tags uniques.

5.2.7- Metadata Games

Inspiré de jeux déjà existants, le projet Metadata Games vise à créer des jeux open source de *gamification* et s'adresse tout particulièrement aux institutions

culturelles. L'expérimentation, portée par le Tiltfactor Laboratory du Dartmouth College a commencé avec la Bibliothèque de Dartmouth College puis l'Université de Washington, la Bibliothèque Publique de Boston, L'université de Buffalo, UC-Santa Cruz et Hong Kong ont rejoint le projet. Durant les 8 premiers mois du projet, 7 jeux ont été développés en HTML5 avec le soutien du National Endowment for the Humanities (NEH) et du American Council of Learned Societies (ACLS). (Flanagan, 2012).

Le jeu Zen Tag propose une simple description d'images en récompensant les contributions par des points et en récompensant d'avantage les tags non déjà saisis sur une image. Le jeu Cattygory propose une indexation plus structurée, par champs. Le jeu Zen Pond consiste à proposer la même image simultanément à 2 joueurs et à leur attribuer des points s'ils trouvent les mêmes tags. Le jeu Guess What! est également un jeu à 2 joueurs. Un premier joueur a une image sous les yeux et il doit faire deviner, grâce à des mots clés, à un deuxième joueur quelle est l'image correspondante parmi 12 images. Ce faisant, il indexe son image et permet de faire valider son indexation par un autre joueur.

Capture d'écran du jeu Guess What:

(d'après http://blogs.loc.gov/digitalpreservation/files/2013/03/guesswhat-320x266.png, consulté le 23 juin 2016)

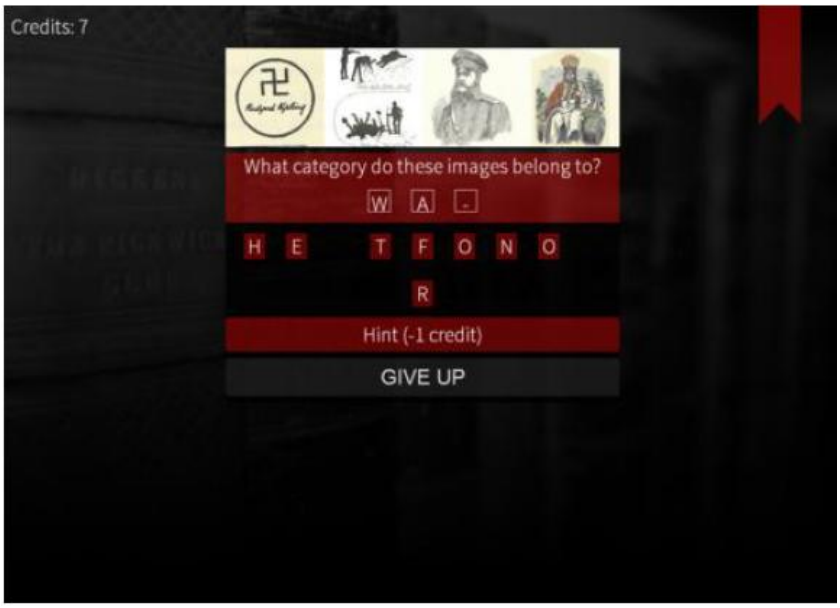
5.2.8- SaveMyHeritage

SaveMyHeritage est un jeu culturel sur les réseaux sociaux qui met en avant la compétition entre les joueurs car, d'après les conclusions de cette expérimentation, ce type de motivation resterait la plus efficace. Ainsi, pour chaque photo sur laquelle il a contribué chaque participant gagne 3 points, ainsi que 1 point par tag ou un élément de métadonnées. Un mini-classement sur le côté de la galerie montre les trois principaux utilisateurs ainsi que deux utilisateurs immédiatement mieux classés et moins bien classés que l'utilisateur actuel. (Havenga, 2012)

Capture d'écran de SaveMyHeritage (d'après Havenga, 2012)

5.2.9- Picaguess

Développé par la British Library, ce jeu Android de devinettes des images que la bibliothèque a diffusé sous Flickr. A partir de lettres, un joueur est invité à deviner une catégorie thématique pour 4 images telle que définie par d'autres joueurs. Il peut également demander des indices.



Capture d'écran de Picaguess

d'après http://britishlibrary.typepad.co.uk/digital-scholarship/2015/01/picaguess-a-crowdsourcing-app-from-the-british-library-experiment.html (consulté le 23 juin 2016)

Dans cette annexe au panorama des projets qui se voulait la plus exhaustive possible, nous n'avons toutefois pas évoqué les projets suivants. Nous les mentionnons donc ci-après :

- Addressing history, University of Edinburgh & National Library of Scotland (UK)
- Alto Editor IMPACT Centre of Competence
- Civil War Diaries & Letters Transcription Project The University of Iowa Libraries
- Crowd4U (JAPON) : plateforme mutualisée de projets *crowdsourcing* lancée en 2010
- Dickens Journals Online (UK) : correction participative d'OCR
- Family Search Indexing Family Search (depuis 2004, 780 000 volontaires, 100 000 volontaires actifs par mois, 1 500 088 741 notices indexée en juillet 2012)
- FieldData Atlas of Living Australia/Gaia Resources
- Harold "Doc" Edgerton Project MIT?
- Islandora TEI Editor UPEI, Robertson Library
- Itineranova-Editor Stadsarchief Leuven/HKI Cologne
- L-Crowd (JAPON) : lancé en 2012 par des bibliothèques universitaires japonaises pour la correction de métadonnées
- Metadata Games : jeu développé par le Tiltfactor Laboratory (Dartmouth Collegep) et proposant aux internautes de tagger des photos, des enregistrements audio ou vidéos de bibliothèques ou d'archives.
- Metadata Games project : jeu développé par l'Université de Munich et proposant un jeu de tagging sur des photos.
- Marine Lives (UK) : transcription de manuscrits de la marine anglaise du 17e siècle
- National Archives Transcription Pilot Project U.S. National Archives
- North American Bird Phenology Program USGS
- OpenScribe
- Prism, University of Virginia (USA)

- Project Runeberg (http://runeberg.org, consulté le 23 juin 2016)
- PyBOSSA Citizen Cyberscience Centre/OKFN
- Scribe Zooniverse
- Scripto Center for History and New Media at George Mason University
- Son of Suda On-Line Integrating Digital Papyrology
- TextLab John Bryant,et al, Hoftsta University
- Unbindery Ben Crowder
- VdU-Editor Monasterium.net/HKI Cologne
- Velehanden.nl
- Veridian DL Consulting
- Virtual Transcription Laboratory Poznań Supercomputing and Networking Center
- Wiki::Score ?
- Word Soup (http://cat.iti.upv.es/wordsoup consulté le 23 juin 2016)
- World Archives Project Ancestry.com

Annexe 2- Une analyse du panorama des projets de crowdsourcing en bibliothèques

A partir du panorama des projets que nous avons réalisé et dont une bonne partie se trouve en annexes, nous avons cherché à produire des analyses historiques, géographiques et taxonomique.

Histoire des projets de *crowdsourcing* en bibliothèques

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
1994	UK	Impression à la Demande	*Crowdfunding*	Higher Education Resources ON Demand
1996	USA	Mise en ligne participative	*Crowdsourcing* explicite	Internet Archive
1997	France	Numérisation à la demande	*Crowdfunding*	Le livre à la carte, Phénix éditions
2000	USA	Correction de l'OCR	*Crowdsourcing* explicite	Distributed Proofreaders
2003	USA	Correction de l'OCR	*Crowdsourcing* explicite	Wikisource
2003	USA	Indexation	*Gamification*	ESP Game
2004	France	Numérisation à la demande	*Crowdfunding*	Chapitre.com
2005	USA	Indexation	*Crowdsourcing* explicite	steve.museum
2005	USA	Tous types de tâches	*Crowdsourcing* rémunéré	Amazon Mechanical Turk Marketplace

2006	Autriche	Numérisation à la demande	*Crowdfunding*	Ebooks on Demand
2006	USA	Impression à la Demande	*Crowdfunding*	Espresso Book Machine
2006	USA	Indexation	*Gamification*	Google Image Labeler
2006	USA	Indexation	*Gamification*	Peekaboom
2008	Australie	Correction de l'OCR	*Crowdsourcing* explicite	TROVE
2008	UK	Mise en ligne participative	*Crowdsourcing* explicite	l'Oxford's great war archive
2008	USA	Correction de l'OCR	*Crowdsourcing* implicite	reCAPTCHA
2008	USA	Indexation	*Crowdsourcing* explicite	Flickr: The Commons
2009	Israël	Correction de l'OCR	*Gamification*	COoperative eNgine for Correction of ExtRacted Text
2009	USA	Impression à la Demande	*Crowdfunding*	Amazon BookSurge
2009	USA	Indexation	*Gamification*	KissKissBan
2010	UK	Transcription de manuscrits	*Crowdsourcing* explicite	Transcribe Bentham
2011	Australie	Transcription de manuscrits	*Crowdsourcing* explicite	ArcHIVE
2011	Finlande	Correction de l'OCR	*Gamification*	Digitalkoot
2011	France	Impression à la Demande	*Crowdfunding*	*Print on Demand* sur Gallica

2011	France	Numérisation à la demande	*Crowdfunding*	Adopter un livre de la BnF
2011	Singapour	Correction de l'OCR	*Gamification*	TypeAttack
2011	UK	Transcription de manuscrits	*Crowdsourcing* explicite	Ancient Lives
2011	USA	Correction de l'OCR	*Crowdsourcing* explicite	California Digital Newspaper Collection
2011	USA	Indexation	*Gamification*	museumgam.es
2011	USA	Transcription de manuscrits	*Crowdsourcing* explicite	Do it yoursel History
2011	USA	Transcription de manuscrits	*Crowdsourcing* explicite	What's on the menu ?
2012	France	Indexation	*Crowdsourcing* explicite	Les herbonautes
2012	UK	Transcription de manuscrits	*Crowdsourcing* explicite	What's the score
2012	USA	Indexation	*Gamification*	Metadata Games
2012	USA	Indexation	*Gamification*	SaveMyHeritage
2013	France	Numérisation à la demande	*Crowdfunding*	Numalire
2013	Pays bas	Indexation	*Crowdsourcing* explicite	Glashelder! et VeleHanden
2013	USA	Numérisation à la demande	*Crowdfunding*	Revealdigital
2014	France	Correction de l'OCR	*Crowdsourcing* explicite	Correct
2014	France	Mise en ligne participative	*Crowdsourcing* explicite	Europeana 1914-1918
2014	Suède	Indexation	*Gamification*	Art collector

Hormis dans le domaine du *crowdfunding* où la France et l'Autriche ont été précurseurs, les USA dominent largement les débuts du *crowdsourcing* en bibliothèque.

Géographie des projets de crowdsourcing en bibliothèques

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
2008	Australie	Correction de l'OCR	*Crowdsourcing* explicite	TROVE
2011	Australie	Transcription de manuscrits	*Crowdsourcing* explicite	ArcHIVE

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
2006	Autriche	Numérisation à la demande	*Crowdfunding*	Ebooks on Demand

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
2011	Finlande	Correction de l'OCR	*Gamification*	Digitalkoot

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
1997	France	Numérisation à la demande	*Crowdfunding*	Le livre à la carte, Phénix éditions
2004	France	Numérisation à la demande	*Crowdfunding*	Chapitre.com

2011	France	Impression à la Demande	*Crowdfunding*	*Print on Demand* sur Gallica
2011	France	Numérisation à la demande	*Crowdfunding*	Adopter un livre de la BnF
2012	France	Indexation	*Crowdsourcing* explicite	Les herbonautes
2013	France	Numérisation à la demande	*Crowdfunding*	Numalire
2014	France	Correction de l'OCR	*Crowdsourcing* explicite	Correct
2014	France	Mise en ligne participative	*Crowdsourcing* explicite	Europeana 1914-1918

Très peu de projets de *crowdsourcing* ont été identifiés en France et aucun projet de *gamification*, ou de *crowdsourcing* implicite. Par contre, la France occupe une position de pionnière dans le domaine du *crowdfunding* appliqué à la numérisation à la demande.

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
2009	Israël	Correction de l'OCR	*Gamification*	COoperative eNgine for Correction of ExtRacted Text

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
2013	Pays bas	Indexation	*Crowdsourcing* explicite	Glashelder! et VeleHanden

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
2011	Singapour	Correction de l'OCR	*Gamification*	TypeAttack

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
2014	Suède	Indexation	*Gamification*	Art collector

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
1994	UK	Impression à la Demande	*Crowdfunding*	Higher Education Resources ON Demand
2008	UK	Mise en ligne participative	*Crowdsourcing* explicite	l'Oxford's great war archive
2010	UK	Transcription de manuscrits	*Crowdsourcing* explicite	Transcribe Bentham
2011	UK	Transcription de manuscrits	*Crowdsourcing* explicite	Ancient Lives
2012	UK	Transcription de manuscrits	*Crowdsourcing* explicite	What's the score

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
1996	USA	Mise en ligne participative	*Crowdsourcing* explicite	Internet Archive
2000	USA	Correction de l'OCR	*Crowdsourcing* explicite	Distributed Proofreaders
2003	USA	Correction de l'OCR	*Crowdsourcing* explicite	Wikisource
2003	USA	Indexation	*Gamification*	ESP Game
2005	USA	Indexation	*Crowdsourcing* explicite	steve.museum
2005	USA	Tous types de tâches	*Crowdsourcing* rémunéré	Amazon Mechanical Turk Marketplace
2006	USA	Impression à la Demande	*Crowdfunding*	Espresso Book Machine
2006	USA	Indexation	*Gamification*	Google Image Labeler
2006	USA	Indexation	*Gamification*	Peekaboom
2008	USA	Correction de l'OCR	*Crowdsourcing* implicite	reCAPTCHA
2008	USA	Indexation	*Crowdsourcing* explicite	Flickr: The Commons
2009	USA	Impression à la Demande	*Crowdfunding*	Amazon BookSurge
2009	USA	Indexation	*Gamification*	KissKissBan
2011	USA	Correction de l'OCR	*Crowdsourcing* explicite	California Digital Newspaper Collection
2011	USA	Indexation	*Gamification*	museumgam.es
2011	USA	Transcription de manuscrits	*Crowdsourcing* explicite	Do it yoursel History
2011	USA	Transcription	*Crowdsourcing*	What’s on the menu ?

		de manuscrits	explicite	
2012	USA	Indexation	*Gamification*	Metadata Games
2012	USA	Indexation	*Gamification*	SaveMyHeritage
2013	USA	Numérisation à la demande	*Crowdfunding*	Revealdigital

Les tâches externalisées par *crowdsourcing* en bibliothèques

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
2000	USA	Correction de l'OCR	*Crowdsourcing* explicite	Distributed Proofreaders
2003	USA	Correction de l'OCR	*Crowdsourcing* explicite	Wikisource
2008	Australie	Correction de l'OCR	*Crowdsourcing* explicite	TROVE
2011	USA	Correction de l'OCR	*Crowdsourcing* explicite	California Digital Newspaper Collection
2014	France	Correction de l'OCR	*Crowdsourcing* explicite	Correct
2008	USA	Correction de l'OCR	*Crowdsourcing* implicite	reCAPTCHA
2009	Israël	Correction de l'OCR	*Gamification*	COoperative eNgine for Correction of ExtRacted Text
2011	Finlande	Correction de l'OCR	*Gamification*	Digitalkoot
2011	Singapour	Correction de l'OCR	*Gamification*	TypeAttack

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
1994	UK	Impression à la Demande	*Crowdfunding*	Higher Education Resources ON Demand
2006	USA	Impression à la Demande	*Crowdfunding*	Espresso Book Machine
2009	USA	Impression à la Demande	*Crowdfunding*	Amazon BookSurge
2011	France	Impression à la Demande	*Crowdfunding*	*Print on Demand* sur Gallica

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
2005	USA	Indexation	*Crowdsourcing* explicite	steve.museum
2008	USA	Indexation	*Crowdsourcing* explicite	Flickr: The Commons
2012	France	Indexation	*Crowdsourcing* explicite	Les herbonautes
2013	Pays bas	Indexation	*Crowdsourcing* explicite	Glashelder! et VeleHanden
2003	USA	Indexation	*Gamification*	ESP Game
2006	USA	Indexation	*Gamification*	Google Image Labeler
2006	USA	Indexation	*Gamification*	Peekaboom
2009	USA	Indexation	*Gamification*	KissKissBan
2011	USA	Indexation	*Gamification*	museumgam.es
2012	USA	Indexation	*Gamification*	Metadata Games
2012	USA	Indexation	*Gamification*	SaveMyHeritage
2014	Suède	Indexation	*Gamification*	Art collector

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
1996	USA	Mise en ligne participative	*Crowdsourcing* explicite	Internet Archive
2008	UK	Mise en ligne participative	*Crowdsourcing* explicite	l'Oxford's great war archive
2014	France	Mise en ligne participative	*Crowdsourcing* explicite	Europeana 1914-1918

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
1997	France	Numérisation à la demande	*Crowdfunding*	Le livre à la carte, Phénix éditions
2004	France	Numérisation à la demande	*Crowdfunding*	Chapitre.com
2006	Autriche	Numérisation à la demande	*Crowdfunding*	Ebooks on Demand
2011	France	Numérisation à la demande	*Crowdfunding*	Adopter un livre de la BnF
2013	France	Numérisation à la demande	*Crowdfunding*	Numalire
2013	USA	Numérisation à la demande	*Crowdfunding*	Revealdigital

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
2010	UK	Transcription de manuscrits	*Crowdsourcing* explicite	Transcribe Bentham
2011	Australie	Transcription de manuscrits	*Crowdsourcing* explicite	ArcHIVE
2011	UK	Transcription de manuscrits	*Crowdsourcing* explicite	Ancient Lives
2011	USA	Transcription de manuscrits	*Crowdsourcing* explicite	Do it yoursel History
2011	USA	Transcription de manuscrits	*Crowdsourcing* explicite	What's on the menu ?
2012	UK	Transcription de manuscrits	*Crowdsourcing* explicite	What's the score

Taxonomie du *crowdsourcing* en bibliothèques

Date	Pays	Type de tâche	Type de crowdsourcing	Projet
1994	UK	Impression à la Demande	*Crowdfunding*	Higher Education Resources ON Demand
2006	USA	Impression à la Demande	*Crowdfunding*	Espresso Book Machine
2009	USA	Impression à la Demande	*Crowdfunding*	Amazon BookSurge
2011	France	Impression à la Demande	*Crowdfunding*	*Print on Demand* sur Gallica
1997	France	Numérisation à la	*Crowdfunding*	Le livre à la carte,

		demande		Phénix éditions
2004	France	Numérisation à la demande	*Crowdfunding*	Chapitre.com
2006	Autriche	Numérisation à la demande	*Crowdfunding*	Ebooks on Demand
2011	France	Numérisation à la demande	*Crowdfunding*	Adopter un livre de la BnF
2013	France	Numérisation à la demande	*Crowdfunding*	Numalire
2013	USA	Numérisation à la demande	*Crowdfunding*	Revealdigital

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
2000	USA	Correction de l'OCR	*Crowdsourcing* explicite	Distributed Proofreaders
2003	USA	Correction de l'OCR	*Crowdsourcing* explicite	Wikisource
2008	Australie	Correction de l'OCR	*Crowdsourcing* explicite	TROVE
2011	USA	Correction de l'OCR	*Crowdsourcing* explicite	California Digital Newspaper Collection
2014	France	Correction de l'OCR	*Crowdsourcing* explicite	Correct
2005	USA	Indexation	*Crowdsourcing* explicite	steve.museum
2008	USA	Indexation	*Crowdsourcing*	Flickr: The Commons

			explicite	
2012	France	Indexation	*Crowdsourcing* explicite	Les herbonautes
2013	Pays bas	Indexation	*Crowdsourcing* explicite	Glashelder! et VeleHanden
1996	USA	Mise en ligne participative	*Crowdsourcing* explicite	Internet Archive
2008	UK	Mise en ligne participative	*Crowdsourcing* explicite	l'Oxford's great war archive
2014	France	Mise en ligne participative	*Crowdsourcing* explicite	Europeana 1914-1918
2010	UK	Transcription de manuscrits	*Crowdsourcing* explicite	Transcribe Bentham
2011	Australie	Transcription de manuscrits	*Crowdsourcing* explicite	ArcHIVE
2011	UK	Transcription de manuscrits	*Crowdsourcing* explicite	Ancient Lives
2011	USA	Transcription de manuscrits	*Crowdsourcing* explicite	Do it yoursel History
2011	USA	Transcription de manuscrits	*Crowdsourcing* explicite	What's on the menu ?
2012	UK	Transcription de manuscrits	*Crowdsourcing* explicite	What's the score

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
2008	USA	Correction de l'OCR	*Crowdsourcing* implicite	reCAPTCHA

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
2005	USA	Tous types de tâches	*Crowdsourcing* rémunéré	Amazon Mechanical Turk Marketplace

Date	Pays	Type de tâche	Type de *crowdsourcing*	Projet
2009	Israël	Correction de l'OCR	*Gamification*	COoperative eNgine for Correction of ExtRacted Text
2011	Finlande	Correction de l'OCR	*Gamification*	Digitalkoot
2011	Singapour	Correction de l'OCR	*Gamification*	TypeAttack
2003	USA	Indexation	*Gamification*	ESP Game
2006	USA	Indexation	*Gamification*	Google Image Labeler
2006	USA	Indexation	*Gamification*	Peekaboom
2009	USA	Indexation	*Gamification*	KissKissBan
2011	USA	Indexation	*Gamification*	museumgam.es
2012	USA	Indexation	*Gamification*	Metadata Games
2012	USA	Indexation	*Gamification*	SaveMyHeritage
2014	Suède	Indexation	*Gamification*	Art collector

Annexe 3- Résultats de l'enquête auprès des usagers de Numalire

70,59 % des répondants sont des hommes avec des métiers divers (Enseignant chercheur, professeur, égyptologue, universitaire, Étudiant, Cadre supérieur, juge, verrier…)

Pour 51,43 % d'entre eux, les documents ont été demandés dans le cadre de leur travail ou de leur recherches (recherche universitaire / études, recherche dans un cadre professionnel, préparation d'un ouvrage, recherche personnelle, recherche généalogique et historique…). On trouve, par exemple :

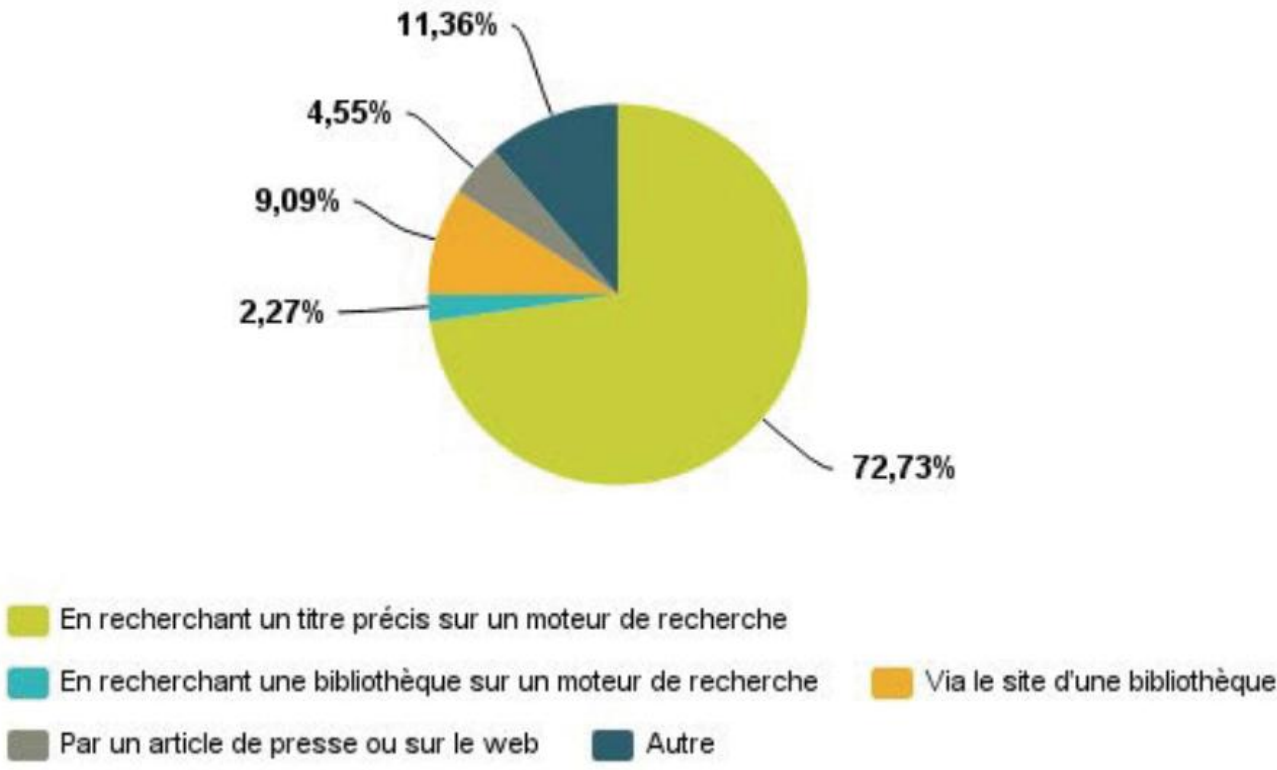
« Demande faites dans le cadre d'une recherche universitaire. Le livre, italien, est très peu diffusé en France et ne semble présent qu'à la bibliothèque Ste Geneviève. »

« Arrière petit fils de l'auteur du document dans l'impossibilité de me procurer ce livre.»

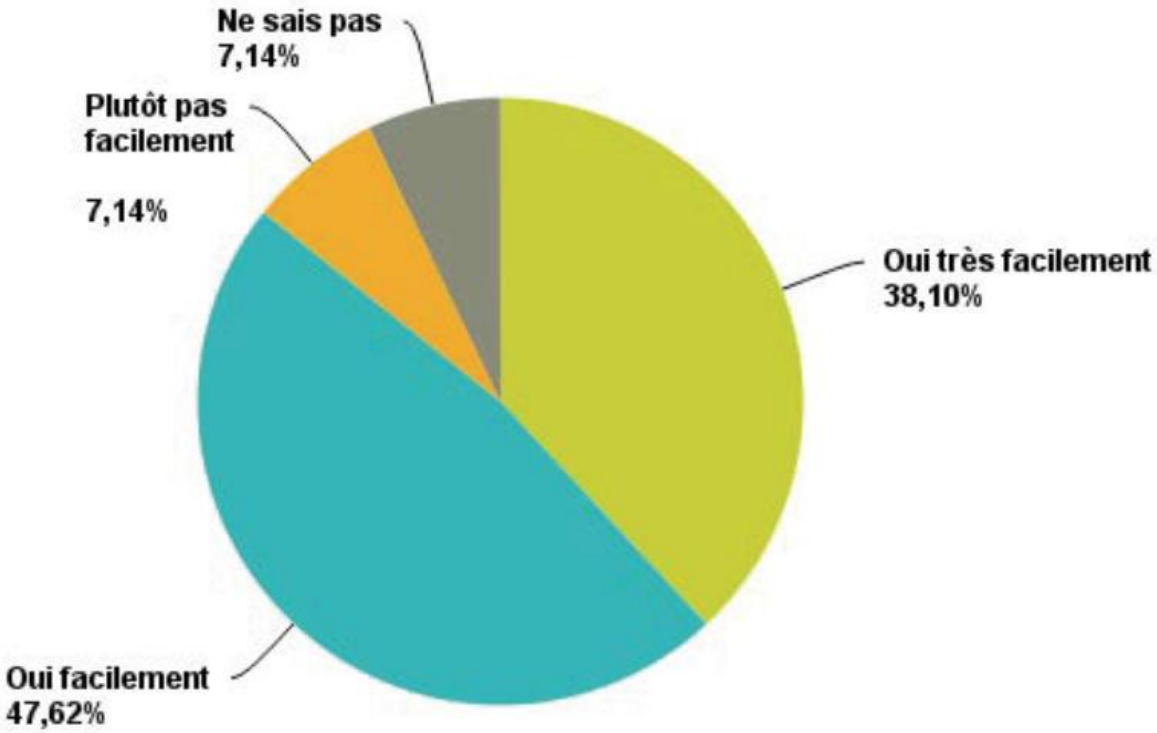
« Publication introuvable autrement, recherche pour une biographie. »

« Recherche sur des procédés disparus. »

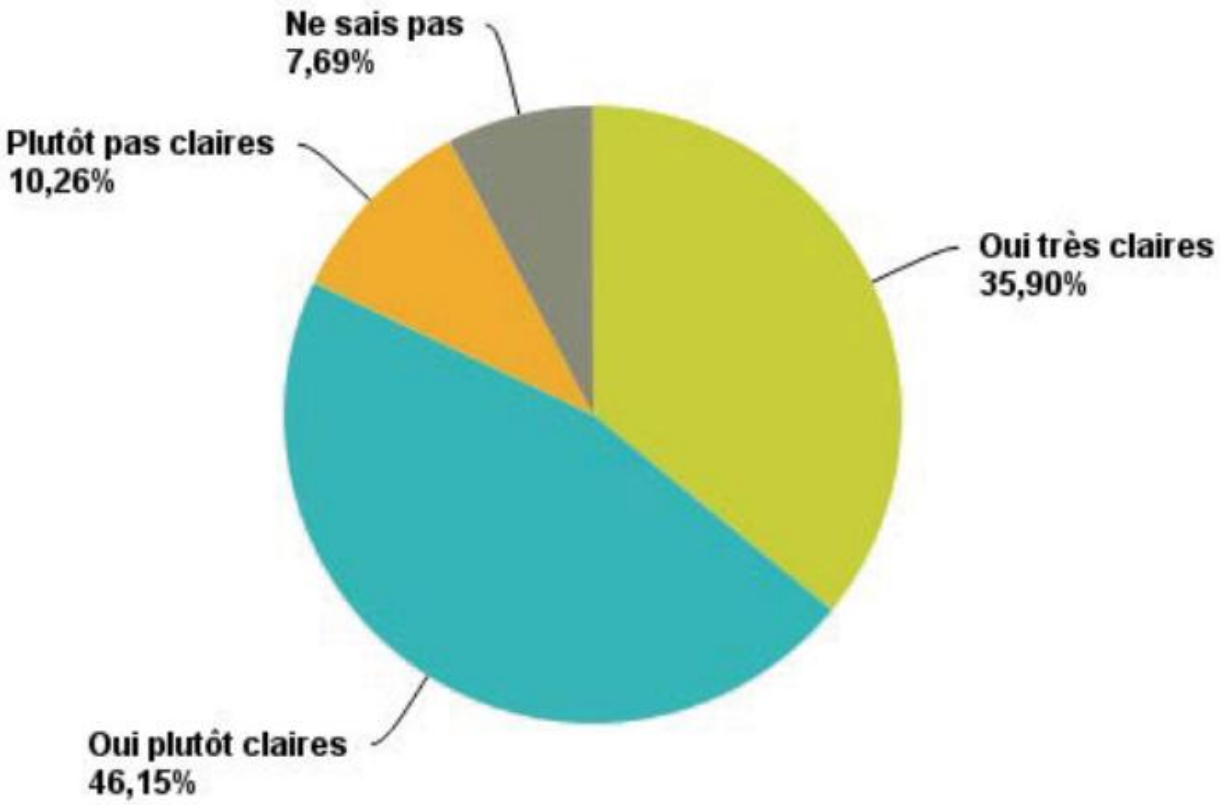
« Simplement d'avoir une copie papier permettant un partage plus simple. »



Répartition des réponses à la question comment avez-vous connu Numalire ?

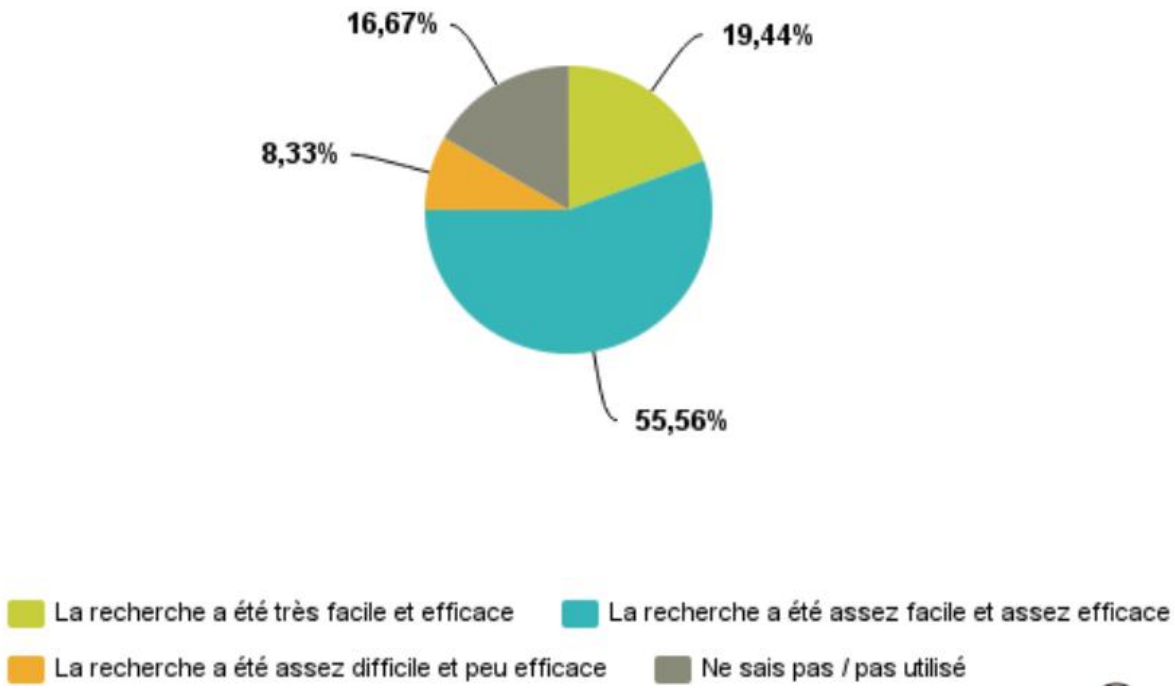


Répartition des réponses à la question “Avez-vous trouvé facilement les informations sur le fonctionnement du site Numalire ?”

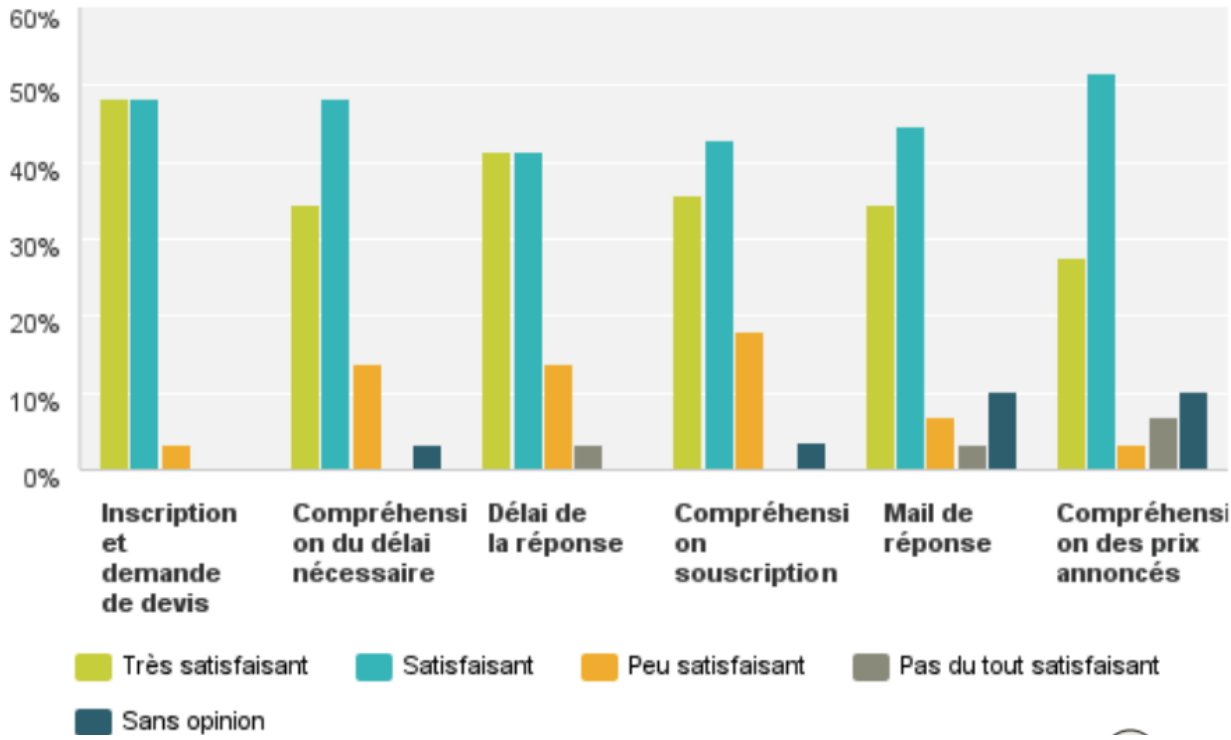


Répartition des réponses à la question "les informations présentant le fonctionnement de Numalire - notamment le système de souscription, vous ont-elles parues claires, suffisamment compréhensibles ?"

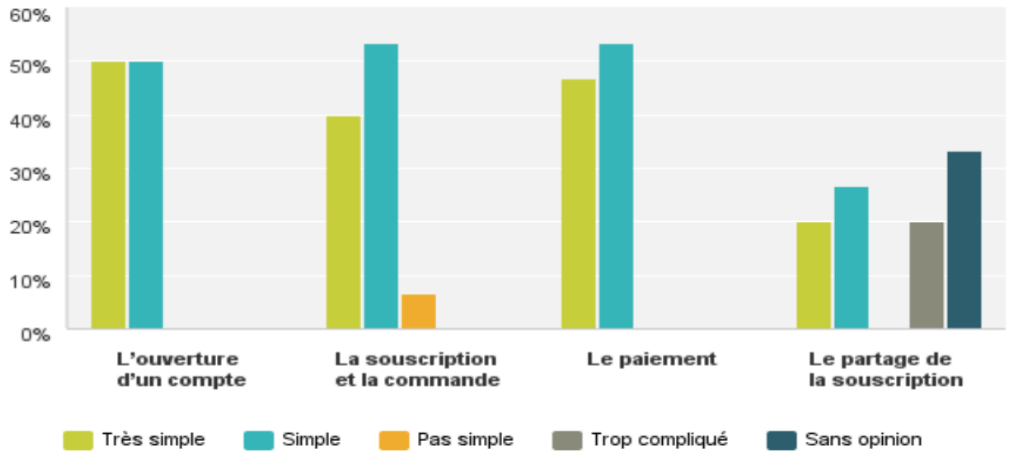
Le principe de souscription a parfois mal été compris, en particulier par ceux qui y ont été invités via les réseaux sociaux ("J'ai tenté de faire participer d'autres personnes qui n'ont rien compris au système…"). Ces fonctionnalités doivent donc être simplifiées et clarifiées.



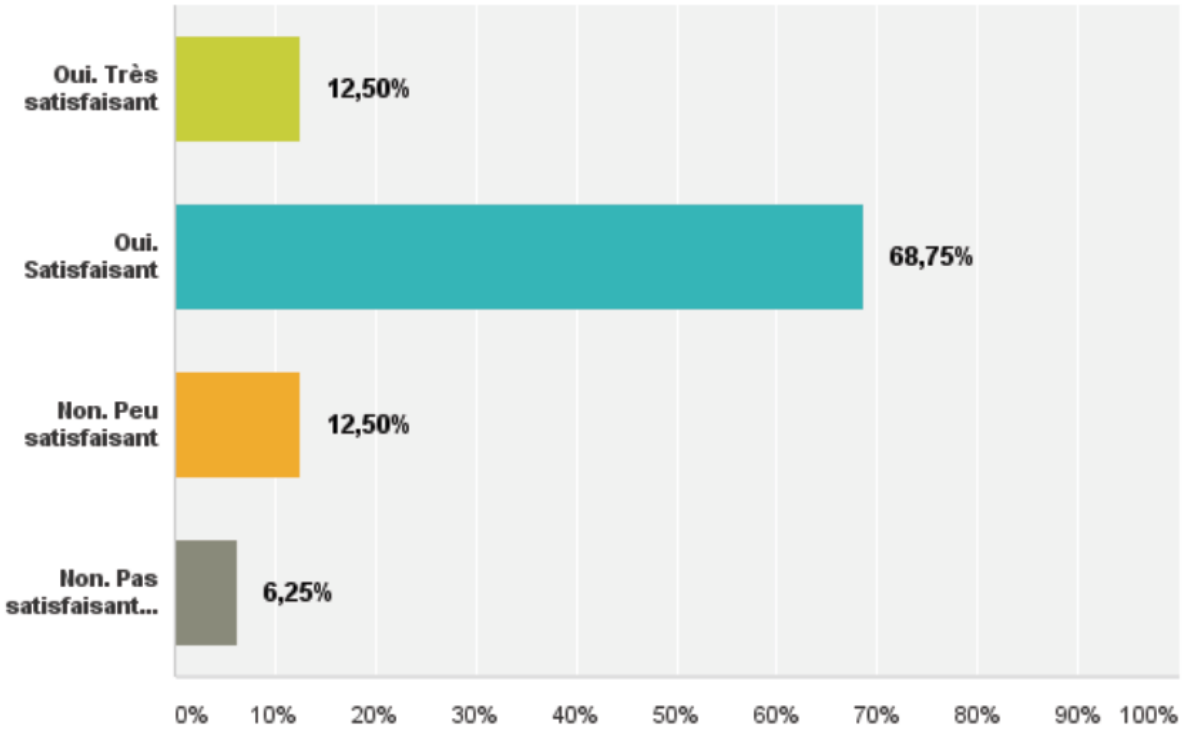
Répartition des réponses à la question "Au cours de vos visites sur le site Numalire, avez-vous eu l'occasion d'utiliser la fonction recherche ? Si oui, pouvez-vous préciser votre expérience ?"



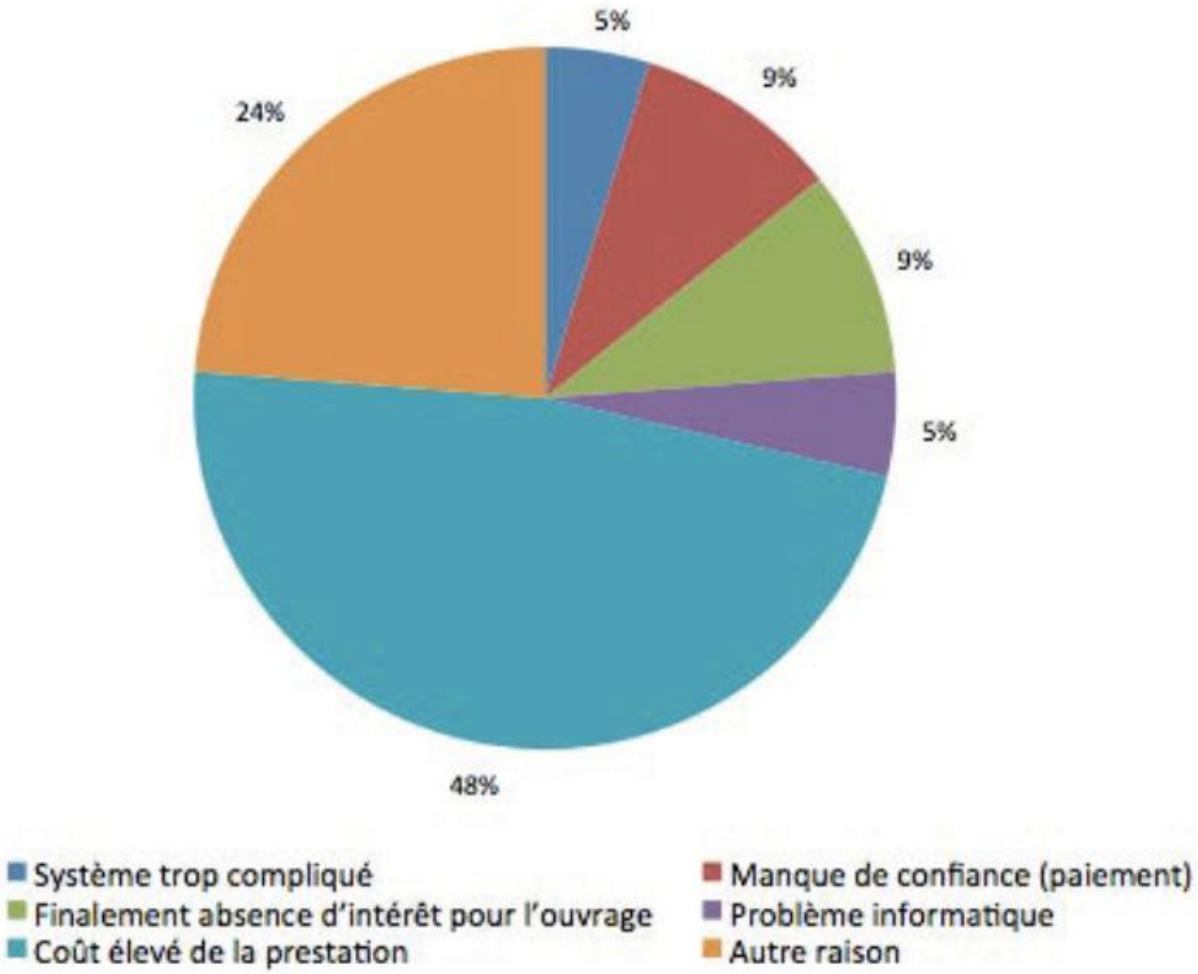
Réponses à la question "Au terme de votre recherche, vous avez souhaité demander un devis. Quelques questions sur cette étape."



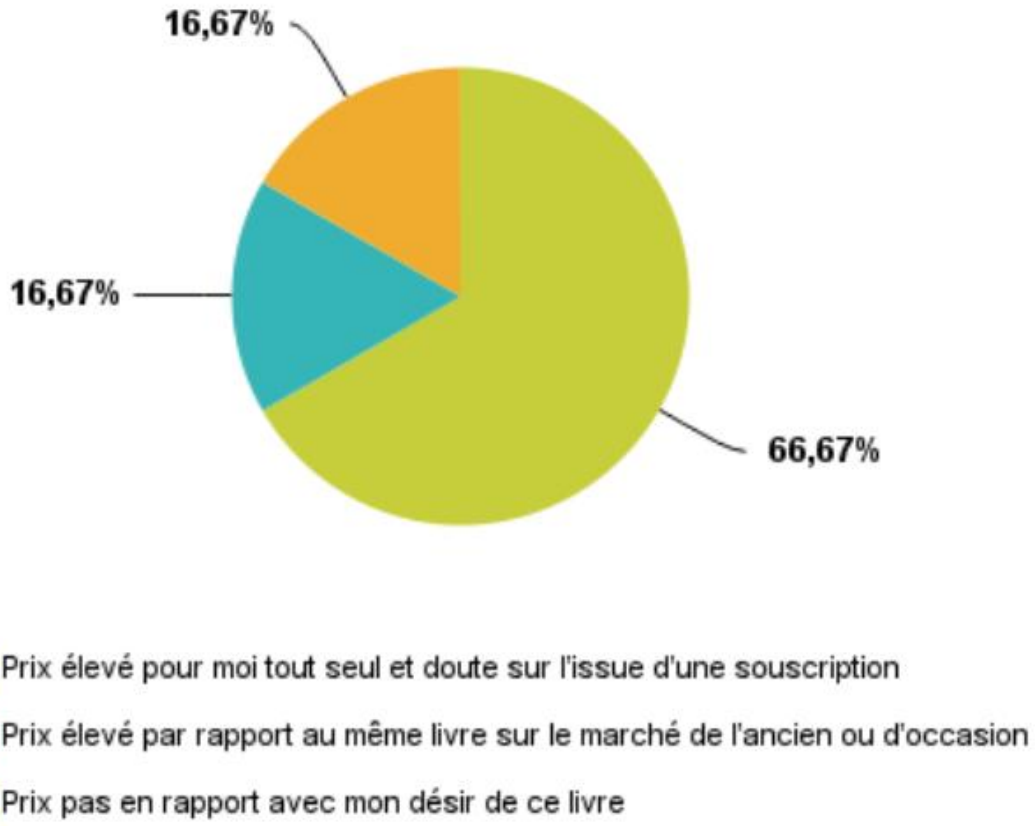
Réponses à la question "A la réception du devis vous avez lancé une souscription et/ou passé une commande. Pouvez-vous juger de la simplicité des outils mis à votre disposition ?"



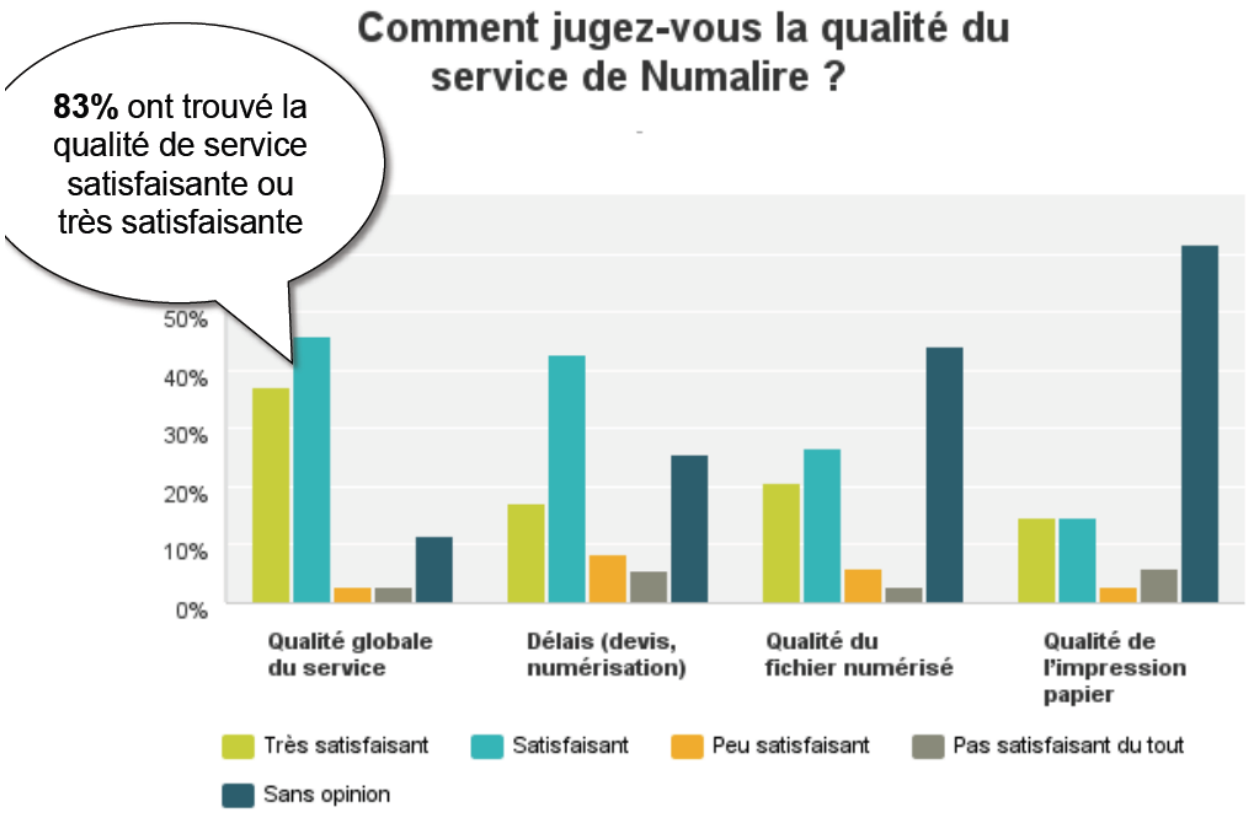
Réponses à la question “Les informations de suivi de la souscription sur le site et par mail ont-elles été suffisantes ?



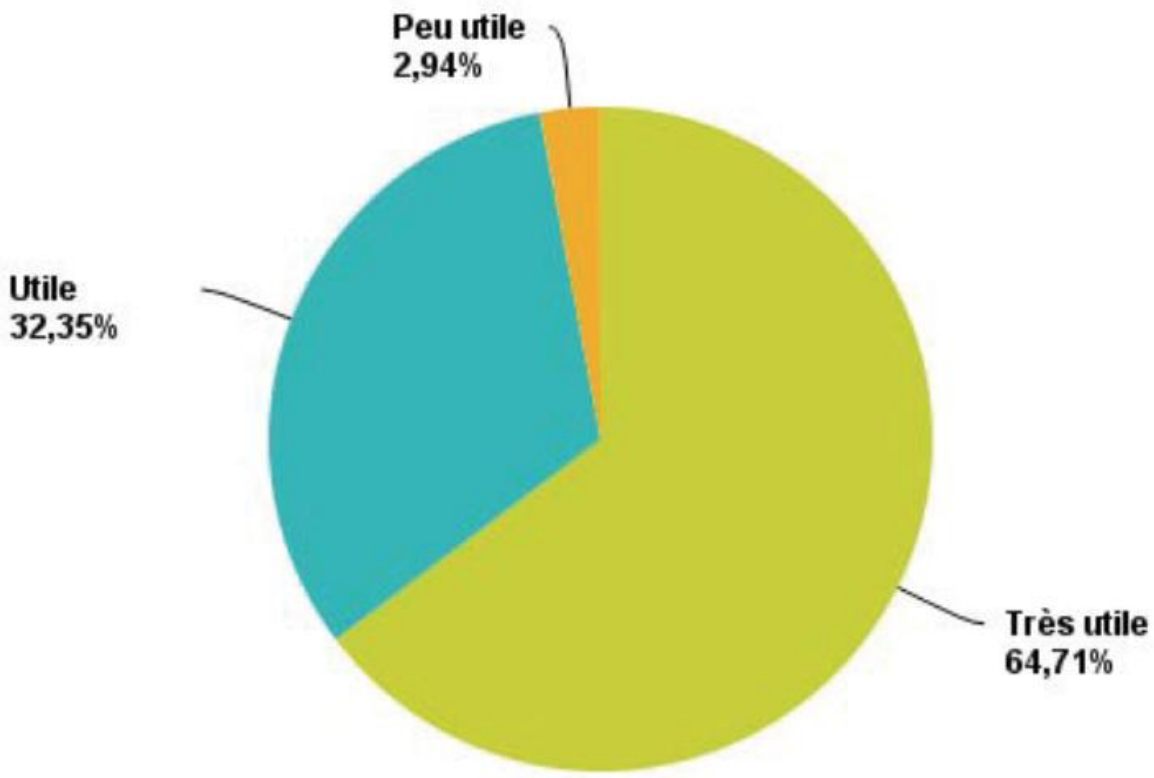
Réponses à la question "A la réception du devis vous n'avez pas lancé de souscription. Pour quelle raison ?"



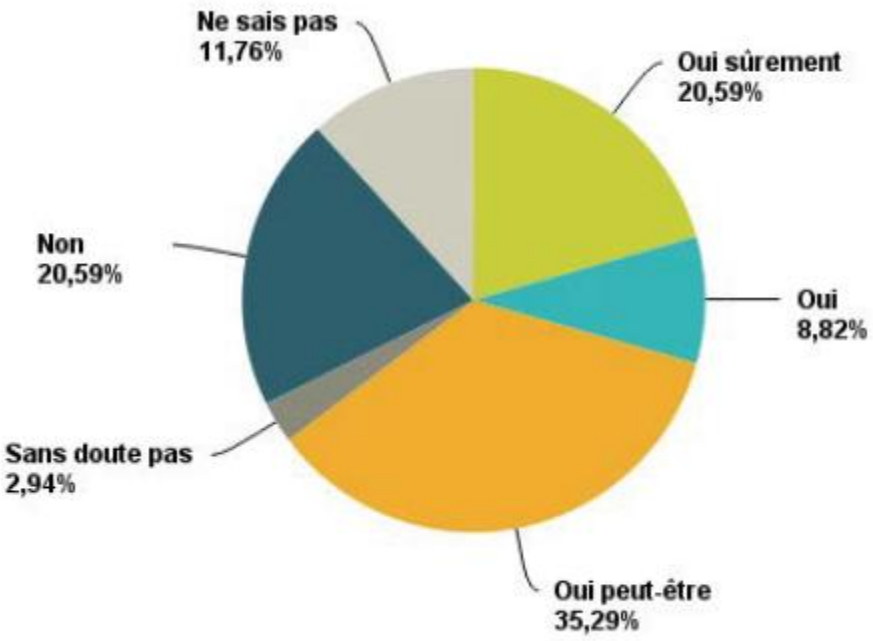
Réponses à la question “Si vous avez répondu que le coût de la prestation vous semblait trop élevé à la question précédente, pouvez-vous nous préciser ?”



Réponses à la question "Comment jugez-vous la qualité du service de Numalire ?"



Réponses à la question “Jugez-vous le service Numalire…”



Réponses à la question "Vous avez participé à la souscription pour la réédition d'un document appartenant au domaine public, sans doute parce que des liens particuliers vous relient à cette œuvre. Accepteriez-vous d'apporter vos connaissances (amatrices ou professionnelles) à la bibliothèque pour des "travaux d'intérêt commun" concernant cette œuvre ? Par exemple, qualification de l'œuvre au travers de mots clés, commentaire écrits sur l'importance de l'œuvre, participation à des opérations de correction de l'océrisation (opération qui permet de passer d'un fichier image à un fichier texte)

Enfin, il est à noté que 65 % des répondants seraient prêts à collaborer avec les bibliothèques au delà du *crowdfunding* via des travaux de *crowdsourcing*.

Annexe 4- Equations de recherche utilisée pour constituer le corpus

1- *Crowdsourcing* et bibliothèques

Equation web of Science (A1) :

(TS=librar* AND TS=crowd*) OR TS="digitization on demand" OR TS="digitisation on demand" OR TS="numérisation à la demande" OR TS=books2ebooks OR TS="ebooks on demand" OR TS="Phénix Editions" OR TS=Librissimo OR TS="Henri Le More" OR TS="Juan Pirlot de Corbion" OR TS=Youscribe OR TS="Filippo Gropallo" OR TS="Denis Maingreaud" OR TS=Yabé OR ((TS="adopter un livre" OR TS="adoptez un livre") AND (TS="bibliothèque nationale" OR TS="BnF")) OR TS="Amateur Scanning League" OR (TS=OCR AND TS=crowd*) OR TS=wikisource OR (TS=recaptcha AND TS="Google Books") OR (TS=TROVE AND TS=austral*) OR TS="Distributed Proofreaders" OR TS="Digitalkoot" OR (TS=ozalid and (TS="bibliothèque nationale" OR TS=bnf)) OR (TS=manuscript* AND TS=crowd*) OR TS="T-Pen" OR TS="Citizen Archivist Dashboard" OR TS="National Archives Transcription" OR TS="Transcribe Bentham" OR TS="1001 Stories Denmark" OR TS="steve.museum" OR TS="mechanical turk marketplace" OR TS=Mtagger OR TS=PennTags OR TS="Social OAC" OR TS="Google Image Labeler game" OR TS="herbonautes"

Refined by: Web of Science Categories=(INFORMATION SCIENCE LIBRARY SCIENCE OR COMPUTER SCIENCE INFORMATION SYSTEMS OR COMPUTER SCIENCE THEORY METHODS OR COMPUTER SCIENCE SOFTWARE ENGINEERING OR MULTIDISCIPLINARY SCIENCES OR COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE OR COMPUTER SCIENCE HARDWARE ARCHITECTURE OR COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS OR SOCIAL SCIENCES INTERDISCIPLINARY)

Timespan=All Years. Databases=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, CCR-EXPANDED, IC.

Equation ScienceDirect (B1) :

(TITLE-ABSTR-KEY(librar*) AND TITLE-ABSTR-KEY(crowd*)) OR TITLE-ABSTR-KEY(“digitization on demand”) OR TITLE-ABSTR-KEY(“digitisation on demand”) OR TITLE-ABSTR-KEY(“numérisation à la demande”) OR TITLE-ABSTR-KEY(books2ebooks) OR TITLE-ABSTR-KEY("ebooks on demand") OR TITLE-ABSTR-KEY(“Phénix Editions”) OR TITLE-ABSTR-KEY(Librissimo) OR TITLE-ABSTR-KEY(“Henri Le More”) OR TITLE-ABSTR-KEY(“Juan Pirlot de Corbion”) OR TITLE-ABSTR-KEY(Youscribe) OR TITLE-ABSTR-KEY(“Filippo Gropallo”) OR TITLE-ABSTR-KEY(“Denis Maingreaud”) OR TITLE-ABSTR-KEY(Yabé) OR ((TITLE-ABSTR-KEY(“adopter un livre”) OR TITLE-ABSTR-KEY(“adoptez un livre”)) AND (TITLE-ABSTR-KEY(“bibliothèque nationale”) OR TITLE-ABSTR-KEY(“BnF”))) OR TITLE-ABSTR-KEY(“Amateur Scanning League”) OR (TITLE-ABSTR-KEY(OCR) AND TITLE-ABSTR-KEY(crowd*)) OR TITLE-ABSTR-KEY(wikisource) OR (TITLE-ABSTR-KEY(recaptcha) AND TITLE-ABSTR-KEY(“Google Books”)) OR (TITLE-ABSTR-KEY(TROVE) AND TITLE-ABSTR-KEY(austral*)) OR TITLE-ABSTR-KEY(“Distributed Proofreaders”) OR TITLE-ABSTR-KEY(“Digitalkoot”) OR (TITLE-ABSTR-KEY(ozalid) and (TITLE-ABSTR-KEY(“bibliothèque nationale”) OR TITLE-ABSTR-KEY(bnf))) OR (TITLE-ABSTR-KEY(manuscript*) AND TITLE-ABSTR-KEY(crowd*)) OR TITLE-ABSTR-KEY(“T-Pen”) OR TITLE-ABSTR-KEY(“Citizen Archivist Dashboard”) OR TITLE-ABSTR-KEY(“National Archives Transcription”) OR TITLE-ABSTR-KEY(“Transcribe Bentham”) OR TITLE-ABSTR-KEY(“1001 Stories Denmark”) OR TITLE-ABSTR-KEY(“steve.museum”) OR TITLE-ABSTR-KEY(“mechanical turk marketplace”) OR TITLE-ABSTR-KEY(Mtagger) OR TITLE-ABSTR-KEY(PennTags) OR TITLE-ABSTR-KEY(“Social OAC”) OR TITLE-ABSTR-KEY(“Google Image Labeler game”) OR TITLE-ABSTR-KEY(“herbonautes”)

Equation Google Scholar (C1) :

Equation C1.1 : library crowdsourcing digitization

Equation C1.2 : bibliothèque crowdsourcing numérisation

Equation C1.3 : “digitization on demand” OR “digitisation on demand” OR “numérisation à la demande” OR books2ebooks OR "ebooks on demand" OR “Phénix Editions” OR Librissimo OR “Henri Le More” OR herbonautes

Equation C1.4 : “Juan Pirlot de Corbion” OR Youscribe OR “Filippo Gropallo” OR “Denis Maingreaud” OR Yabé OR “Amateur Scanning League” OR wikisource OR “Distributed Proofreaders” OR “Social OAC” OR “Google Image Labeler game”

Equation C1.5 : “Digitalkoot” OR “T-Pen” OR “Citizen Archivist Dashboard” OR “National Archives Transcription” OR “Transcribe Bentham” OR “1001 Stories Denmark” OR “steve.museum” OR “mechanical turk marketplace” OR Mtagger OR PennTags

Equation C1.6 : OCR crowdsourcing correction

Equation C1.7 : manuscripts crowdsourcing transcription

Equation C1.8 : “adoptez un livre” “bibliothèque nationale”

Equation C1.9 : recaptcha “Google Books” OCR

Equation C1.10 : TROVE australian

Equation C1.11 : ozalid BNF

2- *Print on demand*

((((digitis* OR digitiz*) AND librar*) OR (numéris* AND biblioth*)) AND (“print on demand” OR POD)) OR “Amazon BookSurge” OR “lulu.com” OR “Espresso Book Machine” OR “ondemandbooks.com”

Equation web of Science (A2) :

((((TS=digitis* OR TS=digitiz*) AND TS=librar*) OR (TS=numéris* AND TS=biblioth*)) AND (TS=”print on demand” OR TS=POD)) OR TS=“Amazon BookSurge” OR TS=“lulu.com” OR TS=“Espresso Book Machine” OR TS=“ondemandbooks.com”

Equation Google Scholar (C2) :

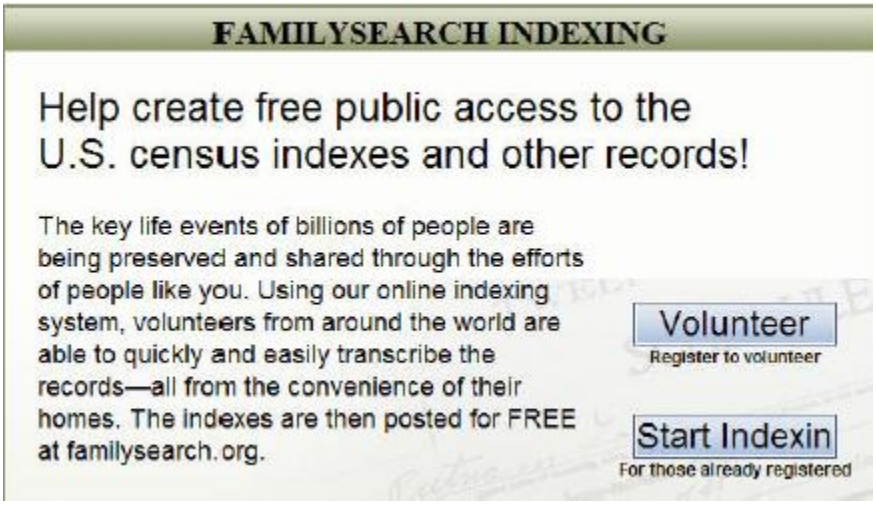
Equation C2.1 :"print on demand" "digital library" digitization

Equation C2.2 : “Print on demand” numérisation bibliothèque

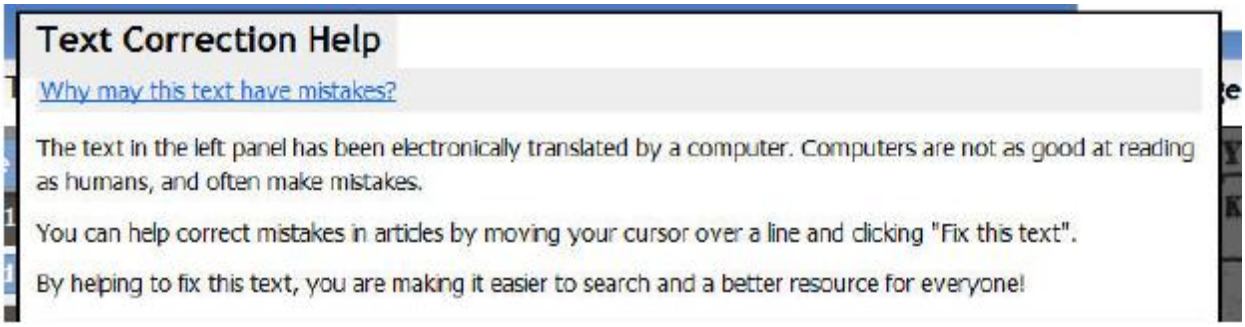
Equation C2.3 : “amazon booksurge” OR “espresso book machine” OR “ondemandbooks.com”

Annexe 5- Illustration de la manière dont communiquent les projets

Voici des exemples de communication sur divers sites web.



Exemple de texte pour inviter les internautes à s'engager pour le projet Familysearch indexing. D'après (Holley, 2009)



Exemple de texte pour inviter les internautes à s'engager pour le projet Australian Newspapers. D'après (Holley, 2009)

We have **458,832** pages of documents. **24,524** of you have reviewed **215,305** of them. Only **243,527** to go...

Start reviewing

Please read our privacy policy to find out how we use your data. You must also read our terms of service. By reviewing pages, you are agreeing that you have read the terms of service, and that you agree to them.

Thanks everyone for your valiant efforts so far.

Des Browne MP's expenses

Start reviewing pages

Labour MP for Kilmarnock & Loudoun

Guardian politics profile

Documents

Document	Total reviewed	Progress
Additional Costs Allowance 2004/05	12 of 12 pages reviewed	
Incidental Expenses Provision 2004/05	19 of 148 pages reviewed	
Additional Costs Allowance 2005/06	16 of 16 pages reviewed	
Incidental Expenses Provision 2005/06	19 of 222 pages reviewed	
Additional Costs Allowance 2006/07	15 of 15 pages reviewed	
Incidental Expenses Provision 2006/07	29 of 195 pages reviewed	
Additional Costs Allowance 2007/08	24 of 24 pages reviewed	
Communication Allowance 2007/08	28 of 37 pages reviewed	
Incidental Expenses Provision 2007/08	102 of 182 pages reviewed	

Recent investigations

- eatmypoverty
- gdw
- anon-16048
- anon-15556
- chopper

Exemples de textes pour inviter les internautes à s'engager pour le projet The Guardian MP's expenses. D'après (Holley, 2009)

Current Progress

16,644 in Completed. XML RSS

These books have been processed through our site and posted to the Project Gutenberg archive.

2,509 In Progress. XML RSS

These books are undergoing their final checks before being assembled into a completed e-book.

793 in Proofreading. XML RSS

These books are currently being processed through our site; sign in and start helping!

Our community of proofreaders, project managers, developers, etc. is composed entirely of volunteers.
658 active users in the past twenty-four hours.
1,425 active users in the past 7 days.
2,991 active users in the past 30 days.

Completed Gold E-Texts

Gold | Silver | Bronze

Below is the list of Gold e-texts that have been produced on this site. Gold e-texts are books that have passed through all phases of proofreading, formatting, and post-processing. They have been submitted to Project Gutenberg and are now available for your enjoyment and download. These e-texts are the product of hundreds of hours of labor donated by all of our volunteers. The list is sorted with the most recently submitted e-texts at the top. You can sort them based upon your own preferences by clicking below. Enjoy!!

Exemples de textes pour inviter les internautes à s'engager pour le projet Distributed Proofreader. D'après (Holley, 2009)

English Wikipedia right now

Wikipedia is running MediaWiki version 1.16alpha-wmf(r58524).

It has 3,087,147 articles, and 18,492,486 pages in total.

There have been 345,137,431 edits.

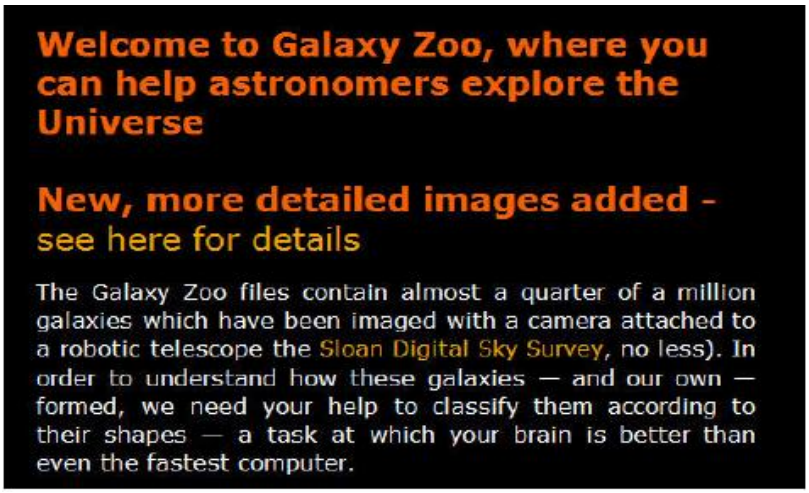
There are 867,587 uploaded files.

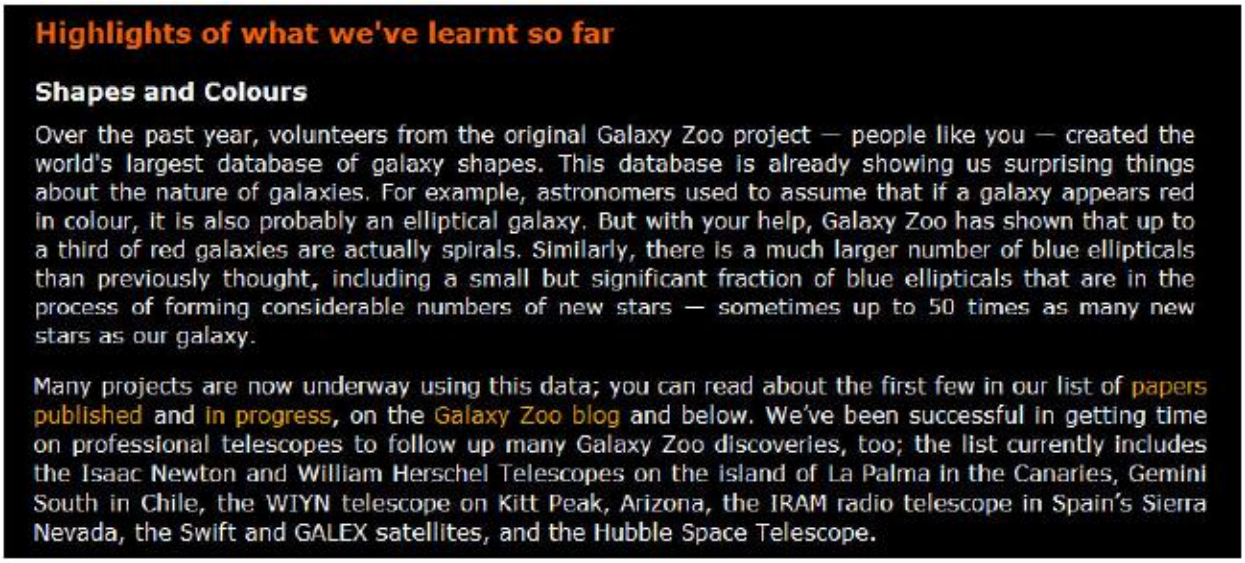
There are 10,929,377 registered users, including 1,693 administrators.

This information is correct as of 06:27 on November 7, 2009.

Update

Exemple de texte pour inviter les internautes à s'engager pour le projet Wikipedia. D'après (Holley, 2009)





Exemples de textes pour inviter les internautes à s'engager pour le projet Galaxy Zoo. D'après (Holley, 2009)

Annexe 6- Articles publiés dans le cadre de la thèse

Bien avant le commencement de ce travail de recherche autour du *crowdsourcing* appliqué à la numérisation des bibliothèques, un approfondissement de connaissances, d'expériences et de compétences avait été développé dans le domaine de la numérisation des bibliothèques dans le cadre d'un poste de chef de projets de numérisation à la Bibliothèque Sainte-Geneviève et initiateur d'un projet de bibliothèque numérique mutualisée et *crowdsourcing* pour le PRES Sorbonne Paris-Cité. Cette expérience à donné lieu aux publications suivantes :

- **Andro, M., Chaigne, M., Smith, F. (2012). "Valoriser une bibliothèque numérique par des choix de référencement et diffusion. L'expérience de la Bibliothèque Sainte-Geneviève". Les Cahiers du numérique 9(3): 75-90.**
- **Andro, M. (2013) "Diffusion, référencement et valorisation des factums numérisés sur le web". Revue du Centre Michel de l'Hospital 3: 136-138.**
- **Andro, M. (2013). "L'expérience de numérisation à la Bibliothèque Sainte-Geneviève" in Manuel de constitution de bibliothèques numériques. Paris, Éditions du Cercle de La Librairie. ISBN 978-2765414131. p. 66-68.**
- **Andro, M. (2011). "Genèse d'un projet de plateforme mutualisée pour la diffusion des documents numérisés". Arabesque 64: 8**

Ayant constaté, en particulier, qu'il existait peu d'études au sujet des solutions concrètes afin de diffuser des documents numérisés et que cela pouvait expliquer, en partie, qu'une proportion importante d'entre eux n'était jamais mise en ligne, la publication d'un livre, aux éditions de l'ADBS, au sujet des solutions de diffusion des documents numérisés, des plateformes et des logiciels pour développer des bibliothèques numériques avait été initiée :

- **Andro, M., Asselin, E., Maisonneuve, M. (2012). "Bibliothèques numériques : logiciels et plateformes". Paris, ADBS. 351 p. ISBN 978-2-84365-140-3**

Une version plus courte de cette étude a ensuite été publiée en français dans une revue nationale puis en anglais dans une revue internationale :

- **Maisonneuve, M., Andro, M., Asselin, E. (2012). "Méthodes, techniques et outils. L'offre de logiciels pour les bibliothèques numériques". Documentaliste, Sciences de l'Information 49(2): 10-11.**
- **Andro, M., Asselin, E., Maisonneuve, M. (2012). "Digital libraries : comparison of 10 software". Library Collections, Acquisitions, and Technical Services 36(3-4): 79-83.**

Dans le prolongement de cette étude, une autre étude au sujet de la mutualisation de la diffusion via des plateformes et des statistiques de consultation des bibliothèques numériques a été proposée :

- **Andro, M., Tröger, G. (2013) Statistiques et visibilité des bibliothèques numériques : quelles stratégies de diffusion ?**

Remarquée par les éditeurs de la revue Archimag, cette étude a ensuite été actualisée et revue pour être publiée dans leur revue :

- **Andro, M. (2015). Bibliothèques numériques : fréquentation et prospective. Archimag guide pratique 52, "Bibliothèques, les nouveaux modèles", p. 22-25**

Dans le cadre de formations au sujet de la numérisation données aux journées « Quoi de neuf en bibliothèques ? » à l'Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB), l'article suivant a été publié :

- **Andro, M. (2011) "Actualité de la numérisation". Supplément au Bulletin des Bibliothèques de France: 27-29.**

Dans le cadre de la thèse, une première étape d'autoformation, d'état de l'art et de synthèse a été réalisée. Au cours de cette étape, la traduction suivante a été publiée :

- **Andro, M. (2013) "Faire appel à la charité" : Trove, les journaux australiens et la foule des internautes. Traduction de : Ayres, M.-L. (2013) Singing for their supper': Trove, Australian newspapers, and the crowd. Paper presented at: IFLA World Library and Information Congress, 17 - 23 August 2013, Singapore.**

- Et une présentation en anglais dans le cadre du projet européen Ebooks on Demand a été proposée :

- **Andro, M. (2014). Crowdsourcing and digitization: Presentation for the Ebooks on Demand Conference 2014 (Innsbruck University, April 11th 2014). HAL**

Le premier chapitre, conceptuelle, de la thèse est résumé dans l'article suivant présenté au 17e colloque international sur le document électronique au Maroc

- **Andro, M., Saleh, I. (2014). Bibliothèques numériques et crowdsourcing : une synthèse de la littérature académique et professionnelle internationale sur le sujet. in ZREIK Khaldoun, AZEMARD Ghislaine, CHAUDIRON Stéphane, DARQUIE Gaétan. Livre post-numérique : historique, mutations et perspectives. Actes du 17e colloque international sur le document électronique (CiDE.17), 2014, 152 p.**

Le second chapitre de la thèse, comprenant un panorama des projets et un état de l'art a fait l'objet d'un article sur la correction participative de l'OCR, d'un autre sur l'impression à la demande et d'un dernier sur la *gamification* :

- **Andro, M., Saleh, I. (2015). Bibliothèques numériques et gamification : panorama et état de l'art. I2D - Information, données & documents. (sous presse)**

- **Andro, M., Saleh, I. (2015). La correction participative de l'OCR par crowdsourcing au profit des bibliothèques numériques. Bulletin des Bibliothèques de France, Contribution du 16 juin 2015.**

- **Andro, M., Klopp, S. (2015). L'impression à la demande et les bibliothèques. Bulletin des Bibliothèques de France, contribution du 13 février 2015. 7 p.**

Le chapitre des analyses du point de vue des sciences de l'information a fait l'objet de l'article suivant :

- **Andro, M., Saleh, I. (2016). Le crowdsourcing appliqué aux bibliothèques numériques. Bibliothèque(s)**

Le chapitre expérimental, de la thèse a fait l'objet d'un article faisant état des résultats de l'expérimentation de numérisation à la demande Numalire et d'une idée d'expérimentation :

- **Andro, M., Rivière, P., Dupuy-Olivier, A., Gropallo, F., Maingreaud, D. (2014). Numalire, une expérimentation de numérisation à la demande du patrimoine conservé par les bibliothèques sous la forme de financements participatifs (crowdfunding). Bulletin des Bibliothèques de France, contribution du 2 octobre 2014, 9 p.**
- **Andro, M. (2015). Le crowdfunding pour financer l'accès à la science ? Publier la science, 7, p. 3**

Enfin, une participation à un groupe de travail sur les sciences participatives autour du Président de l'Inra missionné par le gouvernement pour remettre un rapport et des préconisations sur le sujet peut être retrouvée dans le rapport suivant :

- **Houllier, F., Merilhou-Goudard, J.-B., Andro, M., Charbonnel, F., Cointet, J.-P., Frey-Klett, P., Joly, P.-B, Leiser, H., Mambrini-Doudet, M., Hologne, O., Launay, J.-F., Le Gall, O., , Masson, J., Morcrette, N., Pujol, J.-L., Roturier, C. (2016). *Les sciences participatives en France.* Rapport au Ministre de l'Enseignement Supérieur (3 livrets, 41, 28, 24 p.)**

Et un résumé de l'ensemble de la thèse sera publiée prochainement sous la forme d'une *review* et traduite en anglais.

D'autres articles ont été publiés sur d'autres sujets, en dehors de la thèse et que nous n'évoquons donc pas ici.

Enfin, des interventions à des cours et à des séminaires ont été assurées en relation avec le sujet de thèse :

- Intervention au Master Étude Numériques et Innovation (Université Paris Est) le 15 novembre 2015
- Intervention à l'école Inra : "Évolutions de la recherche et impacts pour l'IST" à propos des sciences citoyennes le mercredi 27 mai 2015

- Formation sur le *crowdfunding*, la webométrie, la culturomique, le *textmining* et l'Open Access dans le cadre du Master Technologies de l'hypermédia (Paris VIII), le 23 octobre 2015
- Formation sur la numérisation, les bibliothèques numériques et le *crowdsourcing* dans le cadre de la licence Licence Pro Créations et Développements Numériques en Ligne (Paris VIII), le 19 janvier 2015
- Intervention à la table ronde du forum GFII 2014 "Le social data est-il l'avenir de l'open data ?" (Lundi 8 décembre 2014)
- Intervention au 17è Colloque International sur le Document Numérique, CIDE 17 sur le *crowdsourcing* et les bibliothèques numériques (Fès, Maroc, 2014).
- Intervention sur le *crowdsourcing* dans le cadre de l'Open Access Week à l'Ecole des Ponts ParisTech, le 21 octobre 2014 matin.
- Intervention en anglais à la conférence du projet européen Ebooks on Demand (Innsbruck University, 10-11 avril 2014). Crowdsourcing and digitization (30 min)
- Intervention au Séminaire de l'axe "traces digitales" (groupe Cortext, Institut Francilien Recherche Innovation Société) : *Crowdsourcing* et numérisation. 6 février 2014 matin
- Compte-rendu d'entretiens avec les bibliothèques participant au projet Numalire aux journées Numalire du 18 juin 1014 à, l'académie de médecine.
- Interventions à l'Ecole Normale Supérieur des Sciences de l'Information et des Bibliothèques (ENSSIB) : participation à aux journées 2011, 2012, 2013, 2014 et 2015 "Quoi de neuf en bibliothèque ?" (pour les Directeurs de Bibliothèques), à une journée de formation des élèves conservateurs sur les solutions logicielles et un cas pratique (7 heures), au module "numérisation et constitution de bibliothèques numériques" (formation continue des professionnels) en 2010 et 2013 et à une table ronde sur la bibliothèque numérique (formation des Bibliothécaires)
- "Réussir la mise en ligne de sa bibliothèque numérique", 5 à 7 ADBS, 24 mai 2012 (avec Marc Maisonneuve)

- Présentation de la sortie du livre ADBS sur les bibliothèques numériques avec Marc Maisonneuve au salon Documation le 21 mars 2012 (60 personnes)
- Formation donnée avec Marc Maisonneuve à l'Unité Régionale de Formation à l'Information Scientifique et Technique (Urfist) de Paris : "les gestionnaires de bibliothèques numériques" le 22 octobre 2012 de 9 h 30 à 12 h 30
- Participation à la journée d'étude "découverte et valorisation d'une source juridique méconnue : le factum ou mémoire judiciaire" organisée par la faculté de droit de Clermont-Ferrand (7 juin 2012)
- Présentation du projet de plate-forme mutualisée du PRES Sorbonne Paris-Cité aux journées ABES 2011
- Présentation d'1 h 30 à l'Université Numérique Vivaldi Paris Ile de France 2011 : "Les programmes de numérisation des bibliothèques d'Île-de-France".
- Participation à la journée d'étude Mediadix "De Gallica à Google : la dématérialisation des collections et des accès à la croisée des chemins" (150 personnes, 8 octobre 2010)
- Intervention au Forum 2010 de l'impression numérique de livres (200 cadres et directeurs de l'édition et de l'imprimerie)

Figures et des tableaux contenus dans la thèse

Figures

Figure 1. Nuage de mots portant sur le contenu de la thèse et réalisé avec tagxedo.com
Figure 2. Analyse temporelle de l'évolution du nombre de publications dans le corpus bibliométrique
Figure 3. Nombre de publications par an dans la bibliographie de la thèse
Figure 4. Répartition du corpus selon les types de documents
Figure 5. Poids des pays des auteurs des articles dans la bibliographie de la thèse
Figure 6. Origine géographique des visites sur le site bibliotheque-numerique.fr d'après Google Analytics
Figure 7. Capture d'écran du site www.bibliotheque-numerique.fr
Figure 8. L'œuvre d'art Ten Thousands Cents
Figure 9. Œuvre d'art de juxtaposition de moutons
Figure 10. Épée du 13è siècle dont la photographie a été publiée par la British Library
Figure 11. Evolution des dons de particulier à Wikimedia France entre janvier 2009 et octobre 2012
Figure 12. Evolution du nombre de recherche du mot "*crowdsourcing*" dans Google selon les pays d'après Google Trends
Figure 13. Pays représentés dans l'enquête conduite par l'OCLC à propos des métadonnées sociales (d'après Smith-Yoshimura, 2011)
Figure 14. Evolution du nombre de publications indexés par Google Scholar sur le *crowdsourcing* appliqué à la numérisation des bibliothèques
Figure 15. Relations entre *human computation*, *collective intelligence* et *crowdsourcing*, d'après (Harris, 2013)
Figure 16. Positionnement du *crowdsourcing* parmi les domaines voisins (d'après Schenk, 2010)
Figure 17. Première forme de *crowdfunding* ? D'après http://gallica.bnf.fr/ark:/12148/btv1b8509563b
Figure 18. Pourcentage de wikipédiens par date de naissance, d'après Wikipedia

Figure 19. Capture d'écran d'un texte en OCR brute
Figure 20. Capture d'écran d'un journal numérisé et de son OCR
Figure 21. Localisation des membres du réseau Ebooks on Demand au 8 juillet 2014
Figure 22. Extrait d'un rapport d'activités EOD (d'après Klopp, 2014)
Figure 23. Commandes par tranches de prix pendant la période 2009-2001 à la Bibliothèque Nationale de Slovénie (d'après Brumen, 2012)
Figure 24. La forme sous laquelle les usagers préfèrent consulter les documents (d'après l'enquête rapportée par Mühlberger, 2009)
Figure 25. la perception positive / négative selon les prix et les délais (d'après l'enquête rapportée par Mühlberger, 2009)
Figure 26. Centres d'intérêts des usagers d'après (Gstrein, 2011)
Figure 27. Raisons pour lesquelles les usagers ont commandé, d'après (Gstrein, 2011)
Figure 28. Statistiques des commandes EOD de la Bibliothèque InterUniversitaire de Santé d'après (Klopp, 2014)
Figure 29. Photographie d'une Espresso Book Machine (d'après http://ondemandbooks.com)
Figure 30. Répartition des EBM dans le monde d'après http://www.ondemandbooks.com/ebm_locations.php le 9 juillet 2014
Figure 31. Évolution du nombre de corrections de lignes sur TROVE d'après les statistiques obtenues sur le site lui-même (http://trove.nla.gov.au/system/stats?env=prod)
Figure 32. Capture d'écran de TROVE
Figure 33. Budget du projet Transcribe Bentham d'après (Causer, 2012)
Figure 34. Évolution du nombre de comptes, de manuscrits transcrits et complétés entre le 8 septembre 2010 et le 8 mars 2011, d'après (Causer, 2012)
Figure 35. Boutons utilisés par Transcribe Bentham
Figure 36. Interface de transcription de Transcribe Bentham d'après (Brokfeld, 2012)

Figure 37. Diagramme représentant comment les internautes ont découvert le projet Transcribe Bantham (d'après Causer, 2012)
Figure 38. Diagramme représentant la répartition des contributeurs de Transcribe Bentham selon leur âge (d'après Causer, 2012)
Figure 39. Motivations des volontaires du projet Transcribe Bentham d'après (Causer, 2012)
Figure 40. Capture d'écran du jeu Mole Hunt
Figure 41. Capture d'écran du jeu Mole Bridge
Figure 42. Proportion du travail réalisé par les 1 %, 10 %, 25 %, des meilleurs contributeurs (d'après Chrons, 2011)
Figure 43. Schéma expliquant comment fonctionne reCAPTCHA d'après le site Google.com
Figure 44. Autre schéma expliquant comment fonctionne reCAPTCHA, d'après (Ipeirotis, 2011)
Figure 45. Le joueur d'échec turc. « Tuerkischer schachspieler windisch4 » par Karl Gottlieb von Windisch. 1783.
Figure 46. Nombre de HITS en novembre 2013 d'après le Mechanical Turk tracker
Figure 47. Répartition des travailleurs indiens et américains sur l'AMT le sexe (d'après Ipeirotis, 2010)
Figure 48. Années de naissance des travailleurs sur l'AMT (d'après Ipeirotis, 2010)
Figure 49. Niveau scolaire des travailleurs sur l'AMT (d'après Ipeirotis, 2010)
Figure 50. Temps moyen consacré à l'AMT (d'après Ipeirotis, 2010)
Figure 51. Revenus moyens tirés de l'AMT (d'après Ipeirotis, 2010)
Figure 52. Nombre de travailleurs déclarant que l'AMT est leur première source de revenus (d'après Ipeirotis, 2010)
Figure 53. Types de motivations en fonction de la plus ou moins grande assiduité des travailleurs sur la plateforme AMT d'après (Kaufmann, 2011)
Figure 54. Capture d'écran du jeu Art Collector 1er round, d'après (Paraschakis, 2013)
Figure 55. Capture d'écran du jeu Art Collector Round 2, choix d'une pièce, d'après (Paraschakis, 2013)

Figure 56. Capture d'écran du jeu Art Collector. Round 2, essayer de gagner une œuvre, d'après (Paraschakis, 2013)
Figure 57. Genre et age des joueurs de Art Collector d'après (Paraschakis, 2013)
Figure 58. Nombre de corrections sur TROVE entre 2008 et 2012, d'après (Hagon, 2013)
Figure 59. L'évolution du nombre contenus comparée à celle du nombre de corrections sur TROVE, d'après (Hagon, 2013)
Figure 60. Part des généalogistes parmi les contributeurs d'après une enquête CDNC / Cambridge Public Library
Figure 61. Répartition des bénévoles par classes d'âge d'après une enquête CDNC / Cambridge Public Library
Figure 62. Les types de documents diffusés sur Trove comparés aux types de documents qui y sont corrigés, d'après (Hagon, 2013)
Figure 63. Les types de documents les plus corrigés sur Trove, d'après (Hagon, 2013)
Figure 64. Le top 50 des plus gros contributeurs de Wikipedia
Figure 65. Classement des contributeurs selon le nombre de lignes corrigés pour les projets TROVE et CDNC d'après (Zarndt, 2014)
Figure 66. Part du travail accompli par chaque contributeur du projet Old Weather proposant de transcrire des observations météorologiques (d'après Brumfield, 2013)
Figure 67. Taxonomie du *crowdsourcing*, d'après (Harris, 2013)
Figure 68. La taxonomie des 4C du *crowdsourcing* d'après (Renault, 2014 bis)
Figure 69. Evolution temporelle depuis 2011 et prévision du marché futur de la *gamification* (d'après Ollikainen, 2013)
Figure 70. *Serious games* et *gamification* d'après (Deterding, 2011)
Figure 71. Capture d'écran de la communication de What's on the menu?
Figure 72. Taxonomie des motivations des bénévoles dans un projet de *crowdsourcing*
Figure 73. Pyramide des besoin de Maslow (d'après http://fr.wikipedia.org/wiki/Pyramide_des_besoins)

Figure 74. Diagramme illustrant qu'une poignée d'internautes est à l'origine de la majorité des contributions (Brumfield, 2013)
Figure 75. Répartition des activités du personnel manageant des projets de *crowdsourcing* d'après (Smith-Yoshimura, 2011)
Figure 76. Le temps de travail du personnel des projets de *crowdsourcing* d'après (Smith-Yoshimura, 2011)
Figure 77. Fréquence avec laquelle les sites mettent en ligne du nouveau contenu d'après (Smith-Yoshimura, 2011)
Figure 78. Les critères de succès d'après (Smith-Yoshimura, 2011)
Figure 79. Nombre de visiteurs uniques par mois pour les projets de *crowdsourcing* d'après (Smith-Yoshimura, 2011)
Figure 80. Nombre de contributeurs par mois pour les institutions culturelles d'après (Smith-Yoshimura, 2011)
Figure 81. Page d'accueil du site www.numalire.com
Figure 82. Ouvrage dont la numérisation a été financée sur le site www.numalire.com
Figure 83. Statistiques journalières de consultation du site numalire.com entre le 1er octobre 2013 et le 31 mai 2014 d'après Google Analytics
Figure 84. Statistiques mensuelles de consultation du site numalire.com entre le 1er octobre 2013 et le 31 mai 2014 d'après Google Analytics
Figure 85. Evolution du nombre de sessions sur le site numalire.com ayant pour origine la saisie du mot numalire dans Google d'après Google Analytics
Figure 86. Evolution du nombre de sessions sur le site numalire.com ayant pour origine le clic sur un lien vers numalire.com depuis un site web d'après Google Analytics
Figure 87. Evolution du nombre de sessions sur le site numalire.com ayant pour origine les réseaux sociaux d'après Google Analytics
Figure 88. Âge des visiteurs du site numalire.com d'après Google Analytics
Figure 89. Genre des visiteurs du site numalire.com d'après Google Analytics
Figure 90. Origine géographique des visiteurs du site numalire.com d'après Google Analytics

Figure 91. Origines régionales des visiteurs français du site numalire.com d’après Google Analytics
Figure 92. Capture d'écran d'un bouton Ebooks on Demand
Figure 93. Capture d'écran de la manière dont Google affiche une notice Numalire

Tableaux

Tableau 1. Sources utilisées pour constituer le corpus analysé dans la thèse
Tableau 2. Tarifs appliqués par les bibliothèques du réseau EOD (d’après http://books2ebooks.eu/fr/prices le 9 juillet 2014)
Tableau 3. Statistiques collectées dans la littérature à propos du projet TROVE
Tableau 4. Statistiques collectées dans la littérature à propos du projet Transcribe Bentham
Tableau 5. Statistiques collectées dans la littérature à propos du projet Digitalkoot
Tableau 6. Statistiques collectées dans la littérature à propos du projet reCAPTCHA
Tableau 7. Statistiques collectées dans la littérature à propos du projet Amazon Mechanical Turk Marketplace
Tableau 8. Coûts comparés entre une correction d'OCR via l'AMT et via un prestataire
Tableau 9. Statistiques collectées dans la littérature à propos du projet Flickr The Commons
Tableau 10. Tarifs pratiqués par diverses institutions pratiquant la numérisation et l'impression à la demande
Tableau 11. Estimation du coût non dépensé en prestation de correction d'OCR grâce au recours au *crowdsourcing*
Tableau 12. Modèle de participations du public inspiré de (Bonney, 2009)
Tableau 13. Activités d'un projet de numérisation croisés avec les types de *crowdsourcing*
Tableau 14. Types existants de *crowdsourcing* appliqués à la numérisation
Tableau 15 des types restant à inventer de *crowdsourcing* appliqués à la numérisation

Tableau 16. Taxonomie du *crowdsourcing* appliqué à la numérisation
Tableau 17. Données collectées dans la littérature à propos de la sociologie des contributeurs des différents projets
Tableau 18. La répartition du temps de travail du personnel des projets de *crowdsourcing* en fonction des activités et des missions d'après (Smith-Yoshimura, 2011)
Tableau 19. Utilisation faite par les institutions culturelles des métadonnées socialesd'après l'étude OCLC (Smith-Yoshimura, 2011)
Tableau 20. Statistiques avant et après *crowdsourcing* pour la California Digital Newspaper Collection (d'après Geiger, 2012)

Bibliographie

Au total, quelques 833 documents ont été consultés. Seule une partie d'entre eux, ceux qui ont été directement utilisés et sont cités dans le corps de la thèse apparaît dans cette bibliographie.

1. Acar, O. A., Van den Ende, J. (2011). *Motivation, reward size and contribution in idea crowdsourcing*. 29 p.
2. Akila, G., El-Menisy, M., Khaled, O., Sharaf, N., Tarhony, N., Abdennadher, S. (2015). *Kalema: Digitizing Arabic Content for Accessibility Purposes Using Crowdsourcing*. CICLing 2015, Part II, LNCS 9042, pp. 655–662.
3. Alabau, V., Leiva, L. A. (2012). *Transcribing Handwritten Text Images with a Word Soup Game*. CHI'12, May 5–10, 2012, Austin, Texas, USA. 6 p.
4. Alam, S. L. Campbell, J. (2012). *Crowdsourcing motivations in a not-for-profit GLAM context : the Australian newspapers digitisation program. In:* Proceedings of the 23rd Australasian Conference on Information Systems 2012, ACIS, [Geelong, Vic.], pp. 1-11.
5. Alam, S. L., Campbell, J. (2013). *Dynamic Changes in Organizational Motivations to Crowdsourcing for GLAMs Completed Research Paper*. Thirty Fourth International Conference on Information Systems, Milan 2013, 17 p.
6. Alam, S. L., Campbell, J. (2013 bis). *A conceptual framework of influences on a non-profit GLAM crowdsourcing initiative: A socio-technical.* 24th Australasian Conference on Information Systems Socio-technical model of crowdsourcing influences, 4-6 Dec 2013, Melbourne.
7. Alcalá Ponce de León, M. (2015). *Crowdsourcing in the memory institutions : mass transcriptions*. BiD: textos universitaris de biblioteconomia i documentació, 35, [22] p.
8. AlRouqi H., Al-Khalifa H. S. (2014). *Making Arabic PDF Books Accessible Using Gamification*. In Proceeding W4A '14 Proceedings of the 11th Web for All Conference. Article No. 28. 4 p.
9. Anderson, R. (2010). *The Espresso Book Machine:The Marriott Library Experience*. Serials, 23(1):39-42.

10. Andro, M., Saleh, I. (2014, 1). Bibliothèques numériques et crowdsourcing : une synthèse de la littérature académique et professionnelle internationale sur le sujet. in ZREIK Khaldoun, AZEMARD Ghislaine, CHAUDIRON Stéphane, DARQUIE Gaétan. Livre post-numérique : historique, mutations et perspectives. Actes du 17e colloque international sur le document électronique (CiDE.17), 2014, 152 p.
11. Andro, M., Rivière, P., Dupuy-Olivier, A., Gropallo, F., Maingreaud, D. (2014, 2). Numalire, une expérimentation de numérisation à la demande du patrimoine conservé par les bibliothèques sous la forme de financements participatifs (crowdfunding). Bulletin des Bibliothèques de France, contribution du 2 octobre 2014, 9 p.
12. Andro, M., Saleh, I. (2015, 1). Bibliothèques numériques et gamification : panorama et état de l'art. I2D - Information, données & documents. (sous presse)
13. Andro, M., Saleh, I. (2015, 2). La correction participative de l'OCR par crowdsourcing au profit des bibliothèques numériques. Bulletin des Bibliothèques de France, Contribution du 16 juin 2015.
14. Andro, M., Klopp, S. (2015, 3). L'impression à la demande et les bibliothèques. Bulletin des Bibliothèques de France, contribution du 13 février 2015. 7 p.
15. Armstrong, T. K. (2010). *Rich texts: Wikisource as an open access repository for law and the humanities*. University of Cincinnati College of Law, Public Law & Legal Theory Research Paper Series, 10-09, 11 p.
16. Arlitsch, K. (2011). *The Espresso Book Machine: a change agent for libraries*, Library Hi Tech, Vol. 29, Issue 1, p. 62 - 72.
17. Askin, N. (2015). *Collaboration and Crowdsourcing: the Future of LAM Convergence*. The UBC School of Library, Archival and Information Studies Student Journal, No 1, [12] p.
18. Ayres, M.-L. (2013). *'Singing for their supper': Trove, Australian newspapers, and the crowd*. IFLA World Library and Information Congress. Singapore. 9 p.
19. Bainbridge, D., Twidale, M. B., Nichols, D. M. (2012). *Interactive context-aware user-driven metadata correction in digital libraries*. International Journal on Digital Libraries, 13, pp. 17-32.

20. Bar-Ilan, J., Shoham, S., Idan, A., Miller, Y, Shachak, A. (2008) *Structured versus unstructured tagging: a case study*, Online Information Review, Vol. 32 Iss: 5, pp.635 - 647
21. Barbrook, R. (2000). *Cyber-communism: how the americans are superseding capitalisme in cyberspace*. Science As Culture, [13] p.
22. Bartlett, J. A. (2014), "*Internet Reviews: Crowdsourcing in Libraries and Archives*". Kentucky Libraries, voume. 78, number 2, p. 6-8.
23. Bauer, A. (2010). *Sciences participatives et biodiversité : implication du public, portée éducative et pratiques pédagogiques associées*. Les livrets de l'Ifrée n°2. 107 p.
24. Bauwens, M. (2015). *Sauver le monde : vers une économie post-capitaliste avec le peer-to-peer*. Les liens qui libèrent, 265 p.
25. Benyayer, L.-D. (2014). *Open Models : les business models de l'économie ouverte*. 226 p.
26. Beuth Hochschule für Technik, Wikimedia Deutschland (2014). *Charting diversity: working together towards diversity in Wikipedia*, ISBN 978-3-9816799-0-8. 21 p.
27. Biella, D., Sacher, D, Weyers, B, Luther, W, Baloian, N, Schreck, T. (2015). *Crowdsourcing and Knowledge Co-creation in Virtual Museums*. Conference Name Proc. International Conference on Collaboration and Technology (CRIWG), 18 p.
28. Birchall, D., Henson, M., Burch, A., Evans, D., Haley Goldman, K. (2012). *Levelling Up: Towards Best Practice in Evaluating Museum Games*. 11 p.
29. Blasselle, B. (1997). *Le livre à la carte*. Bulletin des Bibliothèques de France, 6, pp. 29-29
30. Blasco, A., Boudreau, K. J., Lakhani, K. R., Menietti, M., Riedl, C. (2013). *Do Crowds have the Wisdom to Self-Organize?* 7 p.
31. Blummer, B. (2006). *Opportunities for Libraries with Print-on-demand Publishing*. Journal of Access Services 3, no. 2 : 41-54.
32. Boeuf, G., Allain, Y.-M., Bouvier, M. (2012). *L'apport des sciences participatives dans la connaissance de la bioviversité : rapport remis à la Ministre de l'Ecologie*. 29 p.

33. Bonnefont, A., Giraud, M. (2000). *Réflexion sur le lien entre achat impulsif et modèles de communication*. 23 p.
34. Bonney, R., Ballard, H., Jordan, R., McCallie, E., Phillips, T., Shirk, J., and Wilderman, C. C. (2009). *Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. A CAISE Inquiry Group Report.* Washington, D.C.: Center for Advancement of Informal Science Education (CAISE).58 p.
35. Bouyé, E. (2012). *Le web collaboratif dans les services d'archives publics : un pari sur l'intelligence et la motivation des publics*. 12 p.
36. Brabham, D. C. (2010). *Moving the crowd at threadless*. Information, Communication & Society, 13: 8, 1122-1145.
37. Brabham, D. C. (2012). *The myth of amateur crowds: A critical discourse analysis of crowdsourcing coverage*. Information, Communication & Society 15 (3), 394-410
38. Breton, E. (2014). *Co-construire les collections avec les usagers*. Mémoire de Conservateur de Bibliothèques. 89 p.
39. Breton, E. (2015). *Penser les collections avec les usagers, les bibliothèques à l'heure de la co-construction*. Ar(abes)ques 80, p. 22-23.
40. Brigham, T. J. (2015). *An Introduction to Gamification: Adding Game Elements for Engagement*, Medical Reference Services Quarterly, 34:4, 471-480
41. Brokfeld, J. (2012). *Die digitale Edition der „preußischen Zeitungsberichte": Evaluation von Editionswerkzeugen zur nutzergenerierten Transkription handschriftlicher* Quellen. Master Informationswissenschaften, 148 p.
42. Brumen, M., Blatnik, A. (2012). *Recent developments and results of the european library project Ebooks on Demand (EOD), National and University Library, Slovenia*. Преглед НЦД 21, p. 87-93.
43. Budzise-Weaver, T., Chen, J., Mitchell, M. (2012). *Collaboration and Crowdsourcing: The Cases of Multilingual Digital Libraries*. The Electronic Library, Vol. 30, Issue 2, 11 p.
44. Bureau Van Dijk, Information Management, Bibliothèque nationale de France (2015). *Réalisation d'une étude d'usages des utilisateurs de la plateforme expérimentale Correct : rapport final.* 77 p.

45. Cardon, D.. *La démocratie Internet. Promesses et limites*, Éditions du Seuil, coll. « La république des idées », 2010, 102 p
46. Cardon, D., Casilli, A. (2016). *Qu'est-ce que le digital labor ?: Les enjeux de la production de valeur sur Internet et la qualification des usages numériques ordinaires comme travail. (Etudes et controverses)*. INA, 104 p.
47. Carletti, L., Giannachi, G., Price, D., McAuley, D. (2013). *Digital Humanities and Crowdsourcing: An Exploration*. MW2013: Museums and the Web 2013: The annual conference of Museums and the Web, April 17-20, 2013, Portland, USA. 18 p.
48. Casemajor Loustau, N. (2011). *La contribution triviale des amateurs sur le Web : quelle efficacité documentaire ?*. Études de communication 2011 (1):39-52.
49. Causer, T., Wallace, V. (2012). *Building A Volunteer Community: Results and Findings from Transcribe Bentham*. Digital Humanities quarterly, 6(2), 26 p
50. Causer, T., Tonra, J., Wallace, V. (2012). *Transcription maximized; expense minimized? Crowdsourcing and editing The Collected Works of Jeremy Bentham*. Literary and Linguistic Computing, Vol. 27, No. 2, p. 119-137
51. Chamberlain, E. (2010). Digitisation-on-Demand in Academic Research Libraries. 62 p.
52. Chamberlain, E. (2012). *Investigating Faster Techniques for Digitization and Print-on-Demand*, New Review of Academic Librarianship, 18:1, 57-71
53. Chardonnens, A. (2015). *Collections iconographiques numérisées et crowdsourcing : Possibilités et limites de la co-création de métadonnées par le grand public au travers de trois études de cas*. Université libre de Bruxelles, Mémoire de Master en Gestion Culturelle. 112 p.
54. Chartron, G. (2013). *RX : Recherche eXpérience*. In Germain, M., Pérales, C., Buffard, P., Chaudiron, C., Charaudeau, M.-O., Garnier, A., Chartron, G., Salaün, J.-M., « Les organisations du XXIe siècle », Documentaliste-Sciences de l'Information 2013/4 (Vol. 50), p. 38-47.
55. Chaudiron, S. (2013). *Ordres et désordres numériques*. In Germain, M., Pérales, C., Buffard, P., Chaudiron, C., Charaudeau, M.-O., Garnier, A., Chartron, G.,

Salaün, J.-M., « Les organisations du XXIe siècle », Documentaliste-Sciences de l'Information 2013/4 (Vol. 50), p. 38-47.

56. Christensen, H. S., Karjalainen, M., Nurminen, L. (2014). *What does crowdsourcing legislation entail for the participants? The Finnish case of Avoin Ministeriö.* Paper prepared for IPP2014: Crowdsourcing for Politics and Policy, University of Oxford, 25-26 September 2014. 19 p.
57. Chrons, O. Sundell S. (2011). *Digitalkoot: Making Old Archives Accessible Using Crowdsourcing*. HCOMP 2011: 3rd Human Computation Workshop
58. Chun, S., Cherry, R., Hiwiller, D., Trant, J., Wyman, B. (2006). *Steve.museum: An Ongoing Experiment in Social Tagging, Folksonomy, and Museums*, in J. Trant and D. Bearman (eds.). Museums and the Web 2006: Proceedings, Toronto: Archives & Museum Informatics, [17] p.
59. Y. Citton, *Pour une écologie de l'attention*. Paris: Seuil, 2014
60. Cohen, D. (2015). *Le Monde est clos et le désir infini*. Albin Mcihel, 224 p.
61. Colquhoun, B. (2013). *Making Sense of Historic Photographic Collections on Flickr The Commons: Institutional and User Perspectives*. In Museums and the Web 2013, N. Proctor & R. Cherry (eds). Silver Spring, MD: Museums and the Web.
62. Conteh, A., Tzadok, A. (2009). *User collaboration in mass digitisation of textual materials*. Proceedings: Cultural Heritage on line. Empowering users: an active role for user communities. 7 p.Crowston, K., Prestopnik, N. R. (2013). *Motivation and data quality in a citizen science game: A design science evaluation*. Forty-sixth Hawaii International Conference on System Sciences (HICSS-46), 10 p.
63. Daele, A. (2009). *Les communautés de pratique*. In J.-M. Barbier, É. Bourgeois, G. Chapelle, & J.-C. Ruano-Borbalan (Eds.), Encyclopédie de la formation (pp. 721–730). Paris: PUF.
64. Danowski, P. (2007). *Library 2.0 and User-Generated Content: What can the users do for us?* World Library and Information Congress: 73rd IFLA General Conference and Council 19-23 August 2007, Durban, South Africa. [11] p.
65. Daudey, E., Hoibian, S. (2014). *La société collaborative: mythe et réalité*. CREDOC, Cahier de recherche no. 313, 65 p.

66. De Boer, V., Hildebrand, M., Aroyo, L., De Leenheer, P., Dijkshoorn, C., Tesfa, B., Schreiber, G. (2012). Nichesourcing: Harnessing the Power of Crowds of Experts. Knowledge Engineering and Knowledge Management. Lecture Notes in Computer Science Volume 7603, 2012, pp 16-20

67. Delaine, V. (2014). *L'accompagnement du changement en bibliothèques : une approche managériale. Mémoire de fin d'études, diplôme de conservateur de bibliothèques*. ENSSIB. 84 p.

68. Delcourt, T., Le More, H., (2001). *Un nouveau service pour les lecteurs : la reproduction de livres à la demande à la bibliothèque de Troyes*. Bulletin des Bibliothèques de France, 5 : 94-102.

69. Delfino, M. (2011). *Against BibliOblivion: How modern scribes digitized an old book*. Computers & Education, 57 : 2145–2155.

70. Deodato J. (2014). *The patron as producer: libraries, web 2.0, and participatory culture*. Journal of Documentation, Vol. 70 Iss 5, p. 734-758

71. Deng, X. (N.), Joshi, K. D. (2013). *Is crowdsourcing a source of worker empowerment or exploitation? Understanding crowd workers' perceptions of crowdsourcing career*. Thirty Fourth International Conference on Information Systems, Milan 2013. 10 p.

72. Deterding, S., Dixon, D., Khaled, R., Nacke, L. (2011). *From game design elements to gamefulness: defining "gamification"*. Proceeding MindTrek '11 Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, p. 9-15

73. Deterding, S, Dixon, D, Khaled, R., Nacke, L. E. (2011). *Gamification : toward a Definition*. ACM CHI Gamification Workshop, p. 7–12

74. Deterding, S., Sicart, M., Nacke, L., O'Hara, K. (2011). *Gamification: Using Game Design Elements in Non-Gaming Contexts*. Proceeding CHI EA '11 CHI '11 Extended Abstracts on Human Factors in Computing Systems, p. 2425-2428

75. Djupdahl, M., Eskor, E., Jonsson, O., Runardotter, M., Ruusalepp, R., Sigurðsson, N. (2013). *Review of Crowdsourcing Projects and how they relate to Linked Open Data*. 15 p.

76. Doan, A. Ramakrishnan, R. Halevy, A.Y. (2011). *Crowdsourcing Systems on the World-Wide Web*. communications of the acm, 54(4): 86-96.

77. Dougherty, W. C. (2009). *Print on Demand: What Librarians should know*. The Journal of Academic Librarianship 35(2):184-186.

78. Dunn, S. Hedges, M. (2012). *Crowd-Sourcing Scoping Study Engaging the Crowd with Humanities Research*. Centre for e-Research, Department of Digital Humanities King's College London. 56 p.

79. Dunn, S., Hedges, M. (2013). *Crowd-sourcing as a component of humanities research infrastructures*. International Journal of Humanities and Arts Computing vol. 7, no. 1-2, p. 147–169

80. Dworak, M. (2012). *The Public as Collaborator:Towards Developing Crowdsourcing Models for Digital Research Initiatives*. B Sides : Journal of the University of Iowa School of Library and Information Science.

81. Earle, E. F. (2014). *Crowdsourcing metadata for library and museum collections using a taxonomy of Flickr user behavior*; Thesis, Master of Science, Cornell University, 105 p.

82. Eickhoff, C., Harris, C. G., de Vries, A. P., Srinivasan, P. (2012). *Quality through Flow and Immersion: Gamifying Crowdsourced Relevance Assessments*. SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. p. 871-880.

83. Elie, F., Quesnel, O. (2007). *Indexation collaborative et folksonomies*. Documentaliste - Sciences de l'information vol. 44, no 1, p. 58-63

84. Ellis, S. (2014). *A History of Collaboration, a Future in Crowdsourcing: Positive Impacts of Cooperation on British Librarianship*. Libri 64(1), p. 1-10

85. Emerson Johnson, R. (2012). Crowdsourcing and Library 2.0. 12 p.

86. Estellés-Arolas, E. González-Ladrón-de-Guevara, F. (2012). *Towards an integrated crowdsourcing definition*. Journal of Information Science, 14 p.

87. Estermann, B. (2014). *Diffusion of Open Data and Crowdsourcing among Heritage Institutions: Results of a Pilot Survey in Switzerland*. Journal of Theoretical and Applied Electronic Commerce Research, vol. 9, issue 3, p. 15-31

88. Estermann, B. (2015). *Diffusion of Open Data and Crowdsourcing among Heritage Institutions: Based on data from Finland, Poland, Switzerland, and The Netherlands*. EGPA 2015 Conference, 27 p.
89. Eveleigh, A., Jennett, C., Lynn, S., Cox, A. L. (2013). “ I want to be a Captain ! I want t o be a captain!”: Gamification in the Old Weather Citizen Science Project. Gamification '13: Proceedings of the First International Conference on Gameful Design, Research, and Applications. p. 79 - 82
90. Eveleigh, A. (2014) *Crowding out the archivist? Locating crowdsourcing within the broader landscape of participatory archives*. In: Ridge, M, (ed.) Crowdsourcing our cultural heritage. Ashgate Publishing: Surrey, p. 211-212
91. Flanagan M., Carini P. (2012). *How games can help us access and understand archival images*. The American archivist, Vol. 75, p. 514-537.
92. Fleurbaay, E., Eveleigh, A. (2012). *Crowdsourcing: Prone to Error?* International Council on Archives Conference 2012 Brisbane, Australia. 10 p.
93. Fort, K., Adda, G., Cohen, K. B. (2011). *AmazonMechanical Turk: Gold Mine or Coal Mine?* Computational Linguistics, 37 (2):413-420.
94. Fuchs, C. (2012). *Dallas Smythe Today - The Audience Commodity, the Digital Labour Debate, Marxist Political Economy and Critical Theory. Prolegomena to a Digital Labour Theory of Value*. tripleC 10(2): 692-740
95. Geitgey, Terri (2011). *The University of Michigan Library Espresso Book Machine experience*. Library Hi Tech, Vol. 29, Issue 1, pp. 51-61.
96. Glahn, P. (2015). *Reveal Digital: an open access model empowering libraries to become publishers*, Learned Publishing, 28: 299-302
97. Good, B. M., Su, A. (2011). *Games with a scientific purpose*. Genome Biology, 12:135, 3 p.
98. Göttl, F. (2014). *Crowdsourcing with gamification*. Advances in Embedded Interactive Systems Technical Report, Volume 2, Issue 3, p. [15-19]
99. Gouil, H. (2014). *La place de la coopération dans une conception eudémoniste du travail et de l'échange économique*. Dans L. Hammond Ketilson et M.-P. Robichaud Villettaz (sous la direction de), Le pouvoir d'innover des coopératives : textes choisis de l'appel international d'articles scientifiques, p. 667-680.

100. Grosdhomme Lulin, E. (2013). *La République participative*. 51 p.

101. Groh, F. (2012). *Gamification: State of the Art Definition and Utilization*. Proceedings of the 4th Seminar on Research Trends in Media Informatics, p. 39-46

102. Gstrein, S., Mühlberger, G. (2009). *User-driven content selection for digitisation: the eBooks on Demand Network*. Proceedings of International Conference on Cultural Heritage, 6 p.

103. Gstrein, S., Mühlberger, G. (2011). *Producing eBooks on Demand - A European Library Network*. 12 p.

104. Hagood, J. (2012), *A brief introduction to data mining projects in the humanities*. Bul. Am. Soc. Info. Sci. Tech., 38: 20–23

105. Hamari, J., Koivisto, J., Sarsa, H. (2014). *Does Gamification Work? A Literature Review of Empirical Studies on Gamification*. In proceedings of the 47th Hawaii International Conference on System Sciences, 10 p.

106. Harris, C. G. (2013). *Applying human computation methods to information science*. PhD dissertation, University of Iowa, 2013. 207 p.

107. Hartman, A. (2014). *Identifying medieval manuscript fragments through crowdsourcing*. Archival outlook. p. 8-9

108. Hasan, N. (2014). *Library promotion and resource generation through crowdsourcing*. 5 p.

109. Havenga, M., Williams, K., Suleman, H. (2012). *Motivating Users to Build Heritage Collections Using Games on Social Networks*. In: Proceedings of 14th International Conference on Asia-Pacific Digital Libraries (ICADL '12), Volume 7634 of Lecture Notes in Computer Science, pages 279-288, Springer Berlin / Heidelberg.

110. Haythornthwaite, C. (2009). C*rowds and Communities: Light and Heavyweight Models of Peer Production*. Proceedings of the Hawaii International Conference On System Sciences, January 5-8, 2009, Big Island, Hawaii. [11] p.

111. Herdagdelen, A., Baroni, M. (2010). *The Concept Game: Better Commonsense Knowledge Extraction by Combining Text Mining and a Game with a Purpose*. Commonsense Knowledge: Papers from the AAAI Fall Symposium (FS-10-02). p. 52-57

112. Heerlien, M., Van Leusen, J., Schnörr, S., De Jong-Kole, S., Raes, N., Van Hulsen, K. (2015). *The natural history production line: An industrial approach to the digitization of scientific collections*. ACM Journal on Computing and Cultural Heritage, Vol. 8, No. 1, Article 3, 11 pages.

113. Ho, C. J., Chang, T. H., Lee, J. C., Hsu, J. Y. J., Chen, K. T. (2009). *KissKissBan: A Competitive Human Computation Game for Image Annotation*. in: Proceedings of the 2009 ACM SIGKDD Workshop on Human Computation, pp. 11-14, ACM Press, New York

114. Holley, R. (2009). *A success story: Australian Newspaper Digitisation Program*. Online Currents, Volume 23, Issue 6, p. 283-295.

115. Holley, R. (2009). *Many Hands Make Light Work:Public Collaborative OCR Text Correction in Australian Historic Newspapers National Library of Australia*. 28 p.

116. Holley, R. (2009). *Crowdsourcing and social engagement: potential, power and freedom for libraries and users*. 28 p.

117. Holley R. (2009). *How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs*. D-Lib Magazine, 15(3-4).

118. Holley, R. (2010). Tagging Full Text Searchable Articles: An Overview of Social Tagging Activity in Historic Australian Newspapers August 2008 — August 2009. D-Lib Magazine, 16 (1/2).

119. Holley, R. (2010). *Trove: Innovation in Access to Information in Australia*. Ariadne, 64, 9 p.

120. Holley, R. (2010). *Crowdsourcing: How and Why Should Libraries Do It?*. D-Lib Magazine, 16(3-4)

121. Holley, R. (2011). *Resource Sharing in Australia: Find and Get in Trove – Making "Getting" Better*. D-Lib Magazine, volume 17, number 3/4, 14 p.

122. Houllier, F., Merilhou-Goudard, J.-B., Andro, M., Charbonnel, F., Cointet, J.-P., Frey-Klett, P., Joly, P.-B, Leiser, H., Mambrini-Doudet, M., Hologne, O., Launay, J.-F., Le Gall, O., , Masson, J., Morcrette, N., Pujol, J.-L., Roturier, C. (2016). *Les*

sciences participatives en France. Rapport au Ministre de l'Enseignement Supérieur (3 livrets, 41, 28, 24 p.)

123. Huberman, B.A. Romero, D.M. Wu, F. (2009). *Crowdsourcing, attention and productivity.* Journal of Information Science, 35:758

124. Huvila, I. (2008). *Participatory archive: towards decentralised curation, radical user orientation, and broader contextualisation of records management*. Archival Science 8:15–36

125. Ipeirotis, P. G. (2010). *Analyzing the Amazon Mechanical Turk Marketplace*. XRDS, 17(2) : 16-21.

126. Ipeirotis, P. G. (2010) Demographics of Mechanical Turk. 14 p.

127. Ipeirotis, P. G., Paritosh, P. K. (2011) *Managing Crowdsourced Human Computation*. 20th International World Wide Web Conference, WWW 2011. 5 p.

128. Jockers M. L. Sag M. Schultz J. (2012). *Digital archives: Don't let copyright block data mining*. Nature. 2012;490(7418):29-30

129. Josse, I. (2013). *La bnF engagée dans un projet de R&D pour la conception de la plateforme Correct (Correction et enrichissement collaboratifs de textes)*. Bulletin des Bibliothèques de France, 5:37-38

130. Jovian L. T., Amprimo O. (2011). *OCR Correction via Human Computational Game*. In Proceedings of the 44th Hawaii International Conference on System Sciences. 10 p.

131. Kalfatovic, M. R. Kapsalis, E. Spiess, K.P. Van Camp, A. Edson, M. (2008). *Smithsonian Team Flickr: a library, archives, and museums collaboration in web 2.0 space*. Archival Science, 8:267–277

132. Karnin E. D., Walach E., Drory T. (2010). *Crowdsourcing in the Document Processing Practice (A Short Practitioner/Visionary Paper)*. In Daniel F., Facca F. M. (Eds.): ICWE 2010 Workshops, LNCS 6385, pp. 408–411.

133. Kaufmann, N., Schulze, T., Veit, D. (2011). *More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk*. Proceedings of the Seventeenth Americas Conference on Information Systems, Detroit, Michigan August 4th-7th 2011, 11 p.

134. Kaufman, G., Flanagan, M., Punjasthitkul, S. (2016). *Investigating the Impact of 'Emphasis Frames' and Social Loafing on Player Motivation and Performance in a Crowdsourcing Game*. #chi4good, CHI 2016, San Jose, CA, USA, p. 4122-4128

135. Kittur A., Nickerson J. V., Bernstein M. S., Gerber, E. M., Shaw A., Zimmerman J.,Lease M., Horton J. J. (2013). *The Future of Crowd Work*. CSCW '13, February 23–27, 2013. 17 p.

136. Kleemann, F., Voß, G. G., Rieder, K., Gissendanner, S. S. (2008). *Un(der)paid Innovators: The Commercial Utiliza-tion of Consumer Work through Crowdsourcing*. Science, Technology & Innovation Studies Vol. 4, No. 1, p. 5-26

137. Kleka, P., Łupkowski, P. (2014). *Gamifying science: the issue of data validation*. Homo Ludens 1(6), 13 p.

138. Klopp, S. (2014). *Numérisation et impression à la demande en bibliothèque : un panorama*. Mémoire de conservateur ENSSIB. 133 p.

139. Kowalska, M. (2013). *Crowdsourcing in Libraries*. 17 p.

140. Lagarrigue, M., Rossant, F., Pierrot, A., Gardes, J., Maldivi, C., Petit, E. (2014). *Assessing the quality of digital re-publishing of textual documents through the follow-up of a correction protocol by crowdsourcing*. International Workshop on Computational Intelligence for Multimedia Understanding - IWCIM, Nov 2014, PARIS, France.

141. Lakhani, K. (2013). *Using the crowd as an innovation partner*. Harvard Business review. [12] p.

142. Lallement, M. (2015). *L'Âge du faire : Hacking, travail, anarchie*. Paris, Seuil, 446 p.

143. Lawton, P., Casas, N. A., Moyer, J., Zarndt, F. (2016). *Catholic, Crowdfunded, and Collaborative: A Unique Approach to Newspaper Digitization*. IFLA, 19 p.

144. Lang, A. S. I. D., Rio-Ross, J. (2011). *Using Amazon Mechanical Turk to Transcribe Historical Handwritten Documents*. Code4lib Journal 15, p. 10-31

145. Law, E., Von Ahn, L. (2009). *Input-Agreement: A New Mechanism for Collecting Data Using Human Computation Games*. CHI '09 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, p. 1197-1206

146. Lebraty, J.-F., Lobre, K. (2015). *Crowdsourcing : porté par la foule*. ISTE editions, 134 p.

147. Le Crosnier, H., Neubauer, C., Storup, B. (2013). *Sciences participatives ou ingénierie sociale : quand amateurs et chercheurs co-produisent les savoirs*. Hermès, La Revue no. 67, p. 68-74.

148. Le Deuff, O (2006). *Folksonomies Les usagers indexent le web*. Bulletin des Bibliothèques de France, tome 51, no 4, p. 66-70

149. Le Deuff, O. (2015). *Les humanités digitales précèdent-elles le numérique ?* Imad Saleh. H2PTM 15, Iste éditions

150. Leetaru, K. H. (2011). *Culturomics 2.0: Forecasting large-sacle human behavior using global news media tone in time and space*. First Monday, Volume 16, Number 9, 21 pages.

151. Levi, A. S. (2014). *Memorializing Religion: Crowdsourcing, Minorities, and the Quest for Identity in Online Archives*. Advances in the Study of Information and Religion, vol.1, art. 9. 23 p.

152. Lewis, R. D. (2006). *When Cultures Collide: Leading across Cultures*. Nicholas Brealey International, 624 p.

153. Lewis, D. W. (2010). *The User-Driven Purchase Giveaway Library*. Educause Review, vol. 45, no. 5, P. 10–11

154. Lieberman, H., Smith, D. A., Teeter, A. (2007). *Common Consensus: a webbased game for collecting commonsense goals*. IUI'07, January 28–31, 2007, Hawaii, USA. [6] p.

155. Lièvre, P., Laroche, N. (2014). *Retour sur la notion de communauté épistémique*. 7éme Colloque GeCSO LEST CNRS Université Aix Marseille 4-5-6 juin 2014. 25 p.

156. Liew, C. L. (2014). *Participatory Cultural Heritage: A Tale of Two Institutions' Use of Social Media*. D-Lib Magazine. Volume 20, Number 3/4, [17] p.

157. Ligeon, L. (2012). *Crowd sourcing : labour struggles revisited?* 32 p.

158. Lipinski, M. (2014). *Sciences & citoyens : rapport au Président du CNRS, mission Sciences et Citoyens*. 88 p.

159. Loveluck B. (2008). *Internet, vers la démocratie radicale* ?. Le Débat, 4(151) : 150-166

160. Lutz, C., Hoffmann, C. P., Meckel, M. (2014). *Beyond just politics: A systematic literature review of online participation*. First Monday, vol. 19, No. 7, [36] p.

161. Machiavel, N. (1837). *Œuvres complètes de N. Machiavel*. Tome premier. Paris : Auguste Desrez, 1837

162. Mccarthy S. (2012). *Using gamification as an effective OCR crowdsourcing motivator*. 19 p.

163. McKinley, D. (2012). *A Cognitive Walkthrough of the What's the Score at the Bodleian? task interface to increase volunteer participation*. 14 p.

164. McKinley, D. (2012). *Optimizing crowdsourcing websites for volunteer participation. A case study: What's on the Menu? New York Public Library*. National Digital Forum conference 21 November 2012, Wellington, New Zealand. 14 p.

165. McKinley, D. (2012). *Practical management strategies for crowdsourcing in libraries, archives and museums*. 13 p.

166. McKinley, D. (2013). *Functionality and usability requirements for a crowdsourcing task interface that supports rich data collection and volunteer participation A case study: The New Zealand Reading Experience Database*. Master of Information Studies, School of Information Management, Victoria University of Wellington. 75 p.

167. McKinley, D. (2013). *How effectively are crowdsourcing websites supporting volunteer participation and quality contribution?* Presented at Hamilton City Library, Hamilton, New Zealand. 13 p.

168. McKinley, D. (2013). *Why evaluation isn't a party at the end: Evaluating crowdsourcing websites*. 9 p.

169. McKinley, D. (2015). Heuristics to support the design and evaluation of websites for crowdsourcing the processing of cultural heritage assets [report]. 31 p.

170. McKinsey France (2014). *Accélérer la mutation numérique des entreprises : un gisement de croissance et de compétitivité pour la France*. 134 p.

171. Meade J. E., (1952). *External Economies and Diseconomies in a Competitive Situation*, The Economic Journal, Vol. 62, No. 245 p. 54-67

172. Meyer, M., Molyneux-Hodgson, S. (2011). « Communautés épistémiques » : une notion utile pour théoriser les collectifs en sciences ? », Terrains & travaux 2011/1 (n° 18), p. 141-154.

173. Mező Z., Svoljšak S. Gstrein S. (2007). *EOD - European Network of Libraries for eBooks on Demand*. In Kovács L., Fuhr N., Meghini C. (Eds.): ECDL 2007, LNCS 4675, pp. 570–572

174. McShane, I. (2011). *Public libraries, digital literacy and participatory culture*. In Discourse: Studies in the Cultural Politics of Education, 32(3), p. 383–397.

175. Michel J.B. Shen Y.K. Aiden A.P. Veres A. Gray M.K. Pickett J.P. (2011). *Quantitative Analysis of Culture Using Millions of Digitized Books*. Science, 331(6014) : 176-82.

176. Milena, D., Jennings, E., Devreni-Koutsouki, A. (2015). *Citizen Science and Digital Cultural Heritage: Potential for Wider Engagement with the General Public*. 7 p.

177. Millerand, F., Heaton, L., Proulx, S. (2011). *Émergence d'une communauté épistémique: création et partage du savoir botanique en réseau*. in: Serge Proulx et Annabelle Klein, dir., Connexions: communication numérique et lien social, Presses universitaires de Namur, 14 p.

178. Millot, G., Beubauer, C., Storup, B. (2013). *La recherche participative comme mode de production de savoirs : un état des lieux des pratiques en France*. 94 p.

179. Moatti, A. (2015). *Au pays de Numérix*. Presses Universiatires de France. 166 p. ISBN 978-2130631446

180. Moirez, P. (2012). *Archives participatives*. Dans : Bibliothèques 2.0 à l'heure des médias sociaux. Bulletin des Bibliothèques de France, 187-197

181. Moirez, P. Moreux, J.P. Josse, I. (2013). *Etat de l'art en matière de crowdsourcing dans les bibliothèques numériques*. Livrable L-4.3.1 du projet de

R&D du FUI 12 pour la conception d'une plateforme collaborative de correction et d'enrichissement des documents numérisés. 77 p.

182. Moirez, P., Stutzmann, D. (2013). "*Signaler les ressources numérisées : enrichissement, visibilité, dissémination*" in Manuel de constitution de bibliothèques numériques. Paris, Editions du Cercle de La Librairie. ISBN 978-2765414131. p. 115-174.

183. Moirez, P. (2013). *Bibliothèques, crowdsourcing, métadonnées sociales*. Bulletin des Bibliothèques de France. 5:32-36

184. Moreno, M., Xu, A.(2013). *Innovation during evolution: Document Supply Service digitises library collections*. IFLA WLIC 2013 (Singapore). 12 p.

185. Morvan, A. (2013). *Recherche-action*. in Casillo, I., Barbier, R., Blondiaux, L., Chateauraynaud, F., Fourniau, J.-M., Lefebvre, R., Neveu, C., Salled, D., Dictionnaire critique et interdisciplinaire de la participation, Paris, GIS Démocratie et Participation, 2013

186. Moyle, M., Tonra, J., Wallace, V. (2011). *Manuscript transcription by crowdsourcing*: Transcribe Bentham. Liber Quarterly, the Journal of European Research Libraries. Vol 20, Issue 3/4

187. Mühlberger, G., Gstrein, S. (2009). *eBooks on Demand (EOD): a European digitization service*. IFLA Journal 35(1): 35-43.

188. Nelson, M. J. (2012). *Soviet and American Precursors to the Gamification of Work*. Proceedings of the 16th International Academic MindTrek Conference. p. 23-26

189. Néroulidis, A. (2015). *Le crowdsourcing appliqué aux archives numériques : concepts, pratiques et enjeux*. Mémoire de recherche sciences de l'information et des bibliothèques. 109 p.

190. Neudecker, C., Tzadok, A. (2010). *User Collaboration for Improving Access to Historical Texts*. Liber Quarterly 20 (1), p. 119–128

191. Newby G. B., Franks C. (2003). *Distributed Proofreading.* In Proceeding JCDL '03 Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries, p. 361-363

192. Nguyen L. C., Partridge H. L., Edwards S. L. (2012). *Towards an understanding of the participatory library*. Library Hi Tech, 30(2), p. 335-346.

193. Noordegraaf, J., Bartholomew, A., Eveleigh, A. (2014). "*Modeling Crowdsourcing for Cultural Heritage*." MW2014: Museums and the Web 2014, [16] p.

194. Ollikainen, M. (2013). *On gamification. Master's thesis of the University of Tampere*, School of Information Sciences Computer Science. 70 p.

195. Onnée, S., Renault, S. (2013). *Le financement participatif : atouts, risques et conditions de succès*, Gestion, Vol. 38, p. 54-65.

196. Onnée, S., Renault, S. (2014). *Crowdfunding : vers une compréhension du rôle joué par la foule*, Management & Avenir, 2014/8 N° 74, p. 117-133.

197. Oomen, J., Brinkerink, M., Heijmans, L., Van Exel, T. (2010). *Emerging Institutional Practices: Reflections on Crowdsourcing and Collaborative Storytelling*. In J. Trant and D. Bearman (eds). Museums and the Web 2010: the international conference for culture and heritage on-line. Proceedings. Toronto: Archives & Museum Informatics, [9] p.

198. Oomen, J. Aroyo, L. (2011). *Crowdsourcing in the cultural heritage domain: opportunities and challenges.* In: 5th International Conference on Communities & Technologies. Brisbane, Australia - 29 June – 2 July 2011.

199. Oosterman, J., Bozzon, A., Houben, G.-J., Nottamkandath, A., Dijkshoorn, C., Aroyo, L., Leyssen, M. H. R., Traub, M. C. (2014). *Crowd vs. Experts: Nichesourcing for Knowledge Intensive Tasks in Cultural Heritage*. WWW'14 Companion, April 7–11, 2014, Seoul, Korea. ACM 978-1-4503-2744-2/14/04. p. 567-568

200. Oosterman, J., Nottamkandathy, A., Dijkshoorn, C., Bozzon, A., Houben, G.J., Aroyo, L. (2014 bis). *Crowdsourcing Knowledge-Intensive Tasks in Cultural Heritage*. In F. Menczer, J. Hendler, W. H. Dutton, M. Strohmaier, C. Cattuto, and E. T. Meyer, editors, ACM Web Science Conference, WebSci '14, Bloomington, IN, USA, June 23-26, 2014, pages 267–268. ACM, 2014.

201. Organisciak, P. (2010). *Why Bother? Examining the Motivations of Users in Large-Scale Crowd-Powered Online Initiatives*. A thesis submitted to the Faculty of

Graduate Studies and Research. Master of Arts Humanities Computing - Library and Information Studies, University of Alberta. 159 p.

202. Owens, T. (2013), *Digital Cultural Heritage and the Crowd*. Curator: The Museum Journal, 56: 121–130.

203. Pääkkönen, T. (2015). *Crowdsourcing Metrics of Digital Collections.* Liber Quarterly: the Journal of the Association of European Research Libraries. [8] p.

204. Paraschakis, D. (2013). *Crowdsourcing cultural heritage metadata through social media gaming*. Master Thesis School of Technology, Department of Computer Science. Malmö University. 70 p.

205. Paraschakis, D., Gustafsson Friberger, M. (2014). *Playful crowdsourcing of archival metadata through social networks*. 2014 ASE Bigdata / Socialcom / Cybersecurity Conference, Stanford University, May 27-31, 2014. 9 p.

206. Peccatte, P. (2009). *Flickr et PhotosNormandie : une entreprise collective de redocumentarisation*. Documentaliste Sciences de l'information, 46(1)

207. Petersen, S. M. (2008). *Loser Generated Content: From Participation to Exploitation*. First Monday, 13(3)

208. Peugeot, V. (2012). *Biens communs et numerique : l'alliance transformatrice*. In Lisette Calderan and Pascale Laurent and Helene Lowinger and Jacques Millet. Le document numerique à l'heure du web, ADBS, pp.141-154, 2012, Le document numerique a l'heure du web de donnees; Sciences et techniques de l'information, 9 p.

209. Peugeot, V., Beuscart, J.-S., Pharabod, A.-S., Trespeuch, M. (2015). *Partager pour mieux consommer ? Enquête sur la consommation collaborative*. Esprit 2015/7, p. 19-29

210. Pignal M., Pérez E. (2013). *Numériser et promouvoir les collections d'histoire naturelle*. Bulletin des Bibliothèques de France, tome 58, no. 5, p. 27-31

211. Poetz, M. K., Schreier, M. (2012). *The Value of Crowdsourcing: Can Users Really Compete with Professionals in Generating New Product Ideas?* Journal of Product Innovation Management 29(2):245-256.

212. Quinn, A. J., Bederson, B. B. (2011). *Human Computation: A Survey and Taxonomy of a Growing Field*. CHI '11 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, p. 1403-1412

213. Radice, S. (2014). *Designing for participation within cultural heritage: Participatory practices and audience engagement in heritage experience processes*. Politecnico di Milano, Design Department. 353 p.

214. Randall, P. (2016). *Purposeful Gaming and the Biodiversity Heritage Library,* Journal of Agricultural & Food Information, 17:1, 71-76

215. Rathemacher, A. J. (2015). *Crowdfunding Access to Archives*. Library Journal, February 2015, 3 p.

216. Reighart, R. Oberlander, C. (2008). *Exploring the future of interlibrary loan: generalizing the experience of the University of Virginia, USA*. Interlending & Document Supply, Vol. 36 Iss 4 p. 184-190

217. Renault S. (2014). *Crowdsourcing : La nébuleuse des frontières de l'organisation et du travail*, RIMHE : Revue Interdisciplinaire Management, Homme(s) & Entreprise, 2014/2 no. 11, p. 23-40.

218. Renault, S. (2014 bis). *Comment orchestrer la participation de la foule à une activité de crowdsourcing ? La taxonomie des 4 C*, Systèmes d'information & management, 2014/1 Volume 19, p. 77-105.

219. Revitt, M. (2013). *A Shared Approach to Managing Legacy Print Collections in Maine*. Maine Policy Review, volume 22, issue 1, 65-66

220. Revitt, M., Guthro, C. (2013). *Together we are Stronger: A Cooperative Approach to Managing Print Collections*. IFLA WLIC 2013 Singapore. 11 p.

221. Rifkin, J. (1996). *La Fin du travail : Ou comment l'Europe se substitue peu à peu à l'Amérique dans notre imaginaire*, La Découverte, 1996, 456 p.

222. Rifkin, J. (2014). *La nouvelle société du coût marginal zéro : L'internet des objets, l'émergence des communaux collaboratifs et l'éclipse du capitalisme*, Les liens qui libèrent, 2014, 510 p.

223. Ridge, M. (2011). *Playing with Difficult Objects: Game Designs to Improve Museum Collections*. In J. Trant and D. Bearman (eds). Museums and the Web 2011: Proceedings. Toronto: Archives & Museum Informatics. 83 p.

224. Ridge, M. (2013). *From tagging to theorizing: deepening engagement with cultural heritage through crowdsourcing. Curator*. The Museum Journal, 56(4), pp. 435–450.

225. Rinne, N. (2012). *Teaching with Google Books: research, copyright, and data mining*. 46 p.

226. Roegel, D. (2013). *La numérisation durable*. 231 p.

227. Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., Vukovic, M. (2011). *An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets*. 9 p.

228. Rorissa, A. (2010). *A Comparative Study of Flickr Tags and Index Terms in a General Image Collection*. Journal of the American Society for Information Science and Technology, 61(11):2230-2242

229. Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., Tomlinson, B. (2010). *Who are the Crowdworkers? Shifting Demographics in Mechanical Turk*. Proceeding CHI EA '10 CHI '10 Extended Abstracts on Human Factors in Computing Systems. P. 2863-2872

230. Roth, Y. (2016). *Comprendre la participation des internautes au crowdsourcing : Une étude des antécédents de l'intention de participation à une plateforme créative*. Thèse en Sciences de gestion, Paris 1. 426 p.

231. Rouse, A. C. (2010), *A Preliminary Taxonomy of Crowdsourcing*. ACIS Proceedings. Paper 76. 10 p.

232. Rusbridge, C. (1995). *The Electronic Libraries Programme. Serials* - Vo1.8, no 3, p. 231-240

233. Sabou, M., Bontcheva, K., Scharl, A., Föls, M. (2013). *Games with a Purpose or Mechanised Labour? A Comparative Study*. i-Know '13 Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies. Article No. 19, [8] p.

234. Sagot B., Fort, K., Adda, G., Mariani, J., Lang, B. (2011). *Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé*. TALN 2011, Montpellier, 27 juin – 1er juillet 2011, [12] p.

235. Saint-Luc, F. (2014). *La recherche-action, une recherche à visée formatrice et transformatrice*. 27 p.

236. Sarrouy, O. (2014). *Faire foule. Organisation, communication et (dé)subjectivisation à l'ère hyperindustrielle*. Thèse de l'Université de Rennes 2. 427 p.

237. Saylor, N., Wolfe, J. (2011). *Experimenting with Strategies for Crowdsourcing Manuscript Transcription*. Research Library Issues: a quarterly report from ARL, CNI, and SPARC

238. Schenk, E., Guittard, C. (2010). *Le crowdsourcing : modalités et raisons d'un recours à la foule*. 16 p.

239. Schenk, E., Guittard, C. (2012). *Une typologie des pratiques de Crowdsourcing : l'externalisation vers la foule, au-delà du processus d'innovation*. Management international, vol. 16, 2012, p. 89-100

240. Schildhauer, T., Voss, H. (2014). *Open Innovation and Crowdsourcing in the Sciences*. In Opening Science: the evolving guide on how the Internet is changing research, collaboration and scholarly publishing, pp.255-269

241. Scholz, T. (2008). *Market Ideology and the Myths of Web 2.0*. First Monday, 13(3).

242. Schultz, P. (2005). *The Producer as Poweruser*, in: Engineering Culture: 'on the author as (digital) producer', edited by Geoff Cox, Joasia Krysa, New York 2005, p. 111-125

243. Sharma, A. (2010). *Crowdsourcing Critical Success Factor Model: Strategies to harness the collective intelligence of the crowd*. Working Paper. 22 p.

244. Shirky, C. (2008). *Here comes everybdy: the power of organizing without organizations*. Penguin Books. 344 p.

245. Šimko, J., Tvarožek, M., Bieliková, M. (2013). *Human computation:Image metadata acquisition based on a single-player annotation game*. Int. J.Human-Computer Studies, 71:933–945.

246. Smith, D., Manesh, M. Mehdi Ghar, & Alshaikh, A. (2013). *How Can Entrepreneurs Motivate Crowdsourcing Participants?*. Technology Innovation Management Review, 3(2): 23-30.

247. Smith-Yoshimura, K., Shein, C. (2011). *Social Metadata for Libraries, Archives and Museums Part 1: Site Reviews*. OCLC Research. 174 p.

248. Smith-Yoshimura, K., Godby, C. J., Hoffler, H., Varnum, K., Yakel, E. (2011). *Social Metadata for Libraries, Archives, and Museums:Survey analysis*. OCLC Research. 73 p.

249. Smith-Yoshimura, K. (2012). *Social Metadata for Libraries, Archives, and Museums:Executive Summary.* OCLC Research. 20 p.

250. Smith-Yoshimura, K., Holley, R. (2012). *Social Metadata for Libraries, Archives, and Museums: Recommendations and Readings.* OCLC Research. 78 p.

251. Snow, R., O'Connor, B., Jurafsky, D., Ng, A. Y. (2008). *Cheap and Fast—But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks*. Proceeding EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing, p. 254-263

252. Solemon, B., Ariffin, A., Md Din, M., Md Anwar, R. (2013). *A review of the uses of the crowdsourcing in higher education*. International Journal of Asian Social Science, 3(9):2066-2073

253. Soulé, B. (2007). *Observation participante ou participation observante? Usages et justifications de la notion de participation observante en sciences sociales*. Recherche qualitatives, vol. 27(1), p. [127]-140

254. Spindler, R. P. (2014). *An Evaluation of Crowdsourcing and Participatory Archives Projects for Archival Description and Transcription*. 26 p.

255. Springer, M., Dulabahn, B., Michel, P., Natanson, B., Reser, D., Woodward, D., Zinkham, H. (2008). *For the Common Good: The Library of Congress Flickr Pilot Project*. 50 p.

256. Stambaugh, E. K. (2013). *Reinventing Shared Print: A Dynamic Service Vision for Shared Print Monographs in a Digital World*. CDL Staff Publications. Against the Grain. 68-70

257. Steinbach, L. (2014). *Digital cultural heritage is getting crowded: crowdsourced, crowd-funded, and crowd-engaged*. in: Digital Heritage and Culture: Strategy and Implementation 4th Reading. P. 261-294

258. Stevenson, S. (2010). *When Citizens Become Consumer-Producers: Immaterial Labour and the Unpaid Work of Patrons in the Library as Place and Virtual Space*. 4 p.

259. Stiegler, B. (2015). *La société automatique. 1, l'avenir du travail*. Fayard. 300 p.

260. Stiller, J. (2014). *From Curation to Collaboration A Framework for Interactions in Cultural Heritage Information Systems*. PhD Humboldt-Universität zu Berlin. 268 p.

261. Surowiecki, J. (2004). *La sagesse des foules* (traduction de the wisdom of crowds). 384 p.

262. Szoniecky, S. (2012). *Évaluation et conception d'un langage symbolique pour l'intelligence collective : Vers un langage allégorique pour le Web*, Thèse en Science de l'information et de la communication, Université Paris VIII Vincennes-Saint Denis, 2012

263. Tafuri, N., Mays, A. (2011). *Bullied by Budgets, Pushed by Patrons, Driven by Demand: Libraries and Tantalizing Technologies*. Proceedings of the Charleston Library Conference. p. 317-323

264. Terras, M. (2010). *Digital curiosities: resource creation via amateur digitization. Literary and Linguistic Computing* 25(4): 425-438.

265. Thuan, N. H., Antunes, P. Johnstone, D. (2013). *Factors Influencing the Decision to Crowdsource*. Lecture Notes in Computer Science Volume 8224, p. 110-125

266. Thogersen, R. (2012). *Crowdsourcing for image metadata; a comparison between game-generated tags and professional descriptors. 103 p.*

267. Tonra, J. (2013). *Manuscript transcription: the habits of crowds*. 8 pages.

268. Trainor, C. (2008). *Open Source, Crowd Source: harnessing the power of the people behind our libraries*. Library Faculty and Staff Papers and Presentations. Paper 3. 9 p.

269. Trant, J., Wyman, B. (2006). I*nvestigating social tagging and folksonomy in art museums with steve.museum*. Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland. 6 p.

270. Tweddle, J.C., Robinson, L.D., Pocock, M.J.O., Roy, H.E (2012). *Guide to citizen science: developing, implementing and evaluating citizen science to study biodiversity and the environment in the UK*.29 p.

271. Vandooren F, Gass C. (2008). *Giving new life to out-of-print books: when publishers' and libraries' interests meet*. Learned Publishing 21(3):187-92.

272. Venkatesh, A., Lalitha, M. V., Narayana, J., Mahesh, K. (2015). *Wikiaudia: Crowd-sourcing the Production of Audio and Digital Books*. Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol I, IMECS 2015, 6 p.

273. Vershbow, B. (2013). *NYPL Labs: Hacking the Library*. Journal of Library Administration 53(1): 10–26.

274. Von Ahn, L., Dabbish, L. (2004). *Labeling Images with a Computer Game. ACM Conf. on Human Factors in Computing Systems*, CHI, p.319-326.

275. Von Ahn, L. (2006). *Games With A Purpose*. IEEE Computer Magazine,p. 96-98.

276. Von Ahn, L. Liu, R., Blum, M. (2006) *Peekaboom: A Game for Locating Objects in Images*. Proceeding CHI '06 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. P. 55-64

277. Von Ahn, L., Kedia, M., Blum, M. (2006). *Verbosity: A Game for Collecting Common-Sense Facts*. Proceeding CHI '06 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, p. 75-78

278. Von Ahn, L., Dabbish, L. (2008). *Designing Games With A Purpose*. Communications of the ACM, vol. 51, no. 8. p. 58-67

279. Von Ahn, L., Maurer, B., Mcmillen, C., Abraham, D., Blum, M. (2008). *reCAPTCHA: Human-Based Character Recognition via Web Security Measures*. Science 321: 1465-1468.

280. Von Hippel, E. (2005). *Democratizing innovation*. 204 p.

281. Von Hippel, E., Ogawa, S., De Jong, J. P. J. (2011).*The age of the Consumer-Innovator.* MITSloan Management Review, vol. 53, no. 1, p. 27-35.

282. Wenger, E. (1998). *Communities of Practice : Learning, Meaning,*

283. *and Identity*, Cambridge University Press, 336 p.

284. Wilson-Higgins, S. (2011). *Could print on-demand actually be the "new interlibrary loan"?* Interlending & Document Supply, 39(1):5-8

285. Witten IH, Don KJ, Dewsnip M, Tablan V (2004). *Text mining in a digital library*. International Journal on Digital Libraries. 4(1):56-9.

286. Yuxiang Zhao & Qinghua Zhu (2012). *Evaluation on crowdsourcing research: Current status and future direction*. Inf Syst Front 2012.

287. Zacklad, M. Chupin, L, (2015). *Le crowdsourcing scientifique et patrimonial à la croisée de modèles de coordination et de coopération hétérogènes : le cas des herbiers numérisés*, Canadian Review of Information Science, Vol. 39

288. Zastrow, J (2014). *Crowdsourcing Cultural Heritage: 'Citizen Archivists' for the Future*. The digital archivist, 34(8), 4 p.

289. Zlodi, G., Ivanjko, T. (2013). *Crowdsourcing Digital Cultural Heritage*. INFuture2013: "Information Governance", 199-207.