Cartographie de la Grande Guerre sur le Web

Valérie Beaudouin, Zeynep Pehlivan

Labex « Les passés dans le présent »

Cartographie de la Grande Guerre sur le Web

Rapport final du projet :
« Le devenir du patrimoine numérisé en ligne : l'exemple de la grande Guerre »
Phase 2

Valérie Beaudouin, Zeynep Pehlivan
Télécom ParisTech, UMR I3

Septembre 2016

Ce projet, 2013-2016, implique la BnF, la BDIC, Telecom ParisTech. Il est soutenu par le labex « Les passés dans le présent » (Investissements d'avenir, ANR-11-LABX-0026-01)

Contributeurs du projet :

BnF : Philippe Chevallier (chef de projet), Peter Stirling

Télécom ParisTech : Valérie Beaudouin (resp), Zeynep Pehlivan, Nicolas Auray

BDIC : Lionel Maurel (resp), Josselin Morvan

Table des matières

1 INTRODUCTION

Ce rapport présente les résultats de la phase 2 du projet « Le devenir du patrimoine numérisé en ligne : l'exemple de la Grande Guerre » porté par la Bibliothèque nationale de France (BnF), le département de sciences économiques et sociales de Telecom ParisTech, la Bibliothèque de documentation internationale contemporaine (BDIC). Ce projet d'une durée de trois ans (2013-2016) a été financé par le Labex « Les passés dans le présent ». Il porte sur la cartographie de la Grande Guerre sur le Web en 2014 en portant une attention spécifique au forum Pages 14-18 et à son environnement[1].

Les institutions patrimoniales ont entrepris depuis une vingtaine d'années un travail massif de numérisation et de mise à disposition des publics de documents (livres, journaux, photographies…) numériques. Elles s'interrogent sur la manière dont ces fonds numérisés sont découverts, commentés et partagés sur le web, en particulier dans des collectifs d'amateurs. Le patrimoine numérisé autour de la Grande Guerre est un domaine propice à une telle analyse : c'est un corpus relativement aisé à délimiter et qui a atteint sa pleine maturité en termes de masse documentaire et de consultation, témoignant d'un intérêt bien ancré pour la mémoire et l'histoire de la première guerre mondiale dans le grand public

Le fait que les collections soient accessibles en dehors des murs des institutions permet-il 1) d'atteindre des publics qui ne se déplaçaient pas sur les sites institutionnels, d'élargir l'accès aux archives ? 2) de développer des usages peu observables dans les salles de lecture ?

Quel est le devenir des documents patrimoniaux concernés par cette période, accessibles en ligne et souvent disséminés à travers le web ? Quelles sont les formes d'appropriation et de mobilisation des documents en dehors des interfaces institutionnelles qui y donnent accès ?

L'objectif de cette recherche est d'étudier les pratiques sociales en ligne qui visent à construire une représentation du passé et à entretenir la mémoire de la guerre : quels types de recherche sont menées sur l'histoire de la guerre, comment le fruit de ces recherches est publié, partagé et discuté en ligne, pour construire une mémoire collective de la guerre ? Comment ces recherches se situent-elles, à la croisée entre l'histoire individuelle et collective, entre la généalogie et l'histoire ? Plus

[1] Ce rapport a bénéficié des relectures d'Irène Bastard, de Philippe Chevallier, d'Anne Garréta et de Catherine Perret que nous remercions. Il a été rédigé par Valérie Beaudouin à l'exception de la section 4 co-rédigée avec Zeynep Pehlivan. La chaîne de traitement qui permet d'extraire les métadonnées pertinentes à partir des archives du Web a été mise en œuvre par Zeynep Pehlivan.

spécifiquement, en quoi le numérique renouvelle-t-il les pratiques historiques et mémorielles ?

Notre hypothèse de départ est que l'accessibilité sur le web, avec les fonctionnalités de recherche associées, transforme la manière de faire des recherches historiques et généalogiques, et qu'elle reconfigure les modalités de constitution de collectifs en ligne, voire de « communautés de pratiques » (Lave et Wenger, 1991), dans lesquelles l'utilisation, le référencement, l'agencement des documents numérisés, provenant de sources institutionnelles ou privées, occupe une place centrale.

Ce projet comprend trois phases :

La première phase d'instruction a consisté à dresser un état de l'art sur les travaux concernant l'historiographie de la Grande Guerre et sur les recherches portant sur les pratiques amateurs et ainsi à définir la méthodologie de recherche pour les phases 2 et 3.

La phase 2 visait à cartographier le web consacré à la Grande Guerre et les relations entre les espaces numériques, en accordant une place spécifique au forum Pages 14-18 et à son environnement. Ce forum avait été identifié comme un lieu important d'usages des corpus numérisés de Gallica par l'étude qualitative menée par Muriel Amar et Philippe Chevallier (Amar et Chevallier, 2013).

La phase 3 est expérimentale. Les Albums Valois (ensemble d'albums photographiques constitués par le service photographique des Armées) ont été numérisés et intégralement mis en ligne fin 2014. L'objectif était de suivre de manière empirique mais raisonnée la dissémination de ces albums sur le web, en particulier de mesurer les effets de différentes formes de médiations (échantillon « médiatisé » *versus* échantillon témoin).

Ce rapport concerne la phase 2 du projet. Il porte sur la cartographie du réseau des sites web dédiés pour tout ou partie à la Grande Guerre. L'objectif était d'identifier les sites du web consacrés à la Grande Guerre et d'analyser les liens entre ces sites, à partir des archives du web de la BnF. Une collecte spécifique autour de la Grande Guerre a été mise en place à partir d'un ensemble de sites identifiés par des professionnels. Elle est réalisée à dates régulières, pour permettre des analyses temporelles. La carte représente les sites et les relations entre les sites. Elle permet de caractériser les différents territoires qui organisent la carte (sites institutionnels, sites personnels et associatifs…). Elle révèle également la position des sites qui donnent accès à des ressources numérisées.

Une section méthodologique décrit précisément la démarche mise en œuvre et ses limites, pour qu'elle puisse être reprise par d'autres équipes de chercheurs, tandis que les sections suivantes présentent les résultats et l'interprétation.

Le rapport est nourri par une série d'entretiens menés auprès de participants (auteurs de sites et/ou participants du forum) et par l'analyse des contenus des sites et des fils de discussion. L'approche quantitative est systématiquement mise en perspective grâce à l'exploration ethnographique des espaces de publication et d'interactions et surtout grâce à la rencontre des acteurs participants à cette présence de la Grande Guerre sur le web. Autrement dit, les conclusions qui pourraient être hâtivement tirées d'approches « big data » sont confrontées aux représentations des acteurs.

Le projet vise autant à développer des outils d'analyse du web rigoureux, destinés à fournir des orientations et des critères pour une observation pertinente du web, qu'à fournir aux établissements publics des axes de réflexion pour la valorisation de leur patrimoine en ligne. En articulant une approche quantitative, à base de fouille de texte et de données, avec des approches qualitatives, on se prémunit contre les défauts de chacune des méthodes (absence du sens dans les premières, question de la représentativité pour les secondes).

Ce projet a été une occasion unique d'explorer les potentialités des archives du Web de la BnF comme outil pour les recherches en sciences sociales et nous espérons que la démarche méthodologique et les chaines de traitement que nous avons mises en place et qui sont disponibles à la BnF permettront à d'autres équipes de chercheurs de s'appuyer sur les archives du web pour mener à bien leurs recherches. Travailler sur les archives permet de bénéficier d'un cadre légal déjà établi et favorise la reproductibilité des recherches et l'avancement progressif de la recherche en gardant la trace des corpus analysés sur le temps long.

2 ETAT DE L'ART

La perspective que nous adoptons ici consiste à étudier les pratiques sociales liées à la mémoire ou à l'histoire (nous reviendrons sur ces deux termes) de la Grande Guerre, telles qu'elles se déploient dans l'univers numérique. Il s'agit, si l'on reprend les termes d'Henry Rousso, d'étudier « l'évolution des différentes pratiques sociales, de leur forme et de leur contenu, ayant pour objet ou pour effet, explicitement ou non, la représentation du passé et l'entretien de son souvenir, soit au sein d'un groupe donné, soit au sein de la société tout entière » (Rousso, 1987). Ceci nous conduit à croiser deux perspectives : celle de la sociologie de la mémoire et celle des usages de l'internet. Cette articulation nous conduit à explorer les pratiques mémorielles, en identifiant les procédures et les techniques mises en œuvre.

2.1 LES PASSES DANS LE PRESENT : LA QUESTION DE LA MEMOIRE DANS LES SCIENCES SOCIALES

2.1.1 Mémoire et histoire

Dès la fin des années 70, en Europe comme en Amérique du Nord, on a assisté à une explosion de références à la mémoire sociale ou collective tant dans l'espace social qu'académique (Olick et Robbins, 1998) avec un intérêt renouvelé pour les travaux pionniers de Halbwachs dans les *Cadres sociaux de la mémoire* de 1925 et la *Mémoire collective* de 1950 (Halbwachs, 1994, 1997)

La mise en avant dans l'espace public de la notion de mémoire, dès les années 80, l'invention et l'essor du devoir de mémoire dont Sébastien Ledoux retrace la généalogie (Ledoux, 2012), le « boom mémoriel » depuis les années 1990 (Rousso, 2016) en sont des manifestations. Il revient à Régine Robin dans *La mémoire saturée* (Robin, 2003) d'avoir pointé un paradoxe à cette vogue : en constatant que notre présent serait envahi par une obsession mémorielle, elle s'inquiète qu'un excès de mémoire puisse n'être qu'une figure de l'oubli.

Plusieurs hypothèses ont été proposées pour expliquer ce « boom mémoriel » : la montée du multiculturalisme (et des gender et postcolonial studies), qui critiquant l'historiographie comme une source de domination culturelle, cherche à déconstruire les liens entre histoire, mémoire et pouvoir ; la fin des idéologies tournées vers l'avenir (*ie* le communisme) ; la promotion de la figure de la victime et enfin l'accélération du changement qui fait peser une menace sourde liée à l'oubli (Olick et Robbins, 1998).

Revenons aux textes d'Halbwachs. Dans *les cadres sociaux de la mémoire*, Halbwachs pose que « tout semble indiquer que [le passé] ne se conserve pas, mais qu'on le reconstruit en partant du présent » (Halbwachs, 1994) p. VIII et surtout que « [les cadres collectifs de la mémoire] sont (…) précisément les instruments dont la mémoire collective se sert pour recomposer une image du passé qui s'accorde à chaque époque avec les pensées dominantes de la société ». (p. VIII). Les cadres sociaux tels que définis par le groupe sont le lieu où se réécrit la mémoire à travers les traces. Contre Bergson, Halbwachs affirme de manière très nette la dimension sociale de la mémoire.

Halbwachs, en évoquant la mémoire religieuse, qui fonctionne pour lui comme toute mémoire collective, donne une description précise de ce processus de construction mémorielle : « elle ne conserve pas le passé, mais elle le reconstruit, à l'aide des traces matérielles, des rites, des textes, des traditions qu'il a laissées, mais aussi à l'aide de données psychologiques et sociales récentes, c'est-à-dire avec le présent » (Halbwachs, 1994) p. 221. Nous réutiliserons cette idée d'articulations entre les traces du passé, qui pour nous désignent des matériaux qui font partie d'une archive, et le contexte social présent dans la construction de ce qui est sensé former la mémoire collective, catégorie qui mérite d'être elle-même réinterrogée.

Halbwachs distingue : *l'histoire,* comme souvenir du passé avec lequel on n'a plus de relation sensible ; la *mémoire autobiographique*, celle qui est liée aux événements que l'on a vécus ; la *mémoire historique* qui nous parvient au travers des documents, qui peut être vivante ou morte et enfin la *mémoire collective*, qui constitue un passé actif qui forme nos identités (Halbwachs, 1997). Pour Halbwachs, la mémoire collective est plurielle, liée à nos différents groupes d'appartenance.

Il nous semble important pour la suite de notre propos de revenir sur le texte qui ouvre les *Lieux de mémoire* de Pierre Nora : « Entre mémoire et histoire : la problématique des lieux » (Nora, 1984). Nora part d'une définition de la mémoire et de l'histoire que tout oppose : la mémoire est du côté de la vie, de l'expérience vécue, portée par des groupes vivants tandis que l'histoire est une représentation du passé, universelle certes mais sans lien avec un passé vécu.

Cependant quand l'histoire commence à faire sa propre histoire et entre dans son âge historiographique d'une part et que d'autre part se met en place un nouveau rapport à la mémoire où celle-ci n'est plus vécue comme intime mais reconstituée, la distinction entre mémoire et histoire n'est plus valide : « tout ce que l'on appelle aujourd'hui mémoire n'est donc pas de la mémoire mais déjà de l'histoire ». L'étude des lieux de mémoire résulte pour Nora de cette double évolution :

> Les lieux de mémoire naissent et vivent du sentiment qu'il n'y a pas de mémoire spontanée, qu'il faut créer des archives, qu'il faut maintenir des anniversaire, organiser des célébrations, prononcer des éloges funèbres, notarier des actes, parce que ces opérations ne sont pas naturelles.(p. XXIV)

La mémoire « transformée par son passage en histoire » acquiert trois caractéristiques selon Nora : elle devient *mémoire archive*, en accordant une place centrale aux traces matérielles du passé, aux archives ; elle devient *mémoire devoir*

en rendant l'individu responsable de ce travail de mémoire, l'impératif d'histoire s'étend bien au-delà des historiens ; elle pose une *mémoire distance* où le passé est présenté comme radicalement différent du présent, sans solution de continuité.

Les lieux de mémoire prennent leur place dans cette redéfinition de la mémoire : ils ont, toujours selon Nora, un triple aspect : matériel, fonctionnel et symbolique. Ils ont une dimension matérielle en ce qu'ils constituent une *forme* dans l'espace et dans le temps, fonctionnelle en assurant la cristallisation et la transmission du souvenir, et symbolique par le message qu'ils portent par-delà la représentation.

La perspective des lieux de mémoire, parce qu'elle met l'accent sur les entreprises institutionnelles, néglige la dimension sociale des pratiques de mémoire. L'étude des pratiques mémorielles sur le web souligne au contraire, comme nous le verrons, la dimension collective de la pratique, fondée sur le réseau relationnel.

Marie-Claire Lavabre propose une synthèse des différentes problématiques de la mémoire dans les sciences sociales (Lavabre, 2016). Elle distingue celle des *lieux de mémoire* attachée à Pierre Nora (Nora, 1984), celle du *travail de mémoire* porté par Paul Ricoeur (Ricoeur, 2003) et enfin celle des *cadres de la mémoire* dans la tradition d'Halbwachs (Halbwachs, 1994, 1997). Elle en vient à distinguer l'histoire, « opération intellectuelle qui requiert un exposé systématique, vise la connaissance et l'intelligibilité du passé », la mémoire historique, « une manière d'usage de l'histoire ou d'histoire finalisée, portée par des intérêts ou des « besoins » (Nietzche, Deuxième considération inactuelle), qui ne sont pas ceux de la connaissance, mais de l'exemple ou de la légitimité, de l'identité, de la polémique ou du jugement critique » et la mémoire commune, « c'est-à-dire mémoire de ce qui a été vécu, souvenirs et traces du passé ». La mémoire collective se situe en creux à l'intersection des trois, elle « repose sur le fait – empiriquement vérifié – qu'existent des conditions sociales à l'évocation et à la formulation des souvenirs et des processus – éventuels- d'homogénéisation des représentations du passé (Lavabre, 1994), notamment informées par les mises en récit publiques du passé. » (Lavabre, 2016). Elle propose ainsi une relecture contemporaine des travaux d'Halbwachs, qui peut nous aider à éclaire notre terrain de recherche.

En sociologie, en s'inscrivant dans cette tradition de Maurice Halbwachs, et dans une perspective pragmatiste, de nombreux travaux portent sur les pratiques mémorielles. Ainsi Sarah Gensburger explore la mémoire des camps annexes de Drancy et l'articulation entre mémoire individuelle et collective dans ce travail de remémoration (Gensburger, 2005). Gérôme Truc, à propos des commémoration des attentats du 11 septembre, montre comment les dimensions mémorielle et émotionnelle s'articulent dans ces mises en récit de l'expérience (Truc, 2015). Jonathan Michel, à propos de la mémoire de l'esclavage, introduit la notion complémentaire de mémoire publique, pour désigner « des groupes qui problématisent publiquement un « trouble » mémoriel » (Michel, 2015) . Nous

retenons de ces perspectives pragmatiques, le fait de porter attention à la manière dont se *fabrique* la mémoire, comment se combine art du récit et pratique sociale.

2.1.2 La passion généalogique

Parmi les pratiques mémorielles, celle de la généalogie a connu un essor remarquable et nous verrons qu'elle joue un rôle décisif dans les pratiques numériques autour de la Grande Guerre.

Si la pratique de la généalogie qui consiste à reconstruire la filiation des individus et des familles a été pendant longtemps réservée à une petite frange de familles, elle s'est largement démocratisée au cours du XXe siècle, bien avant la numérisation et mise en ligne des archives territoriales, qui facilitent grandement le travail.

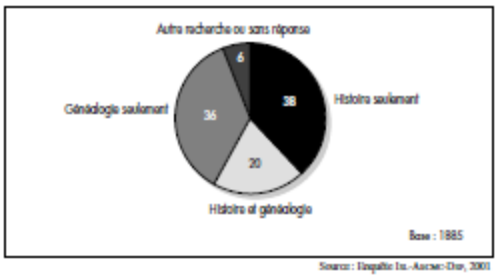
> L'histoire de la généalogie permet de constater qu'elle était jadis réservée à l'étude des grandes familles nobles ou bourgeoises avant de se populariser au cours du XXe siècle, plus particulièrement à partir des années 1960. (Harvey, 2005)

Les enquêtes sur les publics des archives montrent que la généalogie y occupe une place importante.

> En 30 ans, le public des Archives nationales et des archives départementales a plus que quadruplé, atteignant près de 197 000 lecteurs. Cet accroissement, depuis le début des années 1970, s'est accompagné de profonds changements dans la composition du public : les universitaires ont remplacé les « érudits » et ont été, à leur tour, relayés par des amateurs, en particulier les généalogistes curieux de découvrir leurs origines familiales. (DEPS, 2001)

Plus de la moitié des lecteurs fréquentent les archives pour faire des recherches généalogiques en 2000.

Graphique 4 – Répartition des grands domaines de recherche aux archives départementales et communales



Dans cette même enquête, il apparaît que l'histoire est le second grand domaine de recherche avec 20% des lecteurs qui ont un double intérêt pour l'histoire et pour la généalogie.

La mise en ligne d'une grande partie des archives n'a fait que renforcer cette tendance de fond - cet intérêt croissant pour la généalogie -, tout en modifiant profondément les modalités d'accès. En effet, en même temps que la fréquentation des salles décline la consultation des archives en ligne connait une très vive croissance. Les usages des généalogistes se transforment, ils sont moins présents en salle mais beaucoup plus actifs en ligne.

D'après la dernière étude menée par les Archives en 2013-2014 (publiée en 2015), 95% des personnes qui se connectent aux sites des Archives départementales le font dans un but de recherche généalogique (Guigueno, 2015). Si le profil des lecteurs en salle est proche de celui des visiteurs de monuments en termes de formation, le profil des internautes qui fréquentent les archives est beaucoup plus populaire, et tendrait à montrer une véritable démocratisation de la pratique (50% ont un diplôme inférieur au bac, contre 22% des personnes qui ont visité un musée dans les 12 derniers mois) (Guigueno, 2015).

Des travaux en ethnologie sur les pratiques des généalogistes, comme ceux d'Isabelle Sagnes, permettent de resituer dans le contexte spatial local la « passion » pour la généalogie (Sagnes, 1995, 2004).

2.1.3 Histoire et mémoire de la Grande Guerre

Si l'on suit Halbwachs, chaque nouvelle génération qui se penche sur son passé reconstruit la mémoire du passé en fonction de son propre présent. La manière de faire l'histoire de la Grande Guerre et d'en garder la mémoire s'est ainsi vue fortement renouvelée au cours des cent dernières années. Si l'histoire diplomatique et stratégique a prévalu dans l'immédiat après-guerre, puis une histoire culturelle de la vie au front et à l'arrière, se développe depuis ces dernières années une approche plus anthropologique de la guerre, centrée sur l'expérience vécue par les soldats, telle qu'elle peut nous être accessible via des témoignages de combattants de l'époque (carnets, photos, dessins…). Audoin-Rouzeau défend par exemple la nécessité d'avoir une approche anthropologique du combat qui le regarde au plus près au lieu de l'éviter comme les chercheurs en sciences sociales qui avaient vécu l'expérience de la guerre l'ont fait (Audoin-Rouzeau, 2008).

Dans *14-18 aujourd'hui*, Nicolas Offensdadt identifie trois types d'activisme mémoriel autour de la guerre : 1) le récit généalogique 2) le récit local 3) le récit du militant (réhabilitation des mutilés ou fusillés, mais aussi défense d'un point de vue régionaliste) (Offenstadt, 2010). L'intérêt du grand public pour la Grande Guerre emprunte en général l'une de ces trois voies : reconstruire le parcours d'un ancêtre, retracer les batailles, combats dans une zone géographique et enfin lever le voile sur les zones d'ombres de l'histoire, celles que l'Etat a tenté de masquer, à savoir ces mouvements de résistance contre la guerre : fraternisations, mutineries, fusillés…

L'intérêt du grand public pour le premier conflit mondial s'est particulièrement exprimé au moment du lancement de la Grande collecte dédiée à la Grande Guerre : nombreux ont été les participants qui ont confié à la numérisation les souvenirs de leurs ancêtres.

Le travail des historiens est aussi un lieu de *controverses* et de conflits d'interprétations. Ainsi deux lectures de la Guerre s'opposent au tournant du siècle entre ceux qui considèrent que les soldats sont partis au front dans un élan patriotique et ceux qui pensent au contraire qu'ils y sont allés sous la contrainte étatique. *Consentement* ou *coercition* sont les deux mots qui polarisent la guerre de

tranchée des historiens (Birnbaum, 2006), qui si elle s'est apaisée avec le centenaire pourrait avoir des échos dans les espaces en ligne.

Ce projet est l'occasion de réfléchir, à travers le prisme de l'espace public du web, à la manière dont se constitue et s'élabore aujourd'hui la mémoire de la Grande Guerre dans des espaces sociaux d'une nature autre, les espaces numériques, lieux de mémoire d'un genre nouveau. Ce travail va nous permettre de voir comment s'articulent mémoire individuelle et mémoire collective dans ce travail en ligne. Il permet de comprendre les frottements et conflits dans la mémoire de l'événement, comment l'histoire officielle dialogue avec l'histoire individuelle. En quoi l'accessibilité des documents en ligne qu'ils soient patrimoniaux ou privés vient à changer les regards sur l'événement et la manière d'en parler ?

2.2 ACTIVITES PROFANES, ROLE DES AMATEURS SUR LE WEB

Cette recherche nous amène à explorer l'implication des non professionnels, des « amateurs », dans ce travail d'histoire, de mémoire et de généalogie. En effet, comme nous le verrons, une grande partie des sites collectés autour de la Grande Guerre est produite par des individus, hors institution. Elle nous amène à explorer le réseau des relations qui se tissent entre ces acteurs, et avec les autres acteurs que peuvent être les historiens professionnels, les organismes ou institutions officielles ou les associations engagées dans la mémoire de la guerre. Elle pose la question du rôle des dispositifs socio-techniques numériques dans la production de connaissance et de mémoire et dans les échanges entre amateurs.

De nombreux travaux ont souligné la montée en visibilité des amateurs dans l'espace numérique (Flichy, 2010) et en particulier leur contribution à la production de savoir et de connaissance. Les enquêtes sur les pratiques culturelles montrent entre 1973 et 2008 une croissance régulière de la pratique amateur qui passe de 11% à 22% (« avoir pratiqué en amateur une activité autre que musicale au cours des 12 derniers mois ») (Donnat, 2011). La croissance serait plus forte encore si la définition incluait les pratiques en ligne, source d'un fort renouvellement dans les activités amateurs. Autrement dit, la tendance de fond de développement des formes d'activités amateur s'est vue renforcée par le développement du numérique (production et mise en ligne de contenus), qui permet aux individus de donner à voir et d'échanger autour de leurs productions, mais aussi d'acquérir collectivement des compétences.

Utilisateurs-producteurs, audiences actives sont autant d'expressions qui soulignent le tournant de la participation, avec l'idée d'un élargissement des publics qui participent à la constitution des savoirs et des biens culturels. Ce tournant brouille les frontières entre production et consommation, entre experts et profanes.

Des recherches ont montré la contribution croissante des publics à la construction des savoirs scientifiques, comme en botanique (Heaton et al., 2011), le rôle joué par

les « amateurs » dans la construction du savoir encyclopédique (Auray et al., 2009 ; Cardon et Levrel, 2009), le rôle joué par les échanges entre malades pour faire évoluer la vision de la médecine (Akrich, 2010) et ce depuis la naissance de la science moderne.

Ce tournant a été rendu possible aussi par une transformation de notre rapport à la « vérité scientifique » : remise en question du rôle des experts et essor d'un relativisme scientifique qui accorde plus d'importance aux sujets de controverse qu'à la vérité (Auray et al., 2009). Il a été facilité par l'émergence de dispositifs socio-techniques qui favorisent les échanges dans l'espace numérique : lieux qui sont à la fois des lieux de publication et de conversation, des lieux de transmission et de construction de connaissance, où la diversité des points de vue peut se faire entendre, comme dans l'encyclopédie coopérative Wikipedia.

Les travaux sur les collectifs en ligne (forums et listes de discussion) ont permis d'identifier les propriétés principales de ces « communautés ». Ce sont des espaces auto-organisés, où se mettent en place des formes de gouvernance distribuée. Contrôle social et encouragement à la participation sont pris en charge par un réseau important de participants actifs qui ne se limite pas aux modérateurs.

Ce sont des espaces doubles : lieux de construction de savoirs **et** de relations. Le terme de communauté épistémique a été utilisé pour montrer comment ces espaces sont des lieux de construction collective de la connaissance. On y trouve à la fois de la transmission de connaissance (sur le mode Question/Réponse) mais aussi de la construction de connaissance au travers de fils de discussion longs (Conein et Latapy, 2008).

Ces espaces numériques se caractérisent par une très grande hétérogénéité de la participation. Les trois quarts du public sont des visiteurs qui n'interviennent jamais. Parmi les participants, on observe une distribution en loi de puissance avec un noyau de participants très actif qui produit l'essentiel des contributions et une longue traine de participants très occasionnels.

Dans ces communautés épistémiques, on note systématiquement une tension entre les nouveaux et les anciens. Comme ce sont des lieux où se construisent un savoir et une culture commune, les écarts avec les nouveaux se cessent de s'accroître. Les nouveaux se sentent mal accueillis, tandis que les anciens critiquent le comportement des novices, qui perturbent le sentiment de l'entre soi. (Arguello et al., 2006 ; Beaudouin et Velkovska, 1999).

A l'aune de ces résultats bien établis, nous pourrons explorer les caractéristiques du forum 14-18 et de son environnement : quelles sont les spécificités de la pratique mémorielle autour de la Grande Guerre ? Quelle est la nature des lieux mémoriaux virtuels que produisent les amateurs ? Comment agencent-ils documents d'archives publiques et privées pour mettre en récit le passé et éventuellement le publier ? Observe-t-on des conflits d'interprétation et des débats dans l'espace du web ?

2.3 DEMARCHE METHODOLOGIQUE

Pour étudier la présence de la Grande Guerre sur le Web, trois points de vue peuvent être adoptés : celui des producteurs (contributeurs), celui des documents et celui des usages (réception et publics) ; le document étant le point d'articulation de la production et de la réception. Dans la première perspective, celle de la production, il s'agit de mener enquête auprès des contributeurs pour comprendre leur activité, le sens qu'ils lui donnent, l'évolution de leurs parcours, les relations sociales nouées au travers de cette activité. Du côté de la réception, on se situe du côté des utilisateurs ou lecteurs, en interrogeant les acteurs sur leurs pratiques, sur ce qu'ils font des documents et des sources numériques, voire en quantifiant les usages avec de la mesure d'audience). Cette distinction est insatisfaisante dans le contexte du Web dans la mesure où ce média se caractérise par une alternance des positions : les participants les plus actifs sont tour à tour récepteurs et producteurs, lecteurs et auteurs. Ce seront donc souvent les mêmes qui seront interrogés dans les deux perspectives de la production et de la réception.

2.3.1 Le point de vue des documents

Le point de vue des documents (ici les sites Web consacrés à la Grande Guerre) se situe à l'intersection de la production et de la réception. Ces sites sont le résultat d'une activité de production et en même temps le point de départ d'une activité de réception, sachant que les interactions entre les deux sont denses. En effet, des éléments de la réception, comme l'audience, les liens hypertextes, les commentaires sont intégrés dans le site et le contenu même du site évolue en fonction de la réception : le niveau de l'audience et les commentaires des visiteurs modifient la manière d'écrire des auteurs des sites (Licoppe et Beaudouin, 2002).

Dans cette perspective centrée sur les documents, les questions qui se posent sont les suivantes : comment délimiter un corpus ? Comment l'archiver ? Comment l'analyser ?

La *délimitation* d'un corpus est une opération délicate. Il s'agit de mettre en œuvre une démarche raisonnée pour découper au sein de l'univers du web une « sphère du web » (Schneider et Foot, 2004), autrement dit un ensemble cohérent de sites le plus souvent interconnectés.

> We conceptualize a web sphere as not simply a collection of web sites, but as a set of dynamically defined digital resources spanning multiple web sites deemed relevant or related to a central event, concept or theme, and often connected by hyperlinks. The boundaries of a web sphere are delimited by a shared topical orientation and a temporal framework. (Schneider et Foot, 2004).

Cette phase est particulièrement complexe : que faire d'un site dont une partie seulement porte sur la thématique choisie ? Que faire avec les sites de réseaux sociaux dont une partie des échanges porte sur le sujet ? Doit-on conserver tous les sites, même les ébauches de sites ? Comment être sûr qu'un pan entier du web ne nous a pas échappé ? Nous avons pris le parti dans ce projet de travailler à partir d'une

collecte mise en place par des conservateurs de la BnF et d'autres institutions, qui ont déterminé les URL de départ de l'archivage ainsi que la profondeur de l'archivage pour distinguer les sites partiellement consacrés à l'archivage des autres. Nous y reviendrons.

La délimitation d'un corpus va de pair avec la constitution d'une *archive* du corpus. Le web étant éphémère, seul *l'archivage* permet de constituer une image fixe à une date donnée. Mais les défis liés à l'archivage sont nombreux. Niels Brügger les résume de la manière suivante : l'archive du web peut être incomplète (éléments non archivés, comme des documents intégrés dans une page qui viennent d'ailleurs) ou trop riche avec une multiplication des copies du même site ; le temps de l'archivage peut créer des « incohérences temporelles » avec un site archivé qui n'a jamais existé (composé de morceaux provenant d'états différents du site) ; les variations dans la profondeur du crawl peuvent créer des « incohérences spatiales » (Brügger, 2013).

2.3.2 Topographie du réseau

La question de *l'analyse* du corpus archivé est loin d'être triviale. Force est de constater que pour l'analyse des corpus du web, différentes méthodologies sont en concurrence et ont du mal à dialoguer entre elles. Pour simplifier le paysage, dans le domaine quantitatif, deux manières de traiter les corpus peuvent être distinguées : l'une explore les relations entre entités et l'autre le contenu de ces entités. L'approche par les réseaux sociaux est particulièrement fréquente : elle redéploie, pour l'étude du web, les outils développés pour l'analyse des réseaux sociaux. Dans cette approche, l'attention est portée sur les relations entre les sites et laisse de côté les propriétés de ces mêmes entités. L'approche par le contenu cherche à analyser le texte des sites par des traitements automatiques. Dans ce cas, c'est la dimension réticulaire qui est perdue.

Nous adoptons ici une approche « analyse de réseau » que nous enrichissons de manière qualitative afin de donner sens à la carte.

Il ne s'agit en aucun cas d'appréhender les parcours des utilisateurs dans l'espace du web (comment un utilisateur navigue de site en site en suivant des liens), mais bien d'identifier les liens entre les sites tels qu'ils ont été inscrits par les responsables éditoriaux du site. Ainsi c'est bien l'ordre construit par les producteurs qui nous occupe ici, et non celui de la réception. Le lien est une des formes modernes de la citation et de la référence : il est à la fois la trace écrite d'une relation et un acte potentiel de mise en relation, d'association et de rapprochement. La présence d'un lien sur un site est le résultat d'une décision du responsable éditorial du site ou de l'auteur d'un texte dans le site qui choisit de citer et de renvoyer, littéralement, vers une autre entité du web. La présence d'un lien peut avoir plusieurs significations : elle peut marquer l'intérêt pour une source ou un document, le simple partage d'un même centre d'intérêt ou même une proximité relationnelle ; elle peut aussi avoir un sens

négatif (critique, rejet, etc.). De l'extérieur, nous sommes dans l'impossibilité de décrire la signification de ce geste pragmatique qui consiste à inscrire un lien (Saemmer, 2015), nous pouvons juste constater sa présence (ou son absence).

Cet exercice de topographie du web s'inscrit dans la ligne des travaux initiés à l'Université Technologique de Compiègne dans le département COSTECH (Ghitalla et al., 2004 ; Kotras, 2015) et reprise et enrichie par le MediaLab à SciencesPo dont le numéro de *Réseaux* « Topographie et topologies » rend bien compte (Boullier et Lévy, 2016). La méthodologie classiquement utilisée comprend selon (Severo et Venturini, 2016) quatre étapes : 1) identification des acteurs principaux ; 2) exploration des liens pour trouver d'autres sites 3) classification des sites par types d'acteurs 4) visualisation et manipulation des graphes.

Comme nous avons pris l'option de travailler sur les archives du web, à partir d'une collecte établie par la BnF, nous avons adapté la démarche et ceci constitue l'originalité de l'approche. L'identification et la classification des acteurs est prise en charge par les acteurs de la collecte, avec quelques échanges avec les chercheurs (qui peuvent transmettre manques, erreurs ou changements sur certains sites…). Notre travail a consisté d'une part à mettre en place une chaine de traitement pour extraire les liens entre les sites archivés et d'autre part à visualiser et manipuler les graphes. Nous avons testé différents paramétrages pour la délimitation et la représentation du graphe dans une double perspective : vérifier la qualité de la collecte en nous assurant de la présence de tous les sites importants et varier les représentations pour identifier les points de stabilité qui nous permettront d'ancrer l'analyse.

L'étude de la structure des liens entrants, indicateurs de réputation, et des liens sortants, indicateurs d'ouverture sur le web, permettra de différencier les profils des sites. Pour les liens entrants, (Cardon, Fouetillou et Roth, 2014) distinguent les notions de *célébrité*, *popularité*, *notoriété* selon que les liens proviennent du territoire ou de l'extérieur. Ils posent ainsi que la célébrité implique beaucoup de liens provenant de l'intérieur et de l'extérieur du territoire, que la popularité est marquée par beaucoup de liens venant de l'extérieur (sans reconnaissance interne) et la notoriété par une forte reconnaissance interne avec peu de citations provenant de l'extérieur. Dans la même logique, les sites seront qualifiés selon la structure et l'intensité des liens sortants : *curieux* s'ils adressent de nombreux liens dans et hors de leur territoire, *extravertis* s'ils privilégient des liens externes, *introvertis*, s'ils privilégient des liens dans le territoire, *silencieux* s'ils n'ont pas ou peu de liens sortants (Cardon, Fouetillou et Roth, 2014).

2.3.3 Entretiens auprès des acteurs et analyse des sites

Cette démarche d'analyse de réseau a été enrichie par une exploration qualitative des sites visibles sur la carte et par les entretiens menés auprès d'acteurs qui ont été

amenés à réagir face à la carte, à en donner une interprétation et à en pointer les limites. Douze entretiens dont la durée a varié entre 1h et 3h ont été réalisés[2]. Les premiers contacts ont été établis par messagerie privée via le forum, à l'initiative de Nicolas Auray qui devait initialement mener cette partie de l'enquête avant son décès tragique. On obtenait très peu de réponses, même avec relances. Les contacts établis par mail à partir des sites sélectionnés donnaient très peu de résultats. L'accès au terrain était particulièrement difficile. Quelques contacts ont eu lieu en réactivant des liens établis lors de la précédente enquête menée par Philippe Chevallier et Muriel Amar, *Les usages des corpus numérisés de Gallica sur la Grande Guerre* (Amar et Chevallier, 2013). C'est grâce au soutien du fondateur du forum et des modérateurs, que nous remercions chaleureusement, que nous avons pu fixer des rendez-vous. J'ai publié un message public dans le forum sollicitant la participation à un entretien, et ce message, cautionné par deux modérateurs, a conduit une vingtaine de personnes à donner leur accord. Les profils rencontrés sont diversifiés : des membres anciens et actifs sur le forum - y compris des modérateurs - et des membres beaucoup plus récents – des lecteurs du forum qui n'interviennent pas forcément beaucoup ; des membres qui ont créé des sites web dédiés à la Grande Guerre et d'autres qui publient leur recherche sur le forum ou qui les gardent pour eux ou pour leur cercle proche ; et des personnes qui sont très actives dans d'autres sites collaboratifs du web Grande Guerre comme Memorial GenWeb ou Crid 14-18.

Par manque de temps, nous n'avons pas pu rencontrer toutes les personnes qui avaient accepté de participer aux entretiens, ce que nous regrettons d'autant plus que tous les entretiens que nous avons faits ont tous été d'une très grande richesse. Le tableau suivant précise les personnes rencontrées.

Initiales utilisées	Activité	Rôle sur le web autour de la Grande Guerre
YV	Traducteur	Participant Forum
JC	Enseignant	Modérateur Forum, auteur site
ADC	Retraité	Participant Forum
AD	Documentariste-Vidéaste	Participante Forum, auteur site
IM	Retraité, ex militaire	Participant Forum
SA	Dessinateur BD	Participant Forum, auteur site
CM	Professeur des écoles	Participante Forum
AG	Retraité	Participant Forum et Memorial GenWeb
GG	Retraité	Participant Forum
GL et MB	Universitaire et archiviste	Auteurs site
EP	Universitaire	Site association

Tableau 1. Profil des personnes interviewées

La grille d'entretien comprenait quatre temps, le premier portait sur les formes des contributions actuelles autour de la Grande Guerre, le deuxième était centré sur la

[2] Sur les douze entretiens, 9 ont été réalisés par Valérie Beaudouin, 3 par Philippe Chevallier. 5 ont été réalisés à domicile, ce qui permet d'appréhender l'univers de travail, 5 en face à face dans un lieu public et 2 par téléphone.

manière d'utiliser les sources numérisées, le troisième était un retour arrière sur les origines de l'engagement et son évolution et le quatrième temps portait sur la confrontation à la carte du web Grande Guerre.

Les entretiens ont été retranscrits et analysés. Dans ce rapport, sont principalement utilisés les éléments concernant l'activité de production de sites web et les réactions face à la carte.

2.3.4 Questions de recherche

Afin de pouvoir identifier les modes de circulation des documents provenant des collections numérisées, il nous a paru indispensable d'explorer le web en identifiant les sites liés à la Grande Guerre et en représentant le réseau de relations entre ces derniers. Les proximités entre sites reflètent le réseau relationnel entre les acteurs.

Cette carte, une fois élaborée, nous permet d'identifier les acteurs impliqués dans la commémoration (individus, associations, institutions, médias), les relations qui se nouent entre eux et en particulier la place qu'y occupent les institutions patrimoniales.

Cette recherche se situe précisément du côté des producteurs, des acteurs qui produisent ou publient des documents sur la Grande Guerre dans l'espace public du Web et qui ce faisant s'inscrivent dans un réseau de relations. Elle cherche à répondre aux questions suivantes :

- Comment s'organise l'espace du web consacré à la Grande guerre ? Peut-on identifier des sous-espaces dans le réseau, caractérisés par une plus grande densité de connexions ?
- Quels sont les acteurs impliqués dans la présence de la Grande Guerre sur le Web : des amateurs, des chercheurs, des associations, des institutions, des médias… ? Qui est relié à qui ? Quelles sont leurs compétences ?
- Quelle est la place des fonds patrimoniaux des bibliothèques numériques (Gallica, L'Argonnaute, etc.) dans cet environnement ?

Ce faisant, cette exploration exclut l'analyse des réseaux sociaux, qui font l'objet d'études à part entière. Frédéric Clavert, par exemple, analyse les flux de conversation sur Twitter en lien avec des hashtag liés au conflit dans un environnement international, ce qui lui permet de comparer les échanges en France et en Angleterre, le rythme des échanges en fonction des moments de la commémoration (Clavert, 2016)…

La première section porte sur la constitution du corpus de sites liés à la Grande Guerre, corpus qui a fait l'objet d'une collecte spéciale par le service du dépôt légal numérique de la BnF (DLWeb) ; la seconde section présente la méthodologie mise en place tandis que la dernière propose une exploration et interprétation de la carte nourrie par les approches qualitatives. Ceci nous permettra d'explorer les formes que

prennent les activités de publication en ligne autour de la mémoire de la Grande Guerre, et en particulier comment se déploient sur le web des monuments mémoriels contemporains d'un genre nouveau, qui inscrits dans un réseau dense de relations dessinent une modalité originale de constitution d'une mémoire collective.

3 CONSTITUTION D'UN CORPUS : LA COLLECTE GRANDE GUERRE SUR LE WEB DE LA BNF (DLWEB)

L'équipe du dépôt légal du web (DLWeb) de la BnF organise une collecte globale du domaine français, ainsi que des collectes spécialisées (Bonnel et Oury, 2014). En lien avec la commémoration du Centenaire et la Grande Collecte mise en place pour numériser des archives personnelles, la BnF a mis en place une collecte spécifique « Collecte Grande Guerre sur le Web »

3.1 CONSTITUTION DE LA COLLECTE

Cette collecte est pilotée par Agnès Sandras avec Peter Stirling. Elle comprend une phase de sélection et de caractérisation des sources puis une phase de collecte et d'archivage des documents. La sélection est établie par la BnF (départements Philosophie, histoire, sciences de l'homme et Cartes et plans) et ses partenaires (Bnu de Strasbourg, DMPA du ministère de la défense) : une équipe de conservateurs et d'experts est en charge de la sélection. La collecte et l'archivage sont pris en charge par le DLWeb. Nous reprenons les éléments de description de la collecte fournis par Peter Stirling. Une note interne de la BnF « Collecte web de la guerre 14-18 : organisation et politique documentaire » signée de Clément Oury explique la stratégie mise en place pour cette collecte.

L'archivage peut porter sur un site entier ou bien sur une sous-partie d'un site, s'il n'est pas entièrement dédié à la guerre 14-18.

Les sites retenus sont de nature diverse : il peut s'agir de sites personnels, de blogs, de portails, de sites entrepôts de documents…

Les sites archivés sont classés en fonction de leur émetteur. Le choix a été fait d'utiliser le champ « thème » pour caractériser l'émetteur, en réservant le champ « mots-clés » pour la description du contenu[3]. Cinq types d'émetteurs sont distingués *a priori* :

- Officiel : international, national, territorial, européen
- Public : patrimonial, pédagogique, scientifique
- Personnel : dont militaires
- Associatif : dont militaires

[3] Ce champ mot-clés n'a pu être exploité en raison de la trop grande richesse des descriptifs.

– Médias

Comme toute catégorisation, celle-ci pose certaines difficultés. Nous l'avons adoptée en première instance et les résultats nous ont amenées à introduire quelques modifications. La distinction entre les sites « officiels » et « publics » ne nous a pas toujours parue pertinente : nous avons donc regroupé les sites sous l'appellation sites « institutionnels ». De même la catégorie d'un site a pu être modifiée. Par exemple le site Crid 14-18, classé initialement en site public, a été reclassé en site d'association, ce qui correspond plus précisément à son statut.

La base Bcweb (BnF Collecte du Web) comprend les éléments de description des sites ou entités éditoriales retenues pour la collecte :

- URL de départ
- Paramètres (profondeur)
- Thème (émetteur)
- Mots-clés

La liste des sites à collecter a été constituée progressivement. Si les sites institutionnels, en particulier ceux liés à la commémoration, étaient repérés dès la première collecte, l'ajout des sites personnels et associatifs s'est fait progressivement. L'apparition d'un nouveau site dans la collecte ne signifie donc pas forcément que le site vient d'être créé.

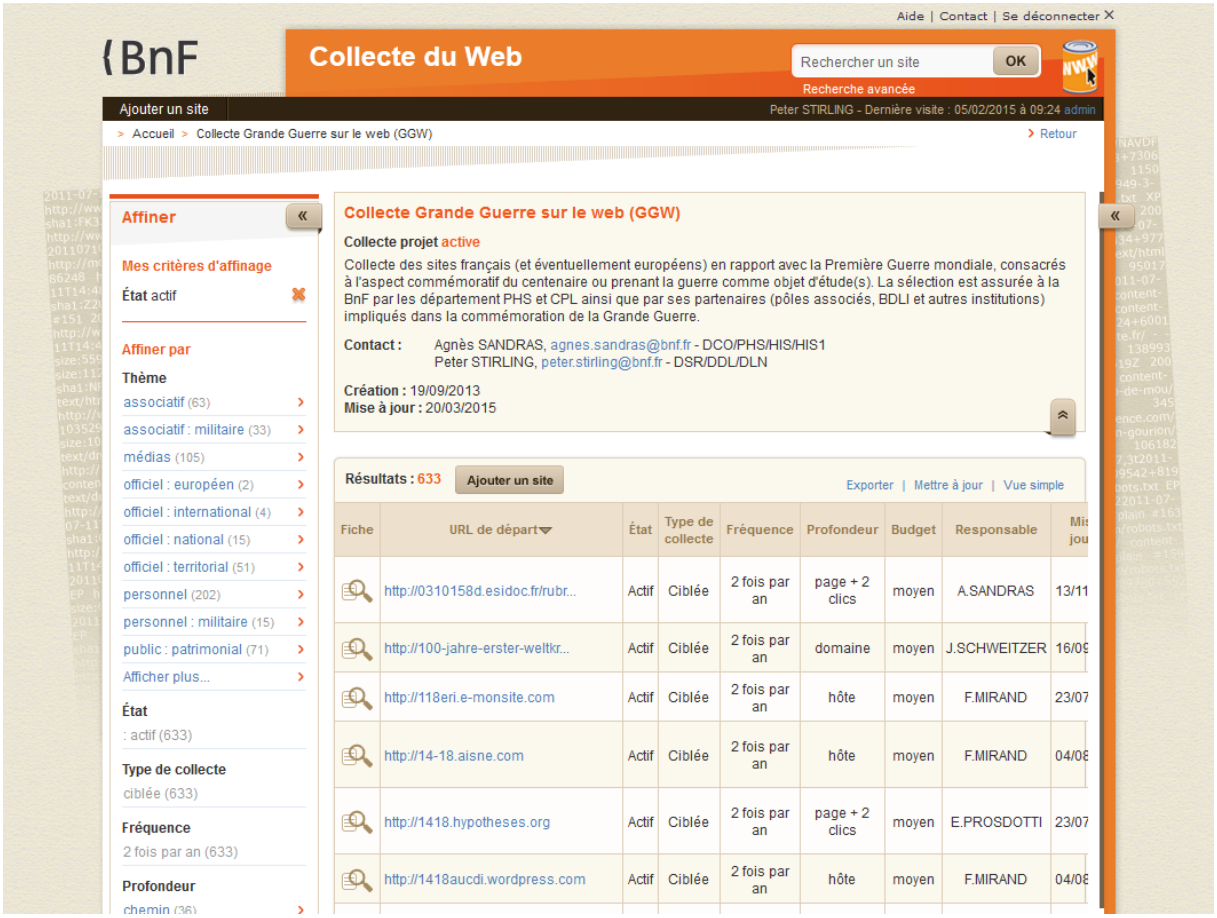


Figure 1. Interface de BCWeb

La collecte de la Grande guerre a commencé en novembre 2013 et s'étendra jusqu'en 2019 avec deux à trois collectes par an. Seules les quatre premières collectes font l'objet d'une analyse dans le cadre de ce projet.

Date collecte	Nombre de sites	Nombre d'URL collectées	Volume
Novembre 2013	41 sites	1 198 723	35,75 Go
Mars 2014	99 sites	1 689 613	47,29 Go
Août 2014	482 sites	7 323 201	402,83 Go
Novembre 2014	555 sites	9 698 633	313,85 Go

Tableau 2. Description des collectes

Les archives de l'internet contiennent des copies des sites tels qu'ils ont été collectés par des robots durant un temps plus ou moins court. Tous les fichiers, dans le champ défini par le paramétrage à partir des URL de départ, sont stockés dans un format « conteneur », WARC. Une interface d'accès permet de restituer les sites tels qu'ils étaient au moment de l'archivage, sous réserve d'absence de problèmes techniques (absence de réponse du site, éléments indisponibles, etc.) et sans la présence des éléments externes appelés, comme les publicités.

Le processus global d'archivage du web est présenté dans (Le Follic, Stirling et Wendland, 2012).

3.2 EXPLORATION DES ARCHIVES DE LA COLLECTE

Le DLWeb possède différentes sources liées à la collecte Grande Guerre comme détaillé dans la section « La collecte Grande guerre par la BNF ». Cette section porte sur les sources et leur utilisation pour cartographier le corpus.

1. BCWeb : C'est la liste des urls choisies par les experts pour démarrer la collecte. Les robots utilisent ces liens pour commencer à « crawler » et suivre les liens sortants pour construire le corpus en fonction de critères prédéfinis (profondeur etc.). Cette liste est fondamentale pour notre travail car elle contient les urls fiables et des éléments de qualification des URLs comme le « thème » (type d'émetteur) ou les mots-clés etc.

2. ARC /WARC[4] : Il s'agit des documents archivés par les robots (« crawler »), qui sont une forme de « copie » des sites. Dans ce projet, qui se limite à cartographier les sites et liens entre sites, le contenu des sites web (images, texte etc) n'a pas été exploité.

3. DAT /WAT[5] : Les fichiers DAT/WAT sont des extraits des fichiers ARC /WARC, contenant toutes les métadonnées relatives aux fichiers archivés (certaines provenant du processus de collecte, d'autres des documents archivés). Comme les liens hypertextes font partie de ces métadonnées, on peut se contenter d'utiliser ces fichiers DAT/WAT, sans avoir besoin d'extraire les données à partir des fichiers ARC /WARC.

4. Crawl Logs : Ces fichiers gardent les traces de l'activité des robots pendant le processus de crawl. Pour une page archivée x, on peut retrouver à partir de quelle url de départ dans BCWeb, le robot a trouvé la page x. Après avoir étudié cette source en détail, nous avons considéré qu'elle n'était pas pertinente pour cartographier les relations entre les sites. En revanche, elle est très utile pour comprendre le processus d'exploration des archives.

Pour conclure, les fichiers DAT /WAT seront utilisés pour cartographier le corpus et BCWeb pour enrichir la cartographie.

[4] ARC/WARC : http://www.bnf.fr/fr/professionnels/dlweb_boite_outils/a.dlweb_formats_fichiers.html.
[5] DAT/WAT : https://webarchive.jira.com/wiki/display/ARS/WAT+Overview+and+Technical+Details.

4 METHODE D'ANALYSE DE LA COLLECTE GRANDE GUERRE

La dimension fondamentalement réticulaire du web fait que la représentation sous forme de graphe des relations entre les sites s'est rapidement imposée comme un moyen de découvrir les marquages intentionnels des acteurs. L'objectif est de cartographier les relations entre les sites ou entités du corpus telles qu'elles sont exprimées via l'usage des liens hypertextes.

La cartographie permet de visualiser les proximités et distances entre les différentes entités du corpus et de représenter les chemins potentiellement offerts aux visiteurs.

D'un point de vue pratique, il s'agit de créer une représentation du graphe des liens hypertextes présents sur l'ensemble des pages web du corpus. Ce corpus peut être constitué à partir du web « vivant » ou à partir des archives du web.

La démarche que nous avons adoptée s'appuie sur les archives du web de la BnF présentées dans la section précédente. La chaine de traitement proposée a été conçue de façon générique pour pouvoir s'appliquer à d'autres collections issues des archives du Web de la BnF et pourra donc être utilisée par d'autres générations de chercheurs.

Ainsi, l'étape de cartographie de ce projet représente un travail pionnier pour la BnF qui souhaite offrir un service de fouille de données (textes et liens) à ses utilisateurs des archives du web.

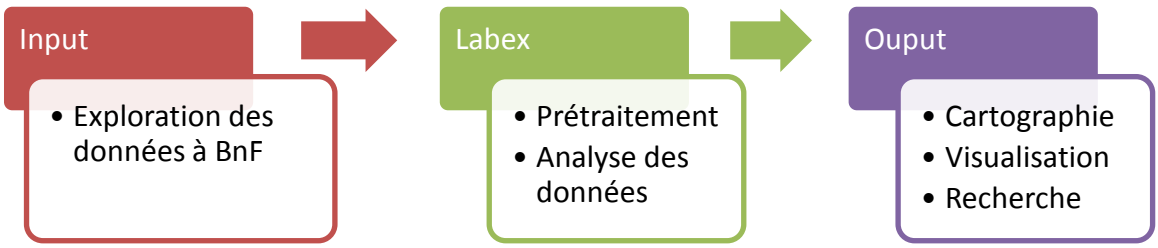


Figure 2. Flux de données

Après avoir rappelé des éléments de terminologie sur la cartographie, nous présenterons les fichiers qui correspondent aux archives du web avant de présenter la chaine de traitement. Enfin sera exposé le problème principal rencontré qui tient au « bruit » dans les données ainsi que la stratégie mise en place pour y remédier.

4.1 CARTOGRAPHIE : TERMINOLOGIE

Un graphe est construit à partir de deux objets : une liste des acteurs (dans notre cas les sites) composant le réseau (les nœuds, notés V) et une liste des relations entre ces acteurs (les liens, notés E). Le label correspond au nom du nœud donc de l'acteur. La Figure 3 est la représentation d'un graphe de 5 nœuds V=(a,b,c,d,e) avec les liens entre eux E=((a,b),(b,a),(e,a),(d,a)…). Un graphe est dit orienté si tous les liens ont une direction (le graphe de la Figure 3 est orienté). Un graphe est pondéré si une fréquence (ou poids) est affectée à chaque lien.

Nous avons opté pour des graphes orientés, pour distinguer les liens entrants et sortants, et pour des graphes non pondérés pour éviter certains biais liés à la norme de construction des sites (par exemple, sur chaque page du site du Centenaire, il y a un pied de page avec des liens sortants vers Facebook, Twitter…).

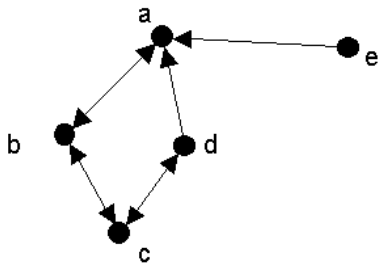
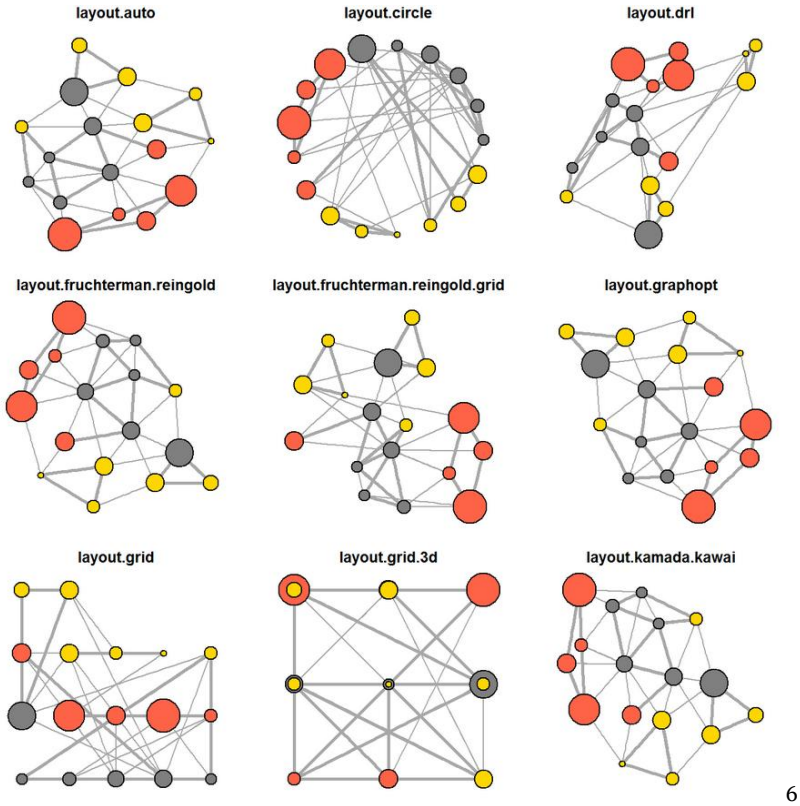


Figure 3. Exemple d'un graphe

Une carte est une représentation dans l'espace du graphe. Pour visualiser les graphes, on utilise des algorithmes de spatialisation (ex : circle, frucherman reingold, grid, spring, force atlas etc.) qui visent à positionner les nœuds en fonction de leurs relations. Un même graphe peut donc avoir plusieurs représentations comme cela est selon l'algorithme de spatialisation utilisé, comme cela est illustré dans la Figure 4.

Nous avons opté pour les algorithmes Force Atlas ou Force Atlas 2 qui après différents essais nous paraissent les plus performants pour différencier des sous-espaces interconnectés. Force Atlas 2 est la version améliorée de Force Atlas dans la manière de traiter les calculs. Son principe est détaillé dans l'article (Jacomy et al., 2012) :

> « Les nœuds se repoussent comme des aimants, tandis que les liens attirent les nœuds qu'ils connectent, tels des ressorts Ces forces mettent les nœuds en mouvement, jusqu'à ce qu'un point d'équilibre soit atteint (les nœuds ne bougent plus). La spécificité de Force Atlas 2 c'est que la force de répulsion est modifiée de façon à ce que les nœuds très connectés et leurs nœuds « satellites » peu connectés se repoussent peu. »



[6]

Figure 4. Spatialisation des graphes

Les paramètres suivants ont été utilisés dans Force Atlas 2 et/ou Force Atlas:

- Force Atlas 2 - Echelle =30 : Contrôle l'intensité des forces de répulsion et d'attraction entre nœuds
- Force Atlas - Force de répulsion = 800
- Force Atlas - Force d'attraction = 3
- Force Atlas - Vitesse = 50

Encadré : Propriétés des graphes

Degré : Le degré d'un nœud correspond au nombre total de liens entre ce nœud et ses voisins.

Degré entrant/sortant : Dans le cas d'un graphe orienté, le degré entrant d'un nœud est le nombre de liens qui aboutissent à ce nœud, et le degré sortant le nombre de liens qui partent du noeud. Dans la Figure 3, le nœud *a* est de degré 4 , avec un degré sortant de 1 et un degré entrant de 3.

[6] http://kateto.net/network-visualization

Degré pondéré : si le degré mesure la simple existence de liens entre un nœud et ses voisins, elle ne mesure pas l'intensité des relations entre les nœuds. Le degré pondéré prend en compte la fréquence des liens entre les nœuds.

Degré de centralité : Cette mesure nous permet d'obtenir l'activité du nœud étudié. Elle est définie par le rapport entre le degré du nœud et le degré maximal possible du nœud (si ce dernier était relié à tous les nœuds du graphe) [7].

Centralité d'intermédiarité (« Betweenness centrality ») : le degré de centralité ne prend en compte que les liens de voisinage direct et ne prend pas en compte les relations indirectes. La mesure de « betweenness centrality » (Freeman, 1977) n'a pas cet inconvénient et favorise les nœuds qui établissent des ponts entre des communautés plutôt que les nœuds marquants de chaque communauté.

Algorithme HITS (Hyperlink-Induced Topic Search) : cet algorithme proposé par Kleinberg est bien résumé dans l'article qui vise à la détection de communautés en ligne (Gibson, Kleinberg et Raghavan, 1998). Soit un corpus de pages web sur une thématique donnée. Deux scores sont attribués à chaque page : un de concentrateur (« hub ») et un d'autorité (« authority ») définis l'un par l'autre dans un système itératif et récursif. Le principe de base est qu'une page qui fait autorité est citée par beaucoup de concentrateurs alors qu'une page qui joue le rôle de concentrateur pointe vers beaucoup de pages qui jouent le rôle d'autorité. Le score de l'autorité correspond à la somme des scores des concentrateurs pointant vers cette page, et le score d'un concentrateur correspond à la somme des scores des autorités vers lesquelles il pointe.

4.1.1 Fichiers de graphes

Tous les fichiers de graphes sont constitués de deux sous-parties : la première donne la liste des nœuds (caractérisés par leur nom et des variables descriptives) tandis que la seconde fournit la liste des liens entre les nœuds avec leurs attributs. Les graphes peuvent être sauvegardés dans différents formats : Csv, graphml, Gexf etc.

Gexf (Graph Exchange XML Format)[8] est le format le plus récent et le plus riche. Il permet aussi de déclarer des attributs dynamiques, qui permettent de visualiser les changements au fil du temps. Comme, dans le contexte de notre projet, des collectes ont été effectuées à différentes dates, il nous paraissait utile d'avoir un format qui puisse rendre compte de la dimension temporelle. On a donc choisi d'utiliser le format gexf pour sauvegarder les graphes.

[7] https://fr.wikipedia.org/wiki/Analyse_des_r%C3%A9seaux_sociaux#Degr.C3.A9_de_Centralit.C3.A9
[8] http://gexf.net/format/

4.1.2 Outils de visualisation

Il existe différents outils de visualisation des graphes. Certains nécessitent une compétence de programmation (R, python etc.), ce qui n'est pas très fréquent chez les chercheurs en sciences sociales. On s'est restreint aux logiciels qui permettent de manipuler/ visualiser les graphes sans programmation. Il faut bien souligner qu'il est impossible de connaitre et de maitriser tous les logiciels, qui évoluent en permanence. L. Beauguitte a regroupé les logiciels de la façon suivante[9] :

- Des logiciels pour débuter : Ucinet, NodeXL et Visone
- Des logiciels pour de grands réseaux : Cytoscape, Gephi et Tulip

Comme nous travaillons sur des données à grande échelle, nous n'avons exploré que les logiciels pour grands réseaux. Tulip est développé depuis 2001 par le LABRI (Bordeaux) en collaboration avec des géographes : c'est un outil en cours de développement. Cytoscape est développé depuis 2003 : il est conçu pour l'analyse de réseaux biologiques. Gephi est développé depuis 2009 par un consortium français (Sciences-Po MediaLab, ISC Paris etc.), mais l'équipe a arrêté le développement depuis environ deux ans, pour se consacrer au développement de nouveaux outils qui permettent de constituer puis de visualiser un corpus du web. Malgré tout, Gephi reste le logiciel le plus utilisé dans le monde entier.

La tendance actuelle est d'aller vers des outils de visualisation sur le web (plutôt que des logiciels « desktop ») qui permettent de partager les graphes : ManyLines développé par le Medialab (http://tools.medialab.sciences-po.fr/manylines), Polinode (https://www.polinode.com/) etc. Malheureusement ces outils ne peuvent visualiser qu'un nombre limité des nœuds à cause des limites de capacité des navigateurs (ex : maximum 10 000 nœuds pour ManyLines).

Tous ces outils peuvent lire les formats de graphe Gexf.

Nous avons principalement utilisé Gephi dans ce projet, en raison de la taille du réseau et de notre connaissance antérieure de l'outil[10].

[9]Les principaux logiciels pour l'analyse de réseau, Laurent Beauguitte, CNRS, UMR IDEES, http://quanti.hypotheses.org/files/2013/06/quantilille-reseaux-02-logiciels-beauguitte.pdf

[10] Utilisation de Gephi dans des projets antérieurs (Beaudouin, 2012) et liens avec le MediaLab (Dime-SHS).

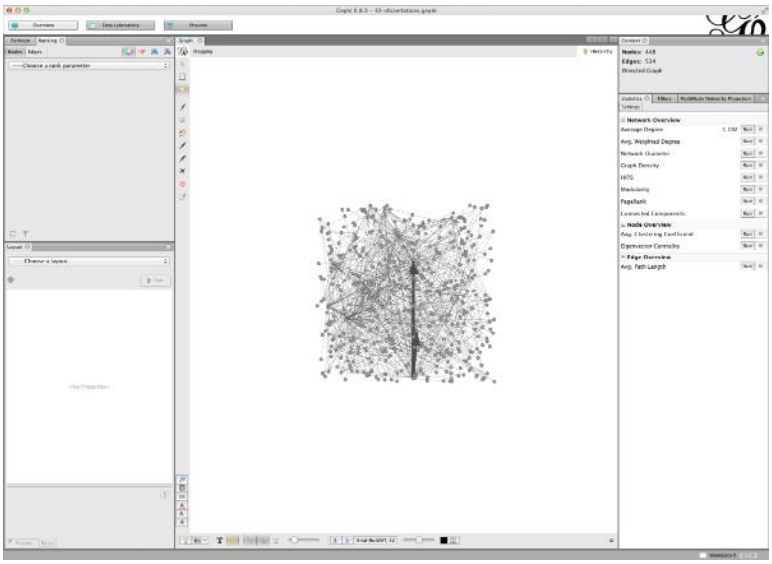


Figure 5. L'interface de Gephi

A partir du fichier du graphe (Gexf dans notre cas), l'outil de visualisation (Gephi) permet :

- D'optimiser la spatialisation des nœuds pour rendre la carte plus lisible ;
- D'adapter la taille des nœuds en fonction de critères choisis par le chercheur (degré entrants, sortants…) ;
- D'associer des couleurs aux nœuds et aux liens en fonction de leurs attributs (par exemple les sites faits par des amateurs seront représentés en bleu) ;
- D'utiliser des filtres pour explorer les graphes (conserver les nœuds avec un degré supérieur à n) ;
- D'appliquer des algorithmes de classification de graphes.

4.2 Chaine de traitement « Labex-Cartographie »

Afin de pouvoir générer des graphes à partir de sources sélectionnées, on a mis en place une chaine de traitement, Labex-Cartographie, détaillée dans la Figure 6.

Figure 6. Framework Labex-Cartographie

4.2.1 Extraction des données

Cette étape consiste à prétraiter les données afin de les préparer pour générer les graphes dans les prochaines étapes. A partir des fichiers de métadonnées (WAT) fournis par BnF (qui sont très riches), on extrait les champs qui nous intéressent :

- Url : l'Url de la page archivée par le robot
- Statut : réponse http (ex :404 – Not found, 200 ok)
- Type Mime de l'Url : les types MIME constituent une façon normalisée de classer les différents types de fichiers sur Internet. (ex : text/html, application/pdf)
- Liens sortants : Liste des liens hypertextes
- Type des liens sortants : catégorisation de chaque lien hypertexte (anchor tags (A), image tags (IMG) etc.)
- Date : Date du crawl

Ces données sont intégrées dans une base de données structurée pour éviter d'avoir à refaire l'extraction (très longue) chaque fois que l'on génère un graphe.

4.2.2 Base de données

Deux solutions différentes ont été testées pour la base de données : NoSql ou utilisation de fichiers plats Hadoop/PigLatin.

NoSql

Les bases de données NoSql sont beaucoup plus flexibles que les bases de données relationnelles : elles s'adaptent beaucoup plus facilement aux changements dans les structures des bases de données. MongoDb est une base de données NoSql de type document qui permet très facilement d'insérer et d'extraire des documents et qui comme toute base de données permet de faire des requêtes classiques (ex : group by , count etc.). Donc, on a mis en place un système qui, à partir des fichiers WAT et Bcweb, extrait les données et les insère dans la base MongoDB en utilisant python.

On a testé MongoDB sur la collecte Novembre 2013 qui a une taille modeste (1 198 723 URL). Les résultats et la manipulation étaient corrects. En revanche sur la collecte d'août 2014 (7 323 201 Url collectées qui pointent vers 240 089 640 Url), on a eu des problèmes de performance insolubles. Comme MongoDB est mono-tâche, il ne distribue pas les calculs et le traitement des données devient un goulet d'étranglement très rapidement. En raison des difficultés rencontrées, nous avons exploré une autre piste et renoncé à MongoDB.

Hadoop/Pig

L'utilisation de Hadoop /Pig est fréquente dans la communauté d'archivage du web. Internet Archive propose différents scripts open-source pour explorer les archives du web[11].

Hadoop est un environnement de développement pour concevoir des applications distribuées à grande échelle. C'est un projet Open Source géré par Apache Software Fundation basé sur le principe Map Reduce[12] et de Google File System. Apache Pig est une plateforme qui permet de créer des programmes pour Hadoop dans un langage procédural appelé *Pig Latin*. Pig est une alternative à Java, dédiée à la création de solutions *MapReduce*.

L'extraction des données et l'agrégation des nœuds sont réalisées en utilisant l'environnement Hadoop/pig. Les fichiers construits à cette étape sont utilisés pour générer le graphe à l'étape suivante.

4.2.3 Générer le graphe et le visualiser

Pour proposer une solution générique de construction des cartographies, applicables à différentes collections d'archives, en fonction des besoins des chercheurs, une interface générique a été développée pour la génération de graphes (Figure 7). Les chercheurs peuvent construire leurs graphes en fonction de leur besoins en utilisant cette interface.

[11] https://github.com/vinaygoel/archive-analysis qui est utilisé aussi dans le projet d'Ian Milligan https://github.com/lintool/warcbase/wiki

[12] Le paradigme **MapReduce** permet de traiter des volumes importants de données en parallélisant les tâches. Il a donné la naissance à Hadoop.

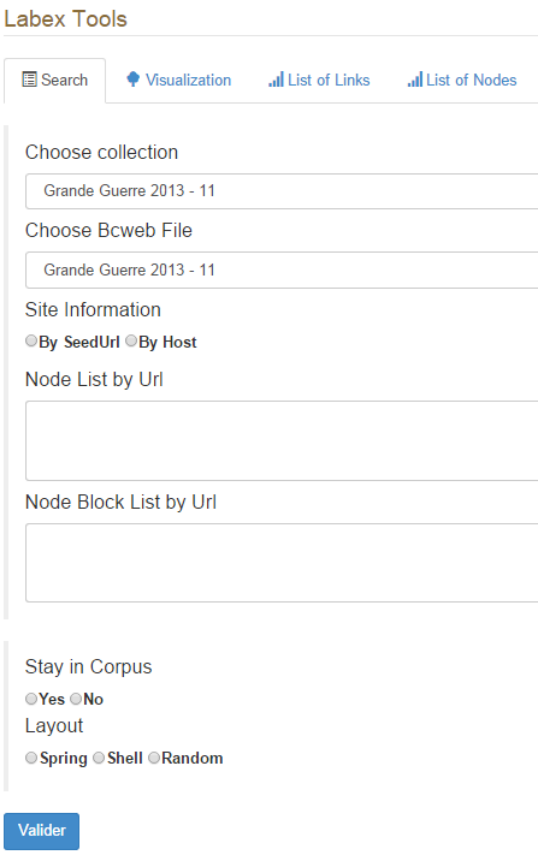


Figure 7. Interface de requête

On peut choisir la collecte sur laquelle on veut travailler, les nœuds (URL) qu'on veut inclure ou exclure des graphes, les types de liens ainsi que la forme de spatialisation du graphe. Une fois que les choix sont faits, le système génère le graphe et ouvre la page de visualisation où on peut manipuler (changer la taille des nœuds, les couleurs, etc.) et visualiser le graphe (Figure 8).

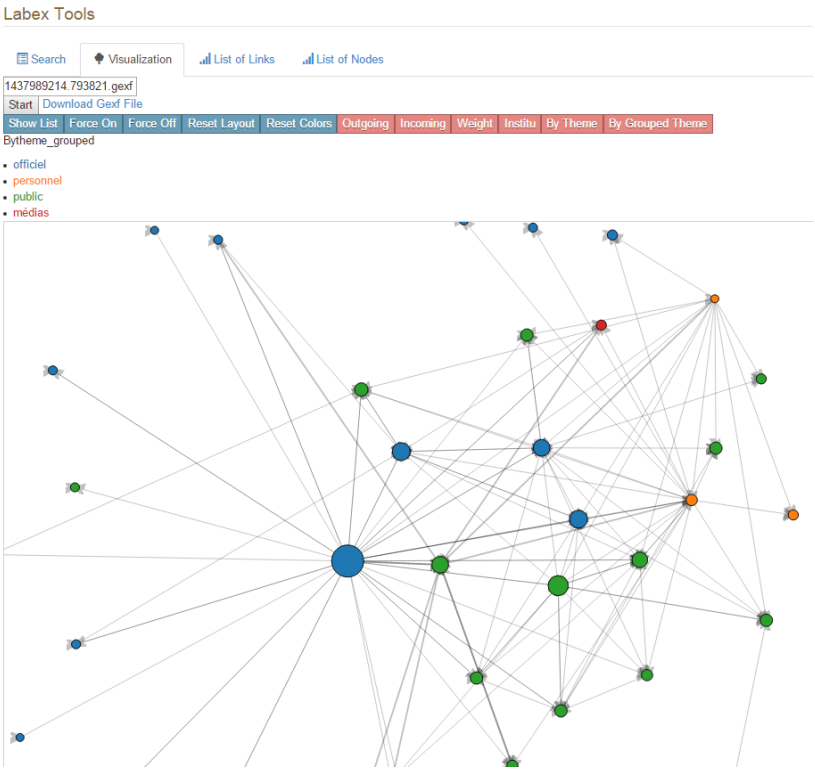


Figure 8. Interface de visualisation

Si l'utilisateur souhaite utiliser un autre outil comme Gephi, Tulipe etc. (par exemple si le graphe est trop grand), il peut télécharger le graphe dans le format gexf compatible avec tous les logiciels.

4.3 PROBLEMES RENCONTRES ET SOLUTION

4.3.1 Problèmes rencontrés

Pendant la mise en place de la chaine de traitement, deux problèmes principaux ont été rencontrés, portant l'un sur la quantité et l'autre sur la qualité des données.

La quantité des données :

Le problème majeur de la visualisation des graphes tient à leur lisibilité. Au-delà d'une certaine densité, ils deviennent ininterprétables. Or, avec les archives du web on arrive très rapidement à des volumes de nœuds et de liens gigantesques. Par exemple, la collecte « Grande Guerre sur le web » datée d'août 2014 contient environ 240 millions d'Urls, ce qui est tout à fait ingérable pour un humain. Il faut simplifier les données pour des raisons de lisibilité et d'interprétabilité.

La qualité des données :

Les robots de crawl prennent une liste d'Urls comme point de départ (BCWeb) et construisent le corpus en respectant les critères de configuration (fréquence de la

collecte et profondeur). Comme nous l'avons vu, les points de départ sont fiables : ils concernent tous la Grande Guerre. En revanche, deux facteurs introduisent du « bruit » dans les cartographies :

- La procédure de crawl peut conduire à archiver des pages qui n'ont plus rien à voir avec la guerre 14-18. C'est en particulier le cas pour les sites qui n'ont qu'une section dédiée à la Première Guerre mondiale : le processus de crawl peine à ne pas sortir au-delà de la section.
- La structure des sites web favorise l'usage des liens hypertextes, dont une partie peut être étrangère au thème collecté. Ces liens, qu'ils soient ou non archivés, figurent dans la liste des liens sortants d'une page et donc potentiellement sur la cartographie.

Prenons l'exemple du Monde (Figure 9) qui a construit un dossier dédié à la Grande guerre : http://www.lemonde.fr/Centenaire14-18. Le point de départ est fiable, mais le robot qui suit les liens pour construire le corpus ou les liens sortants sur la page va introduire un grand nombre d'url qui ne concernent pas la guerre 14-18, comme http://www.lemonde.fr/societe/ etc. Ce sont tous les liens que le Monde propose dans chaque page vers les autres rubriques et articles du site. Pour pouvoir construire une cartographie correcte, il faut pouvoir éliminer ce bruit. D'autant que pour remonter dans la page d'accueil du Monde, les liens sont proliférants (notamment chez les éditeurs qui cherchent à améliorer leur référencement).

Figure 9. Exemple de bruit dans le corpus

4.3.2 Stratégies

On a exploré trois stratégies afin de rendre les graphes plus lisibles et interprétables, qui sont étudiées en détail dans cette section.

Agrégation

Pour rendre les graphes à grande échelle lisibles, la méthode habituellement utilisée est l'« agrégation des nœuds ». Il s'agit de grouper les nœuds qui partagent les mêmes attributs et de les présenter sous forme d'un seul nœud dans le graphe.

Dans notre corpus, cette agrégation consiste à regrouper les URL selon trois techniques.

Pour illustrer, supposons qu'on ait un graphe avec trois nœuds :

- Noeud1 : http://www.lemonde.fr/centenaire-14-18/
- Nœud2 :http://www.lemonde.fr/europe/article/2015/09/07/l-allemagne-fait-une-haie-d-honneur-aux-migrants_4747957_3214.html
- Noeud3 : https://www.facebook.com/lemonde.fr

Noeud1 représente le point de départ et les deux autres nœuds proviennent soit de la capture par le robot, soit sont présents comme lien hypertexte.

Agrégation par Crawllog : Cette méthode agrège les nœuds au point de départ du robot en utilisant le chemin suivi par le robot qui est enregistré dans les fichiers des logs (crawllog). En utilisant cette méthode, tous les nœuds dans notre exemple seront agrégés dans le Noeud1. On a renoncé à cette approche car le niveau d'agrégation est parfois trop élevé : ici une page Facebook est intégrée dans une page du Monde. Par ailleurs dans les crawllogs, il n'y a que l'information correspondant aux pages archivées et donc il n'y a pas les liens hypertextes sortants.

Agrégation par Hôte[13] : Une autre méthode, utilisée souvent pour les graphes du web, est de détecter l'hôte de la page et d'agréger les Urls par rapport à celui-ci. Chaque nœud dans le graphe représente alors un hôte spécifique. Dans notre exemple, les Noeud1 et Noeud2 sont agrégés au nœud lemonde.fr et le nœud 3 à facebook.com.

Agrégation par l'Url de départ : La dernière méthode consiste à utiliser les Url de départ dans BCWeb pour agréger les nœuds, puisque ce sont des points fiables. Les Urls qui commencent par l'Url de départ sont groupées dans un seul nœud. Si l'Url ne correspond à aucune Url de départ, on utilise la méthode « agrégation par hôte ». Pour l'exemple donné, on obtient trois nœuds :

- lemonde.fr/centenaire-14-18/
- lemonde.fr

[13] Dans le contexte d'Internet, un nom d'hôte est le nom d'un sous-domaine. Exemple : pour Url suivant : http://video.google.co.uk:80/videoplay?docid=-7246927612831078230&hl=en#00h02m30s, video.google.co.uk est le nom d'hôte alors que le domaine est google.co.uk.

– facebook.com.

Quel est l'impact de ces méthodes d'agrégation sur les liens ? Les liens des nœuds d'origine sont simplement rattachés au nœud agrégé.

Filtrage

Pour réduire ou supprimer le bruit dans le corpus, on peut aussi utiliser des techniques de filtrage sur le contenu.

L'approche la plus élégante aurait consisté à analyser le contenu des pages pour déterminer si elles relèvent ou non du domaine de la Première Guerre mondiale. Un modèle d'apprentissage aurait pu être mis en place pour ne conserver que les pages qui portent sur la Grande Guerre. Mais cela aurait supposé l'étiquetage manuel préalable d'un corpus volumineux. Par ailleurs, nous n'aurions pu catégoriser les pages citées mais non présentes dans le corpus. Appliquer cette approche dans notre corpus aurait été très couteux.

Comme ce problème est directement lié à la délimitation du corpus, il peut être abordé dans le cadre du développement des outils d'extraction sur lesquels le DLWeb travaille actuellement.

Nous avons choisi d'utiliser une méthode plus rapide : le filtrage par mots-clefs sur les Urls. Un ensemble de mots-clefs (guerre, 1418, 1914 etc.) a été défini en étudiant la liste des Url de départ. Les pages qui ne contiennent pas au moins un de ces mots-clefs dans leur url ont été éliminées. C'est une méthode archaïque dont le résultat est très dépendant du choix des mots-clefs : on obtient un corpus sans bruit, voire même trop silencieux. Pour la collecte d'août 2014, on a environ 240 millions de pages au départ et après le filtrage, il ne reste que 22 millions de pages. 90% du corpus a été considéré comme « hors thème » ce qui parait vraiment excessif. Donnons un simple exemple. Une grande partie des sites de notre corpus concernent des régiments d'infanterie, avec des adresses URL du type : *19emeri.canalblog.com*. Le terme *ri* apparait en général dans l'URL mais cette chaine de caractère de deux lettres ne peut être retenue comme élément de filtrage tant elle est pauvre. Ainsi avec notre filtrage, tous ces sites de régiments ont été supprimés, puisqu'il n'y a aucun autre élément qui signale la thématique.

Corpus

On peut choisir de rester dans l'univers de la collecte (BCWeb) ou inclure des liens extérieurs à la collecte. Si on reste dans l'univers de collecte, on garde seulement les nœuds qui correspondent aux hôtes/Url de départs de BCWeb, sinon on garde tous les nœuds. L'intérêt de sortir de l'univers de la collecte est de pouvoir identifier des sites liés à la grande guerre qui auraient échappé à la vigilance des experts.

En utilisant ces trois méthodes - l'agrégation, le filtrage et le choix du corpus-, on peut générer des graphes plus lisibles en améliorant la qualité de l'information et en diminuant sa quantité.

Notons pour le moment que nous avons exclu d'utiliser une stratégie d'échantillonnage des pages. Nous y reviendrons.

4.3.3 Comparaison des stratégies

Dans cette section, on va comparer les effets combinés des trois stratégies présentées dans la section précédente afin de pouvoir trouver les combinaisons les plus utiles.

A notre connaissance, il n'existe pas de mesures permettant de comparer les différentes stratégies : en général, on ne se base que sur une évaluation qualitative.

Le Tableau 3 montre pour chaque combinaison de stratégies le nombre des nœuds et de liens pour la collecte Août 2014. Cette collecte a été choisie pour comparer les stratégies parce qu'elle est de taille moyenne et contient les urls de départ diverses.

Agrégation	**Filtrage**	**Corpus**	**# nœuds**	**# liens**
Hôte	Filtré	BCWeb	456	3356
Hôte	Filtré	Tous	15148	27968
Hôte	Pas Filtré	BCWeb	483	6603
Hôte	Pas Filtré	Tous	252207	521414
Url de départ	Filtré	BCWeb	419	2274
Url de départ	Filtré	Tous	15310	28910
Url de départ	Pas Filtré	BCWeb	462	3469
Url de départ	Pas Filtré	Tous	252399	525460

Tableau 3. Combinaison des stratégies

Hôte vs Url de départ

Pour comparer les méthodes d'agrégations « hôte » versus « url de départ », on a décidé de rester dans le corpus BcWeb. On a généré deux graphes l'un en utilisant la liste des Urls de départ et l'autre avec les hôtes. Pour pouvoir visualiser les différences entre les graphes, on a filtré les nœuds par rapport à leur degré (les nœuds dont le degré est inférieur à 80 ont été supprimés). Le filtrage par degré n'a pas le même impact sur les deux graphes. L'agrégation par hôte a un effet agrégateur plus fort et donc la distribution des degrés est différente (Hôte avec max degré 387 vs. Url de départ avec max degré 307).

Sur la Figure 10 qui correspond à l'agrégation par hôte, on note la place centrale occupée par des nœuds comme youtube.com, facebook.com qui peuvent être considérés comme des pôles principaux. Prenons l'exemple de facebook.com pour illustrer notre comparaison. Dans les Url de départ (Bcweb), il y a deux Urls choisies par les experts :

1. https://www.facebook.com/pages/Bicyclette-Pliante-G%C3%A9rard/254241421345914
2. https://www.facebook.com/pages/Centenaire-de-la-Premi%C3%A8re-guerre-mondiale-en-Haute-Sa%C3%B4ne/400310596705578

Ces deux nœuds sont présents dans le graphe « Url de départ », mais comme ils ont un degré faible, ils ne sont pas visibles dans la Figure 11.

En revanche, dans la Figure 10, ils sont agrégés dans le nœud facebook.com avec tous les liens qui ont facebook.com comme hôte. Voici quelques exemples d'Url qui se trouvent dans le nœud facebook.com :

http://www.facebook.com/pages/Minist%C3%A8re-de-la-d%C3%A9fense/11091233228
http://www.facebook.com/sharer.php?u=http://centenaire.org/fr/agenda?nid=2015

Aujourd'hui, de très nombreuses pages contiennent des liens sortants vers les réseaux sociaux (dans notre exemple Facebook) pour favoriser le partage des articles. Cela augmente le degré du nœud Facebook. Quand on regarde en détail les URL vers Facebook, on se rend compte qu'une très faible partie des liens pointe vers les urls de départ de Facebook consignées dans BCweb. Seul le site centenaire.org pointe vers une page facebook choisie dans BCWeb.

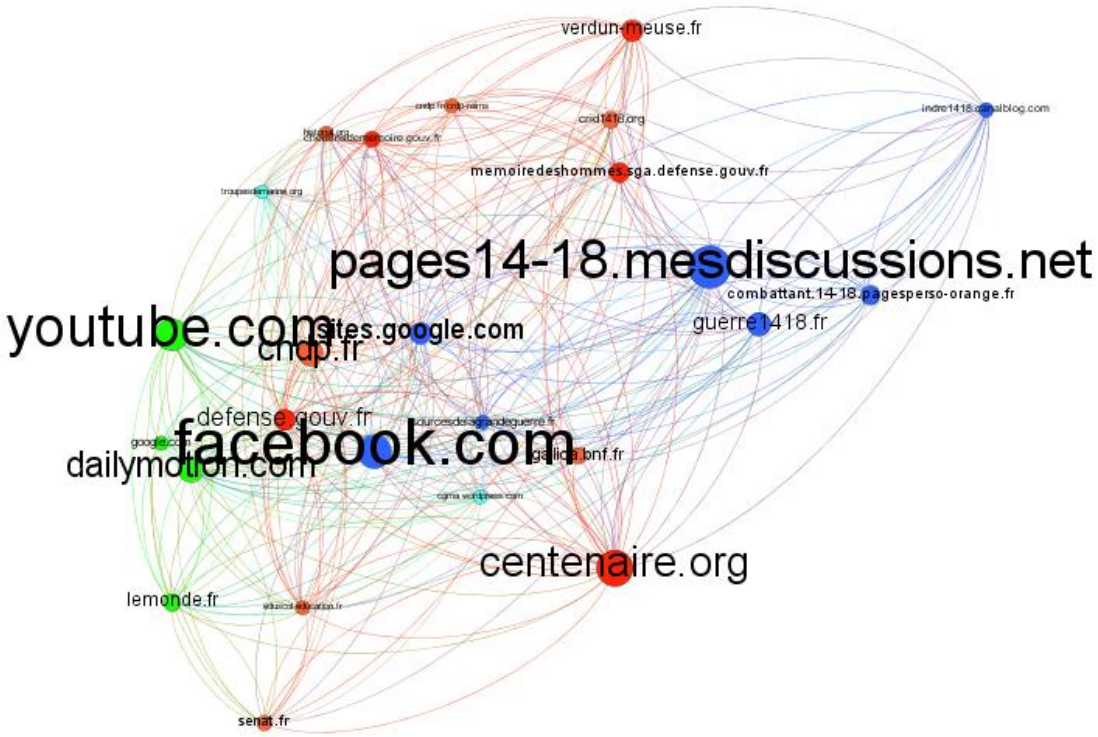


Figure 10. Les nœuds agrégés par Hôte

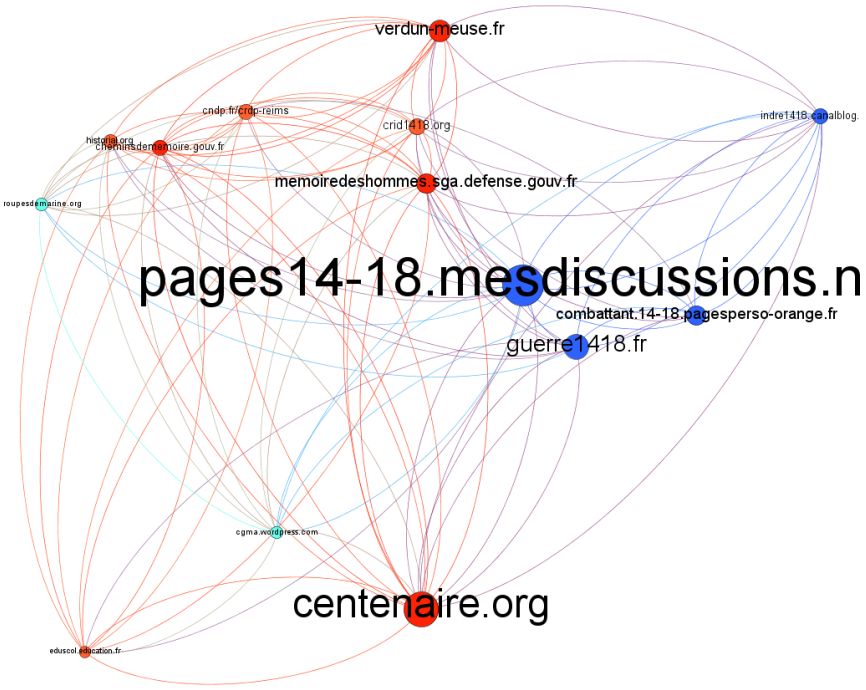


Figure 11. Les nœuds agrégés par Url de départ

Filtré vs non filtré

On compare l'effet du filtrage par mots-clefs (sur les URL) sur le graphe agrégé par hôte. L'utilisation de filtrage augmente la lisibilité du graphe et garantit qu'on élimine toutes les pages qui ne concernent pas la guerre, mais malheureusement aussi beaucoup de pages qui concernent la guerre. La Figure 12 représente un exemple de graphe sans filtrage, très difficile à ouvrir et manipuler avec Gephi. Après avoir filtré le graphe, on obtient la Figure 13, un peu plus lisible et interprétable.

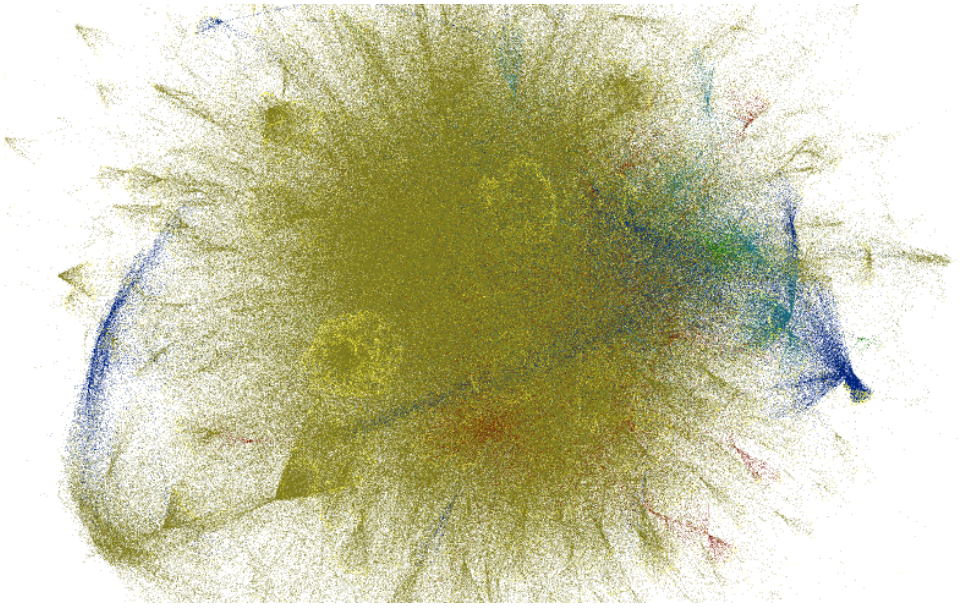


Figure 12. Graphe sans le filtrage

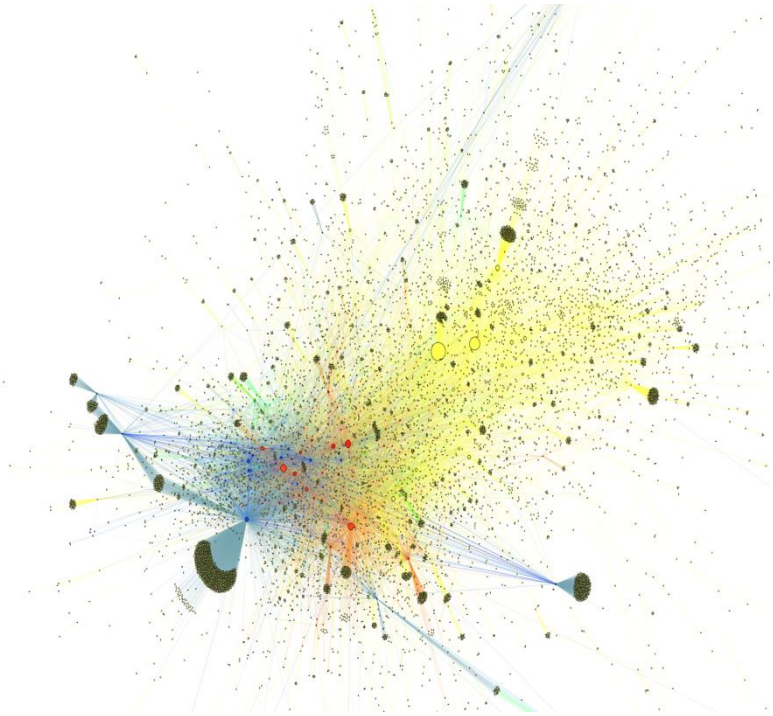


Figure 13. Graphe après filtrage par mots-clefs

Prenons l'exemple de la page : http://www.lemonde.fr/centenaire-14-18/8.html qui pointe vers différentes Urls:

1. http://television.telerama.fr/tele/chaine-tv/france-3,80.php
2. http://www.lemonde.fr/centenaire-14-18-livres/
3. http://www.lemonde.fr/idees/article/2013/10/08/l-europe-ce-n-est-pas-la-paix-c-est-la-consequence-de-la-paix_3492063_3232.html

En utilisant le filtrage, on élimine les Url 1 et 3 qui sont hors sujet et on ne conserve que l'url 2 qui contient le terme « centenaire » et « 14-18 ».

Environ 90% des Urls sont supprimés après le filtrage. Comme nous l'avons vu, cette méthode est très dépendante du choix des mots-clefs : un choix trop sélectif de mots-clefs peut conduire à l'élimination de pages qui sont pourtant pertinentes.

Bcweb vs Tous

Dans cette section, on compare les graphes générés en restant dans le corpus (Bcweb) ou en incluant tous les nœuds cités (Tous). Ce dernier graphe, très difficile à manipuler, correspond à celui de la Figure 12. La Figure 14 représente le graphe en restant dans l'univers du corpus de départ, qui est beaucoup plus lisible.

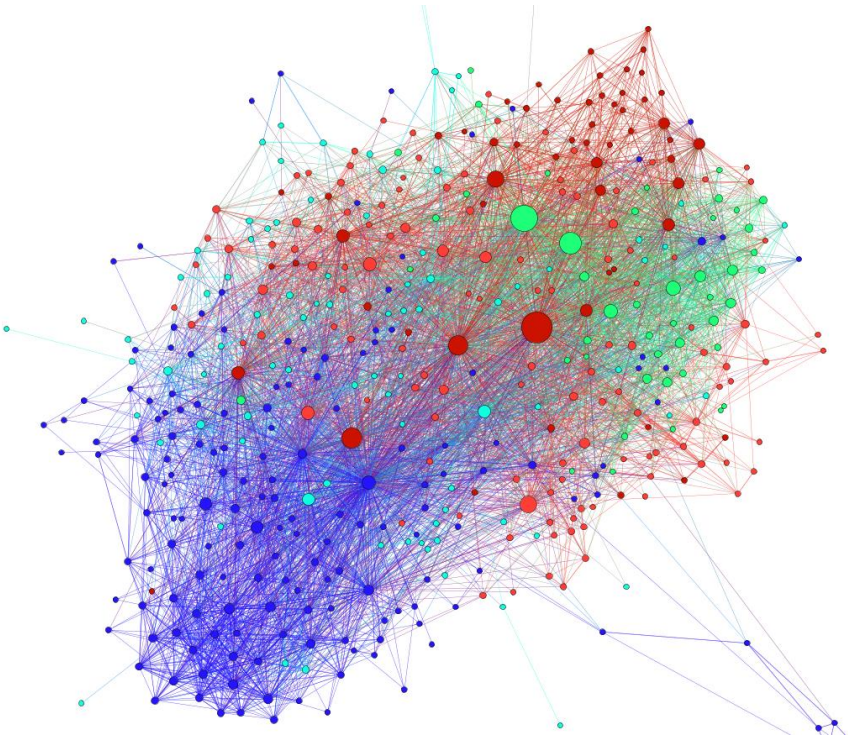


Figure 14. Graphe en restant dans BcWeb (agrégation par hôte)

4.3.4 Décision

Quand on travaille dans le corpus de départ (BcWeb), peu importe que le filtre soit ou non appliqué, le fait de rester dans l'univers de la collecte fonctionne comme un filtre. Si on ne se restreint pas à l'univers de BcWeb, il est important d'utiliser le filtrage

pour avoir un graphe lisible et interprétable, en éliminant le « bruit » qui provient des liens sortants.

Les comparaisons faites sur les modes d'agrégation des nœuds nous conduisent à privilégier l'agrégation sur les Url de départ plutôt que par hôte. En effet, cette dernière méthode accroit artificiellement l'importance de sites partiellement dédiés à la Grande Guerre, comme les sites de réseaux sociaux ou de médias traditionnels.

Les cartes présentées dans les sections suivantes sont restreintes à l'univers du corpus, avec des nœuds agrégés en fonction des URLs de départ et sans aucun filtrage. Ce choix nous parait généralisable à d'autres études portant sur des territoires spécifiques du web. En faisant confiance à ceux qui sont en charge de la collecte, la meilleure représentation sera celle qui se restreint à l'univers de la collecte et qui agrège les documents au niveau de l'URL de départ.

Les différentes explorations que nous avons menées nous ont permis d'évaluer la qualité de la collecte, de vérifier la cohérence des choix et l'absence d'oublis majeurs (un nœud cité par d'autres sites qui n'aurait pas été collecté).

Nous verrons que pour la visualisation des réseaux, il nous faudra aussi introduire un filtre supplémentaire lié au nombre de degrés (nombre de liens entrants et/ou sortants). Il s'agit d'un choix fort, qui favorise les sites les plus interconnectés, ceux qui sont déjà les plus visibles, au détriment des sites isolés. C'est une alternative au modèle de l'échantillonnage aléatoire qui permettrait d'avoir une représentativité de tous les sites quelle que soit leur visibilité. Ce choix s'inscrit dans une évolution étudiée par Kotras dans le monde des sociétés d'études marquée par le déclin du modèle de la représentativité au bénéfice d'un modèle de la visibilité et de l'influence (Kotras, 2013) : certaines opinions comptent plus que d'autres car elles ont une plus grande visibilité et ont donc un potentiel d'influence plus grand. C'est une manière d'intégrer la réception et la lecture dans le processus de cartographie, le réseau des citations étant un reflet de l'état des lectures.

4.4 Évolution de la cartographie de la Grande Guerre

Un des intérêts du travail sur les archives du web est de pouvoir retracer l'évolution temporelle des documents et de leurs relations. Dans notre cas, les différentes collectes devraient permettre d'évaluer comment s'est transformée au cours du temps la structure du réseau de sites consacrés à la Grande Guerre.

> « La plupart des méthodes actuelles de fouille de graphes dynamiques (i.e. qui évoluent dans le temps) reposent sur des séries de graphes statiques représentant soit l'état du graphe à l'instant de chaque capture, soit l'agrégation des nœuds et liens apparus entre cet instant et l'instant de la précédente capture (par exemple le graphe des interactions du mois d'avril dans un réseau social) .»[14]

[14] Investigation visuelle d'événements dans un grand flot de liens (Heyman et Le Grand, 2014).

Avec Gephi, il existe une procédure qui permet, en conservant la spatialisation des nœuds, de voir comment les nœuds et liens apparaissent et disparaissent à chaque étape. Un fichier Gexf est généré pour chaque collecte. Chaque fichier est ouvert dans gephi en indiquant à quelle tranche temporelle il correspond. Ensuite, une fenêtre glissante permet de visualiser l'état du graphe à chaque période. Deux options différentes ont été testées.

4.4.1 Méthode : deux approches

Comme nous l'avons vu quatre collectes différentes ont été faites entre novembre 2013 et novembre 2014. À chaque collecte, le fichier BCWeb s'est enrichi de nouveaux sites) et a été mis à jour (changement d'adresses ou de structures de sites. L'introduction d'un site dans BCWeb ne signifie pas forcément qu'il s'agit d'un nouveau site sur le web : le site pouvait déjà exister mais ne pas avoir été repéré ou avoir été considéré comme non pertinent.

La première option consiste à construire chacun des graphes de la Grande Guerre en ayant comme cadre de référence l'état du fichier BCWeb au moment où la collecte a été faite. Chaque graphe reste alors exclusivement dans le périmètre de la collecte faite.

La seconde option consiste à prendre comme cadre de référence pour les quatre graphes, l'état final du fichier BCWeb. Ceci permet de visualiser les nœuds qui n'avaient pas été retenus mais qui étaient déjà cités lors de la première collecte. Cette seconde option permet d'atténuer les effets liés à la sélection.

On peut également avoir une approche très locale qui consiste à suivre l'évolution d'un sous-graphe lié à des nœuds spécifiques.

4.4.2 Résultats

La Figure 15 illustre la première démarche pour les quatre collectes. On voit que les cartes sont très liées au périmètre de chacune des collectes. Par exemple, dans la première collecte, presque aucun site personnel ou associatif n'était présent dans la liste des Url de départ. Le monde des sites amateur n'avait tout simplement pas encore été repéré par les équipes en charge de la collecte. La carte, à l'image de la collecte, était donc très réduite. Ce n'est qu'à partir de troisième collecte que l'on a une carte dense et presque complète. Cette manière de faire est très dépendante du travail d'identification des sites. Notre but étant d'étudier l'évolution des sites dédiés à la Grande Guerre, il faut pouvoir s'affranchir des limites de la collecte autant que possible.

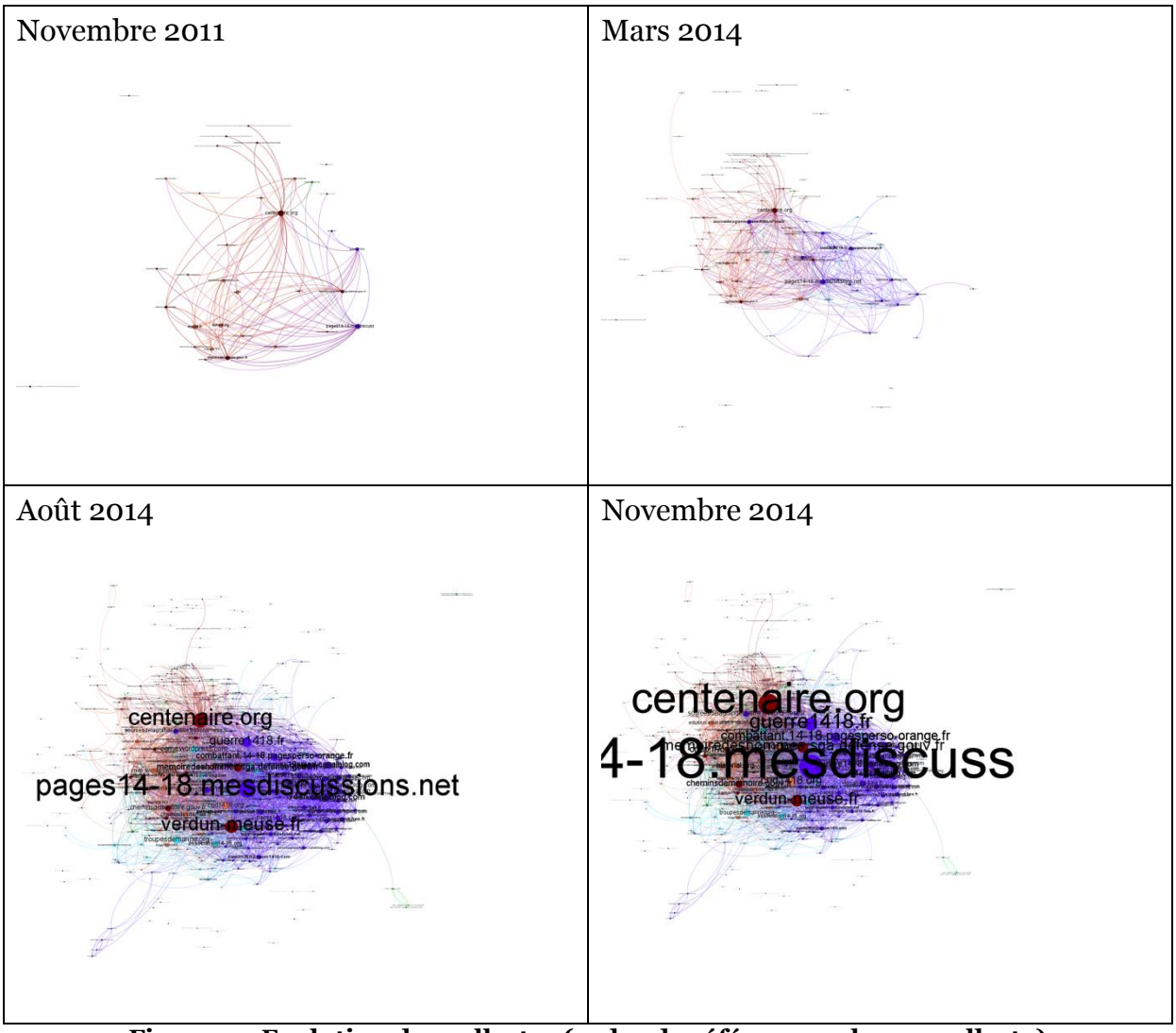


Figure 15. Evolution des collectes (cadre de référence : chaque collecte)

Pour être plus autonome par rapport à la construction de la collecte, on utilise la dernière liste d'Url de départ qui est beaucoup plus complète pour générer toutes les cartes (Figure 16). Même si certains nœuds n'ont pas été collectés à une certaine date, s'ils sont cités par des sites collectés, ils vont pouvoir apparaitre sur la carte. En revanche, on ne pourra pas reconstituer les relations entre les sites non archivés. Cela introduit un autre type de biais. Loin d'être parfaite, cette méthode améliore cependant la représentation temporelle.

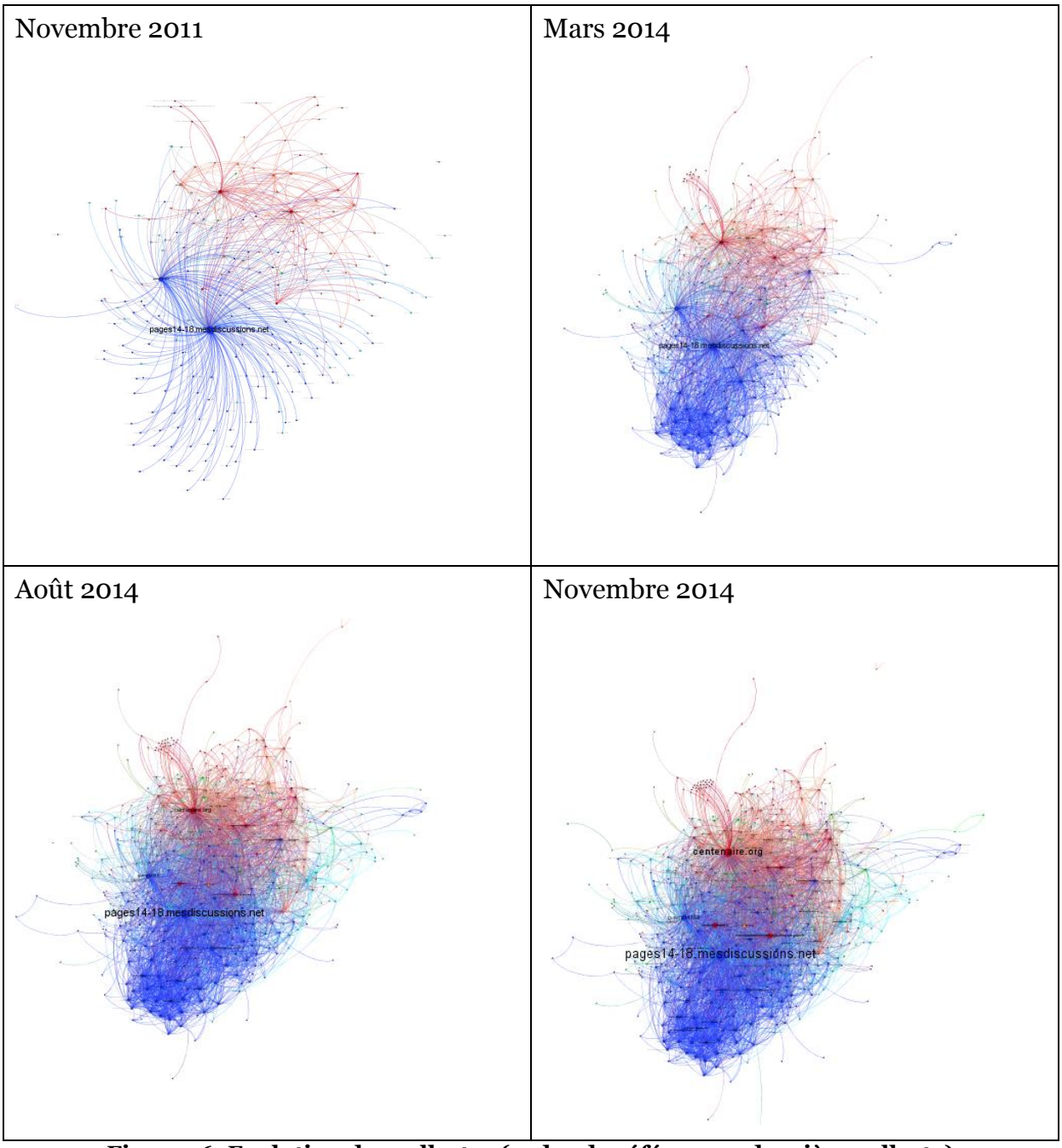


Figure 16. Evolution des collectes (cadre de référence : dernière collecte)

On peut aussi explorer les graphes égocentrés de certains sites pour faire des études approfondies sur l'évolution de la structure des liens entrants et sortants de ces sites.

Représenter la dynamique des sites et des relations entre les sites est un enjeu de recherche important pour appréhender les transformations de l'espace numérique. Mais nous sommes confrontés à une grande incertitude : la densification des relations d'une collecte à l'autre est-elle liée à une meilleure identification des sites pour l'archivage ou à une croissance de l'activité en termes de nombre de sites et de liens entre les sites ? Les graphes sont-ils l'index du territoire ou de l'exercice cartographique ? Dans le cadre de ce projet, nous ne sommes pas en mesure de trancher, même si le contexte du Centenaire nous incite à penser que la densification

du réseau est bien un effet de l'accroissement de l'activité en ligne[15]. Face à l'indécidabilité, nous avons pris comme option de travailler sur une collecte qui rend compte de l'état du web consacré à la Première Guerre mondiale en novembre 2014 tel qu'il a pu être appréhendé par l'équipe en charge de la collecte. Il s'agissait de la quatrième collecte, on peut donc faire l'hypothèse que le territoire était bien exploré.

[15] Si des métadonnées indiquaient la date de publication des documents archivés, il serait possible d'évaluer ce qui relève d'un phénomène ou de l'autre, mais ce n'est pas le cas.

5 CARTOGRAPHIE DU WEB CONSACRE A LA GRANDE GUERRE EN NOVEMBRE 2014

Cette section propose un parcours à travers la carte du web dédié à la Première Guerre mondiale en 2014. Nous en décrirons les reliefs et les curiosités avant de laisser la parole à ceux qui participent à faire vivre la mémoire de la guerre sur le web. Ces acteurs, équivalents d'informateurs locaux, seront à même de commenter l'organisation de la carte et d'en déceler aussi les limites et les manques. Si, pour simplifier la description, nous employons une métaphore spatiale pour décrire cette carte, nous avons conscience qu'il ne s'agit pas d'un espace topographique (Severo et Venturini, 2016).

Nous nous appuyons dans cette section sur le graphe qui correspond à la collecte de novembre 2014, la dernière que nous avons pu exploiter dans le cadre du projet et la plus complète. Les nœuds ont été agrégés à l'URL de départ et le graphe reste dans le périmètre de la collecte, conformément aux choix qui ont été faits suite à l'analyse comparative des graphes (cf. section précédente). Un nœud correspond à un point d'entrée de BCWeb. Il y a un lien du site A vers le site B, si au moins une page du site A contient un lien hypertexte vers le site B.

La couleur du nœud correspond à la nature de l'émetteur (rouge pour les sites officiels, orange pour les sites publics, bleu marine pour les sites personnels, bleu clair pour les sites d'associations et vert pour les sites de média)[16].

La taille du nœud et du label dépend du nombre de liens (ensemble des liens, liens entrants ou sortants).

La couleur du lien dépend de la catégorie de l'émetteur. Un lien est coloré en bleu marine s'il provient d'un site personnel.

Pour la spatialisation du graphe, les algorithmes Force Atlas (1 et 2) ont été utilisés avec les paramètres présentés dans la section consacrée à la méthode.

En amont, comme nous l'avons vu dans la section précédente, il y a autant de graphes que de choix faits pour la sélection et l'agrégation des nœuds. Une fois le graphe délimité, sa représentation varie selon les modes de visualisation des nœuds et selon les algorithmes de spatialisation choisis. Un graphe peut avoir une infinité de représentations. Nous avons fait des centaines de cartes (cf. partie précédente) en faisant varier bien des paramètres. Nous avons finalement opté pour une

[16] Comme nous l'a indiqué une des personnes rencontrées, le choix des couleurs n'est pas ici très heureux, puisque le rouge était employé pour marquer les fronts français, le bleu pour les lignes allemandes sur les différentes cartes géographiques.

représentation qui permet de visualiser au mieux les éléments les plus stables que nous observions dans toutes les cartes. Les points sur lesquels nous insistons dans l'analyse sont ceux qui sont stables quelle que soit la représentation et qui sont confirmés par des calculs supplémentaires et par une observation manuelle du contenu des sites et de leurs réseaux de liens. Le travail d'interprétation ne peut être délégué à une machine. Seul le regard du chercheur peut donner du sens à ce que produit la machine.

5.1 Les acteurs presents en ligne s'organisent selon deux poles

La répartition des sites selon le type d'émetteur, tel que catégorisé par les responsables de la collecte, montre que plus de la moitié des sites proviennent d'individus ou de collectifs d'individus : 37% sont des sites personnels et 17% des sites d'associations. Les sites officiels et publics, que nous avons regroupé sous le label sites institutionnels, sont moins nombreux (35%), mais en général, ils sont cependant de bien plus grande taille. Enfin, 11% des sites ou sections de site proviennent de médias qui ont proposé des dossiers spéciaux sur la première guerre mondiale.

Editeur	Nombre de sites	%
Site ou blog personnel	187	37%
Site d'associations	90	17%
Sites institutionnels (Public ou Officiel)	179	35%
Sites de media	56	11%
Total	512	

54%

Figure 17. Répartition des sites selon émetteurs

La Figure 18 propose une représentation globale du graphe. On y distingue au nord-ouest une zone où dominent les sites publics (nœuds orange) et les sites officiels (nœuds rouge) avec un nœud qui se distingue par sa taille, le site Centenaire.org. Les sites officiels et publics se distinguent peu les uns des autres en occupant le même espace de la carte. Ceci nous conduit à les considérer comme des sites *institutionnels*

sans conserver la distinction entre site public et officiel[17]. Au sud, une zone dense de sites très interconnectés, dominée par des sites personnels, s'organise autour de son plus gros nœud, le forum Pages14-18.mesdiscussions.net. Au nord-est, à la frontière du graphe apparaissent des sites associatifs, plus faiblement connectés aux autres. Les sites des médias, qui représentent pourtant 11% des sites, sont peu visibles sur la carte, ce qui témoigne d'une très faible connexion aux autres sites. Une représentation des sites en fonction de leur audience aurait sans doute donné aux sites de médias une bien plus grande visibilité, mais ici seule est représentée la structure des liens.

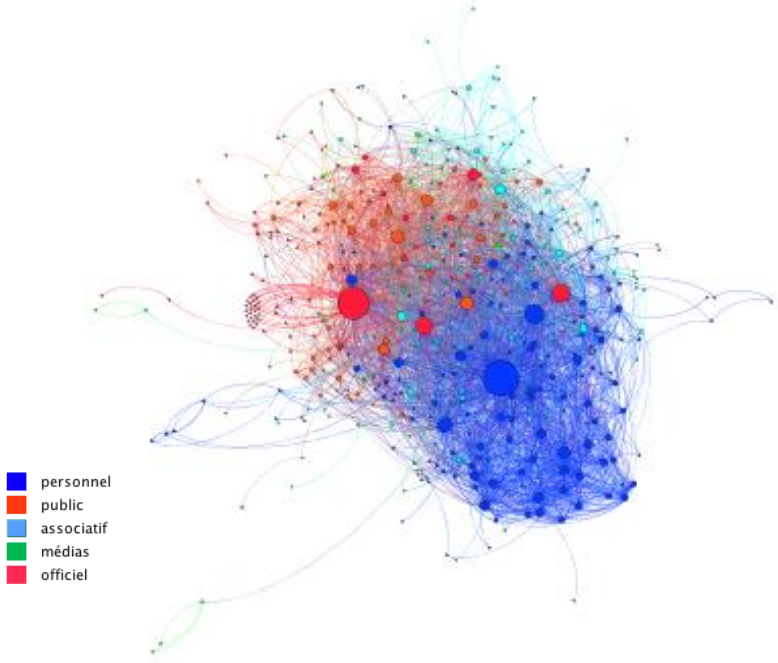


Figure 18. Réseau avec l'ensemble des liens entrants et sortants

Notons que le graphe ne produit pas une dichotomie parfaite entre les sites institutionnels et les sites personnels et associatifs. En effet, quelques sites personnels ou associatifs se trouvent dans la zone institutionnelle. Inversement, quelques sites institutionnels se situent dans la zone des sites personnels. Nous reviendrons sur ces points.

Les sites ont été catégorisés par les experts en charge de la collecte en fonction de la qualité de l'auteur ou du responsable éditorial du site. Un site est personnel s'il est mené en nom propre par une personne physique, il est associatif, s'il représente le site d'une association et il est institutionnel s'il représente une entité publique, nationale, régionale ou une organisation avec financement public.

[17] Le fait que les sites officiels et publics partagent le même territoire sur la carte témoigne de la difficulté à les différencier.

La cartographie qui s'appuie sur la structure des liens entre les sites montre globalement la cohérence de ces catégories *a priori*[18] puisque les sites personnels sont dans leur majorité rassemblés au sud de la carte, en raison de leurs interconnexions, tandis que les sites institutionnels se situent au nord. Nous désignons la zone sud de la carte comme étant la zone des *amateurs*, car l'examen des sites qui la constitue montre que ce sont principalement des chercheurs non professionnels, passionnés de la guerre, qui en sont les auteurs. Les sites personnels qui en raison de leur structure relationnelle se situent en dehors de cette zone nous semblent avoir un statut différent et ne seront pas considérés comme relevant de cette sphère amateur.

Dorénavant, pour faciliter la lecture, ne seront conservés que les nœuds de degré 30 ou plus (avec au moins 30 liens entrants et sortants) comme dans la Figure 19. Comme nous l'avons vu précédemment, ce filtrage favorise les sites les plus en vue et surtout les sites insérés dans des réseaux de collaboration et d'échanges, dont les liens sont la trace.

[18] Quelques modifications ont été faites dans la catégorisation des sites, souvent grâce aux entretiens : ainsi les sites Crid 14-18 et En evor ont été reclassés comme des sites associatifs.

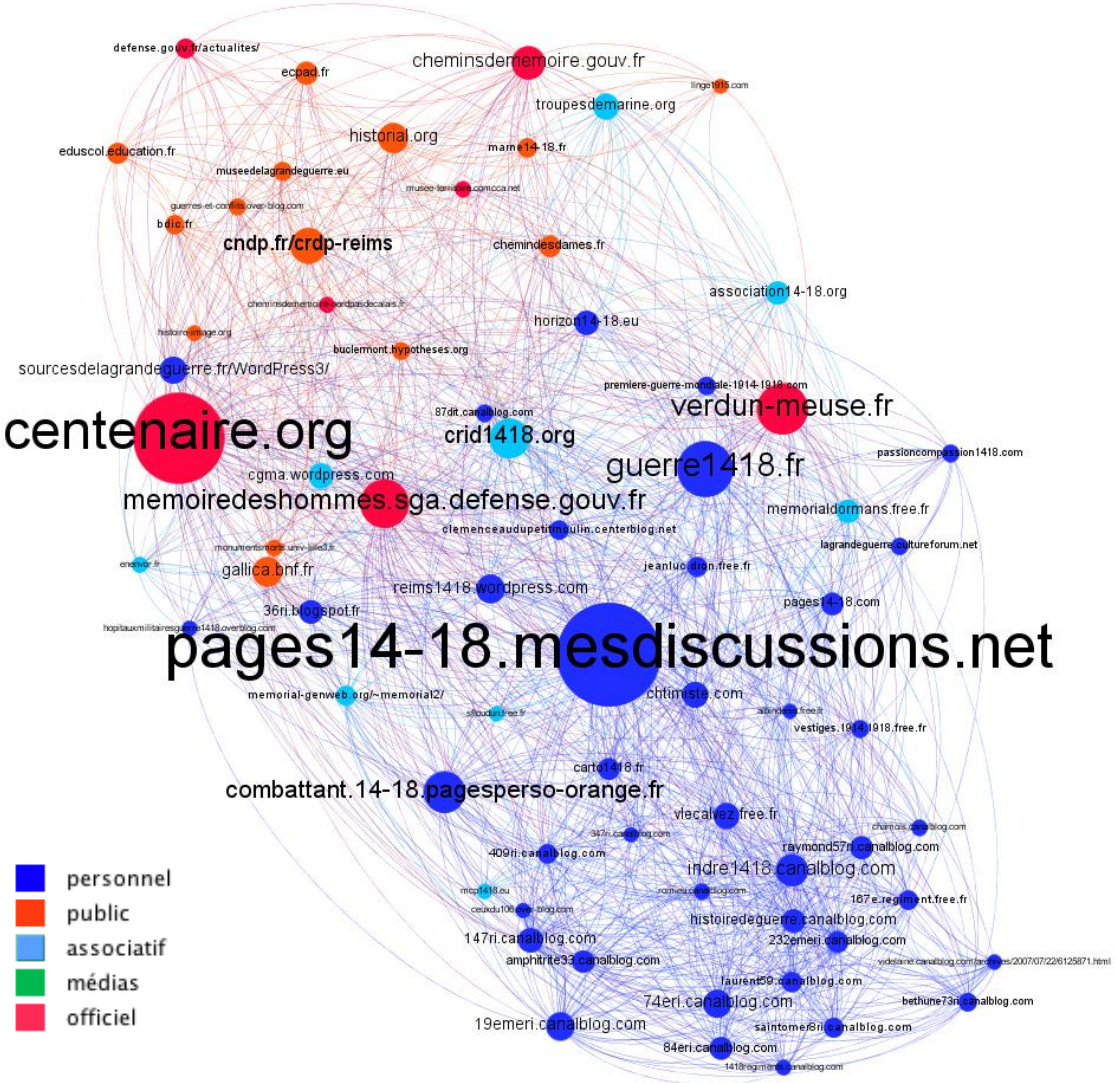


Figure 19. Carte du Web Grande Guerre (collecte de novembre 2014, sites avec degrés >30)

La carte montre l'existence de deux zones distinctes chacune dominée par un des nœuds principaux du graphe : le site du Centenaire d'un côté et celui du forum Pages 14-18 de l'autre.

5.2 Le pole institutionnel

La Mission du centenaire de la Première Guerre mondiale est un groupement d'intérêt public créé en 2012, associant de nombreux ministères et institutions

patrimoniales. Elle est chargée de soutenir et de valoriser les actions de commémoration sur la période 2014-2018. Ses objectifs sont :

- « Organiser, de 2014 à 2018, les temps forts du programme commémoratif du centenaire de la Première Guerre mondiale décidés par le Gouvernement ;
- Coordonner et accompagner l'ensemble des initiatives publiques et privées mises en œuvre en France ou par la France à l'étranger dans le cadre du Centenaire, en proposant notamment un « label Centenaire » et un programme officiel des principales manifestations organisées autour du Centenaire. La Mission du Centenaire s'appuie, dans cette perspective, sur les Comités départementaux du Centenaire (CDC) mis en place dans chaque département afin de coordonner, sous l'autorité des préfets, l'action des services déconcentrés de l'Etat, des collectivités territoriales et des associations, sur les référents « mémoire et citoyenneté » qui président les Comités académiques du Centenaire (CAC) mis en place par l'Education nationale dans chaque académie, sur le réseau des acteurs du tourisme (GIE Atout France, comités régionaux du tourisme, comités départementaux du tourisme et offices de tourisme), ainsi que sur le réseau des ambassades et des Instituts français à l'étranger ;
- Informer le grand public sur les préparatifs du Centenaire et mettre en œuvre une politique de communication autour des principales manifestations organisées dans le cadre du Centenaire et assurer la diffusion des connaissances sur la Grande Guerre, notamment grâce à un portail de ressources numériques de référence. »

Source : Site de la Mission du Centenaire, consulté en décembre 2016.

Le site de la Mission (www.centenaire.org) a été créé en 2013 et constitue la vitrine principale des activités. En à peine deux ans, il est devenu un point nodal du web de la Première Guerre mondiale. Ce site officiel, de par sa mission même, valorise ses partenaires et toutes les initiatives qu'il soutient via un signalement systématique par des billets et des liens numériques.

Comme sa mission est de « coordonner et accompagner toutes les initiatives *privées et publiques* », le site indexe également les sites personnels ou associatifs comme Pages 14-18 et les sites liés. Son ouverture aux initiatives privées, explique sa position proche de la zone « amateur ». Le site joue un rôle de portail vers toutes les initiatives tout en étant un site de référence mentionné dans la grande majorité des sites.

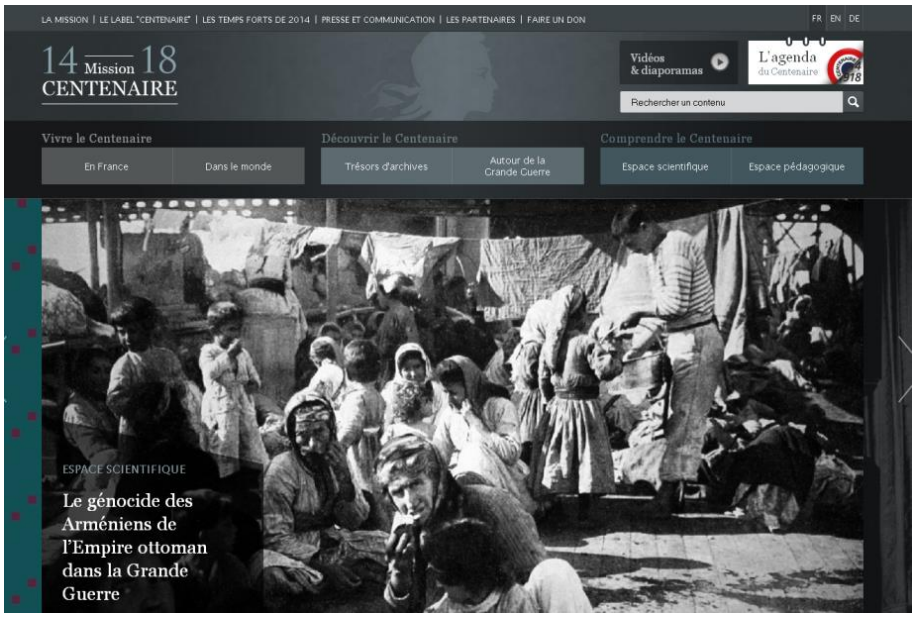


Figure 20. Site de la mission Centenaire (centenaire.org), mars 2015

Autour du site du Centenaire, dominent cependant les sites institutionnels liés à la guerre, qu'ils soient liés au tourisme, à l'éducation ou à la recherche.

5.2.1 Entre tourisme et éducation

Le discours des sites institutionnels sur la première guerre mondiale porte principalement sur la mémoire : le terme *mémoire* est beaucoup plus fréquemment utilisé que le terme *histoire* ne serait-ce que dans les noms des sites, témoignant de la montée en puissance de l'approche mémorielle dans l'espace public. Les sites *Mémoire des hommes* et *Chemins de mémoire* du ministère de la Défense, tous deux très visibles sur la carte, en sont une bonne illustration : le premier recense les fiches de tous les morts pour la France ainsi que tous les journaux des marches et des opérations ; le second recense tous les lieux de commémoration sur le territoire français.

La dimension géographique est centrale dans l'activité de mémoire autour de la Première Guerre mondiale. Une grande partie des sites institutionnels a un ancrage territorial clairement affiché. Sur toute la ligne de front ont émergé dès la fin de la guerre des lieux de mémoire qui entretiennent des liens étroits avec l'activité touristique : musées, mémoriaux, champs de bataille, forts, cimetières… Les sites web, faits à l'initiative de ces lieux de mémoire, des offices de tourisme ou des conseils régionaux, en sont la vitrine. Souvent connectés entre eux, ils reconstruisent une ligne du front virtuelle. Le site régional le plus saillant dans notre cartographie est Verdun-Meuse, le site officiel de la Mission Histoire du Conseil Général de la Meuse. Mais nous trouvons aussi les sites de mémoriaux comme celui du Chemin des dames, de la Caverne du dragon, du fort de Douaumont, ou celui du Linge pour n'en citer que quelques-uns. La plupart de ces sites existaient avant le Centenaire, mais

pour certains le centenaire a été l'occasion d'un renouvellement du lieu et du site lui-même comme pour le mémorial de Verdun dont le site a été entièrement refondu en même temps que le bâtiment rénové et agrandi.

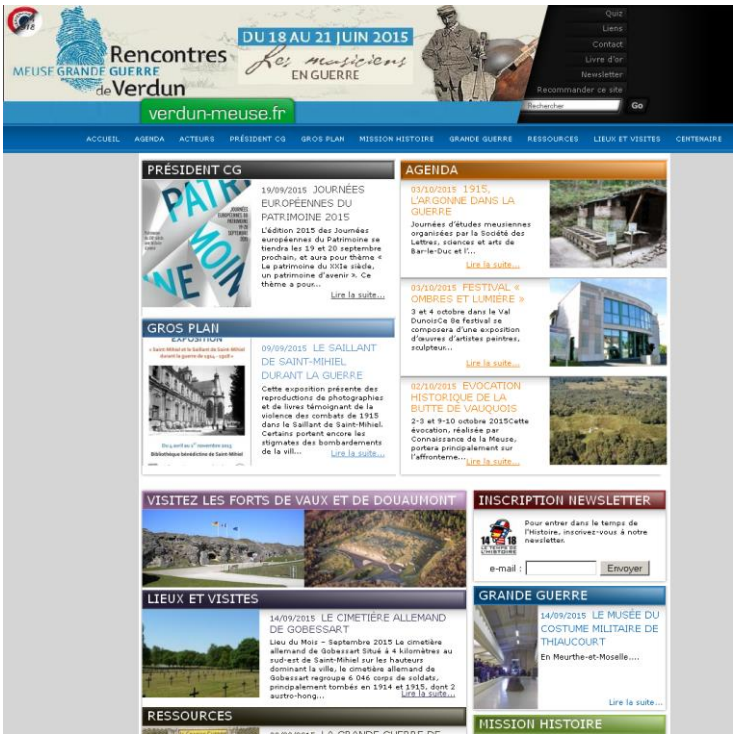


Figure 21. http://www.verdun-meuse.fr/ (septembre 2015)

Dans cette zone institutionnelle, on notera enfin l'importance des sites pédagogiques qui proposent des ressources éducatives autour de la Grande Guerre, en particulier à travers le site Eduscol et le réseau Canopé (éditeur de ressources pédagogiques public, dépendant du ministère de l'Éducation nationale, anciennement CNDP). De nombreuses ressources sont assemblées et proposées autour de la Guerre, en particulier dans les académies qui sont situées sur la ligne de front (Amiens, ville de Reims, etc.).

Les logiques touristiques et éducatives ne sont pas reliées (ce ne sont pas les mêms acteurs, les mêmes types de contenus) mais elles partagent le fait d'être ancrées dans un territoire géographique profondément marqué par la guerre.

Figure 22. Picardie 14-18, CANOPÉ de l'académie d'Amiens (crdp.ac-amiens.fr/picardie1418, mars 2015)

5.2.2 Quelle place pour la recherche académique en ligne ?

La Première Guerre mondiale s'est constituée en objet de recherche universitaire dès la fin du conflit et l'intérêt académique pour ce conflit ne s'est pas démenti, voire s'est renforcé avec le Centenaire, comme en témoigne le nombre de publications de livres, de numéros de revues, de conférences organisées…

Pourtant la recherche académique sur la Grande Guerre est peu visible sur le Web sans doute parce qu'elle y est très dispersée : un blog de chercheur, une page pour un livre chez un éditeur, une page pour un numéro spécial de revue ou pour une conférence, une page pour un projet de recherche… chacune de ces publications citant peu les autres et étant elle-même peu citée. Cet éclatement des publications explique la faible visibilité en ligne de la recherche.

Deux collectifs de chercheurs de nature différente apparaissent cependant sur notre carte. Le premier, hébergé par l'Historial de la Grande Guerre à Péronne, est un centre de recherche international adossé au musée qui « regroupe d'éminents spécialistes universitaires venus du monde entier, qui ont été dès l'origine, associés à la création du musée » (présentation sur le site) et qui dispose en plus du site d'un carnet de recherche sur une plateforme dédiée à la recherche hypotheses.org.

Le second est le CRID 14-18 (Collectif de Recherche International et de Débat sur la Guerre de 1914-1918), né en 2004. Il s'agit d'une « association de chercheurs qui vise au progrès et à la diffusion des connaissances sur la Première Guerre mondiale ». Comme l'indique sa charte, il « a pour origine la rencontre d'universitaires, d'historiens non professionnels et de membres de sociétés savantes sur le terrain

même des combats de 1914-1918 ». Par sa composition et par sa localisation à Craonne, la ville de la chanson éponyme, chanson antimilitariste associée aux mutineries, dans la région de l'Aisne où les tentatives de résistance à la contrainte de la guerre ont été les plus fortes, le CRID 14-18 met l'accent sur les contraintes exercées par l'Etat sur ses populations.

Ces deux collectifs de chercheurs défendent tous deux une histoire culturelle de la guerre, s'interrogeant principalement sur le destin des hommes au combat. Il fut un temps où leurs lectures opposées du conflit, la thèse du patriotisme et du consentement s'opposant à celle de coercition par l'Etat, devint l'objet d'un conflit intellectuel ayant trouvé un écho dans la presse (Birnbaum, 2006).

Le positionnement des deux sites sur la carte est différent. Tandis que le site de l'Historial se situe au nord de la carte, le site du CRID est très proche de la zone amateur. Cette disposition est liée à la structure des liens de ces deux sites : s'ils sont quasiment autant cités l'un que l'autre par une soixantaine de sites dont l'origine est à peu près similaire (les liens proviennent de sites personnels ou associatifs dans les deux tiers des cas), ils se distinguent par le nombre et la structure des liens sortants. Le site de l'Historial cite 15 sites dont 2 seulement sont des sites personnels ou associatifs, tandis que le Crid 14-18 en cite 42, dont 26 personnels et associatifs, soit près des 2/3. Si l'on reprend la typologie de (Cardon, Fouetillou et Roth, 2014), le Crid 14-18 est *curieux* par son ouverture conjointe au territoire institutionnel et amateur, tandis que le site de l'Historial est plus *introverti*, relié principalement à l'espace institutionnel. Cette observation ne vaut que pour novembre 2014 évidemment.

Sa connexion avec le monde des amateurs explique le positionnement du CRID 14-18 sur la zone frontière. Nous l'avons vu, cette structure de lien est aussi à l'image de la structure de l'association qui a fait le choix de « l'ouverture sur des usages et des publics multiples » et qui a tenté de mettre en place un forum ouvert « qui permet le dialogue entre universitaires et « amateurs » [ce qui] a, dans le champ historique français, peu d'équivalents » (Loez, 2008).

Enfin, le CRID 14-18 est plus souvent cité par les amateurs (dans les entretiens en particulier) car il constitue une ressource pour le travail de recherche « pas mal de références bibliographiques, des critiques de bouquins, des évènements qui étaient aussi annoncés » (CM).

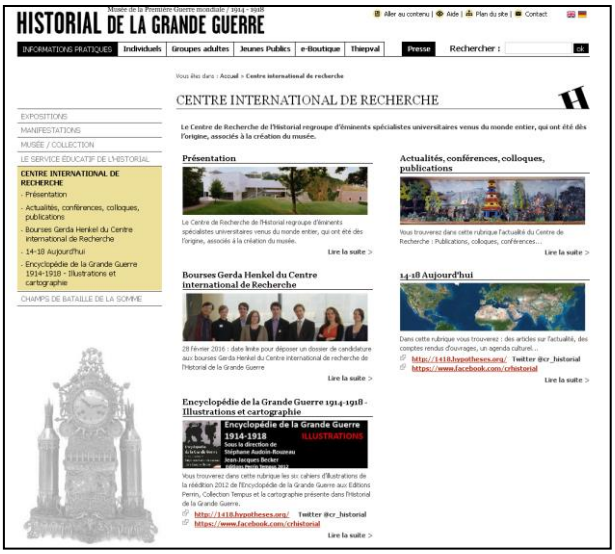


Figure 23. Site du Centre International de recherche de Péronne (décembre 2015)

Figure 24. Site du CRID (Collectif de Recherche International et de Débat sur la Guerre de 1914-1918) (décembre 2015)

En bref, l'activité de recherche académique sur la Grande Guerre, à l'exception du CRID 14-18, n'a pas de visibilité claire sur l'espace du web alors que les projets de recherche et les colloques sont extrêmement nombreux. Cela tient au caractère dispersé des équipes de recherche et publications, structurées par discipline et non par objet de recherche. Finalement, c'est le site du Centenaire, via sa rubrique « espace scientifique », qui fédère l'activité scientifique autour de la Grande Guerre (comptes-rendus d'ouvrages, annonces de séminaires…) et en montre la densité.

La zone institutionnelle visible sur le web, dans laquelle le site du Centenaire joue un rôle pivot, est dominée par des sites liés à des lieux de mémoire (musées, mémoriaux…) en lien avec l'activité touristique -, par des sites éducatifs et par des sites de ressources documentaires (nous y reviendrons dans la section : Les sites patrimoniaux comme entrepôts de documents, p. 69).

5.3 LE POLE « AMATEUR » : LE FORUM ET SON ENVIRONNEMENT

Examinons à présent la zone « bleue » au sud de la carte en commençant par son site principal, le forum Pages 14-18

5.3.1 Le forum Pages 14-18

Pages 14-18 est un forum de discussion créé en 2004 par Joël Huret, un historien amateur, par ailleurs collectionneur et responsable d'un musée. Le fondateur s'est peu à peu entouré de modérateurs pour organiser l'animation et la modération du forum.

Plus de 10 ans après son lancement, en janvier 2015, le forum compte 15 000 inscrits, 400 000 messages publiés. Une centaine de messages sont publiés chaque jour et en termes de visiteurs, plus de 33 000 pages sont vues par jour, dont les trois-quarts par des personnes non inscrites dans le forum.

Il existe bien un autre forum (Histoire et militaria 14-18, http://lagrandeguerre.cultureforum.net/) présent sur la carte (240 000 messages et 2135 inscrits en janvier 2015), mais le forum Pages 14-18 s'est imposé comme un lieu de référence de discussion sur la Grande Guerre en France.

Beaucoup de collectifs en ligne s'appuient sur un forum de discussion pour animer la vie collective du groupe.

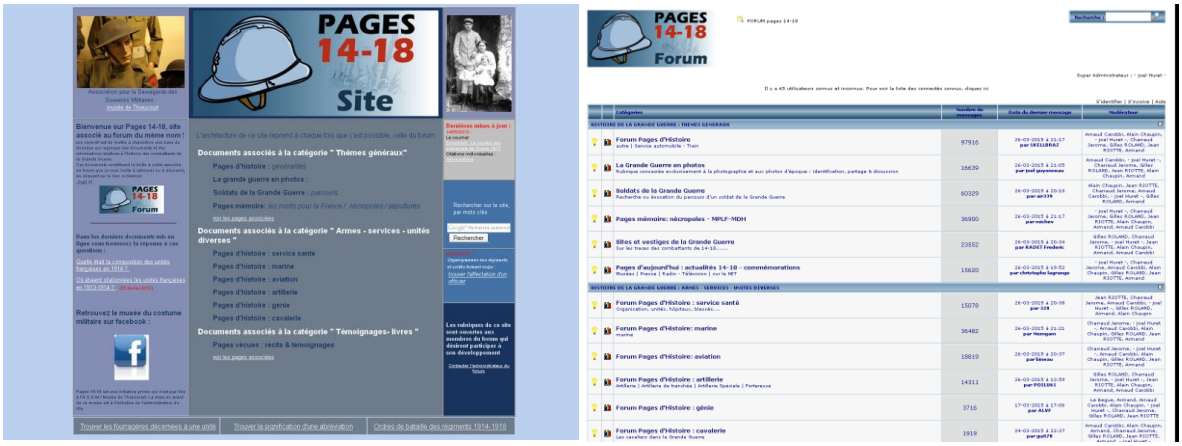


Figure 25. Forum Pages 14-18 (mars 2015)

5.3.2 Le collectif des amateurs

Autour du forum, apparaît une zone très connectée et homogène de sites personnels (zone de la carte où domine la couleur bleu). Au cœur de cette nébuleuse, le forum est à la fois le site le plus référencé et celui qui pointe vers le plus de sites. En effet, dans la page de liens des sites ou blogs personnels, le forum Pages 14-18 figure toujours en

bonne place. Et en sens inverse, au sein du forum, les liens vers les sites voisins sont nombreux. On peut cependant distinguer dans le forum deux usages contrastés du lien hypertexte : la citation et la signature. Les liens dans le cœur du message pointent naturellement vers des ressources proposées par les sites voisins. En clôture de message, les liens ont une autre signification : beaucoup de participants ajoutent une signature standardisée à leur message avec un lien vers leur site personnel, qui joue le rôle de « résidence » numérique. Par ailleurs, les blogs se citent beaucoup entre eux, ce qui explique cette zone dense de sites interconnectés. De fait, les activités dans le forums et les sites alentours sont entrelacées : les auteurs de site sont souvent d'importants contributeurs dans le forum.

Ces sites personnels ont en commun plusieurs éléments. La plupart sont hébergés sur la même plateforme de blog, Canalblog, sans doute par un effet de contagion : l'utilisation du même environnement permet de réduire les coûts d'apprentissage de la publication en ligne. En effet, en étant inséré dans un réseau d'utilisateurs, chacun bénéficie de l'entraide technique, des conseils d'édition et de publication du réseau. Le même phénomène avait été observé pour les écrivains sur web en France qui étaient très nombreux à utiliser la même plateforme de publication Spip (Beaudouin, 2012). De même que les commerces d'un même type sont souvent regroupés dans un même quartier, voire une même rue, on observe une tendance des amateurs à se regrouper sur une même infrastructure numérique.

Ensuite, ils sont nombreux à porter sur un régiment d'infanterie spécifique, ce qu'indique l'adresse du site qui contient souvent le terme RI. Ces sites régimentaires sont des lieux d'archivage de toute la documentation concernant un régiment particulier, ses hommes et ses batailles. On y trouve des documents d'époque provenant tant d'archives publiques (extraits de journaux numérisés, extraits de documents officiels comme les journaux des marches et des opérations, cartes du front…) que privées (photographies, lettres, extraits de carnets personnels). L'objectif du site n'est pas seulement de collectionner toute la documentation mais de la faire revivre par le travail de transcription, commentaire et analyse. Les publications sont organisées par billets, classées de manière thématique. Retracer la vie d'un soldat, le déroulement d'une bataille, analyser une photographie sont des activités communes sur ces sites. L'écriture des billets passe par l'assemblage de sources documentaires hétérogènes, comme des photographies d'époque, des extraits de presse, des cartes géographiques inscrits dans une nouvelle narration portée par l'auteur du site.

Ainsi, sur le site consacré au 74ème RI, dont l'auteur est Stéphan Agosto, on trouve des portraits de combattants, des documents sur différentes périodes de la guerre, une reprise un siècle plus tard du Canard du Boyau, journal de tranchées du 74ème RI et toute une série de liens.

Figure 26. Le 74ème Régiment d'infanterie, site de Stéphan Agosto

Les portraits sont des billets illustrés par une photographie du soldat ou un dessin réalisé par l'auteur du site : chacun tente de reconstituer le parcours d'un combattant à partir des traces assemblées, tout en rendant compte des difficultés rencontrées dans le parcours de recherche : les trous dans les données, les manques, les incertitudes sont commentés et parfois le billet s'achève même sur un appel à contribution à l'attention de descendants ou d'autres amateurs.

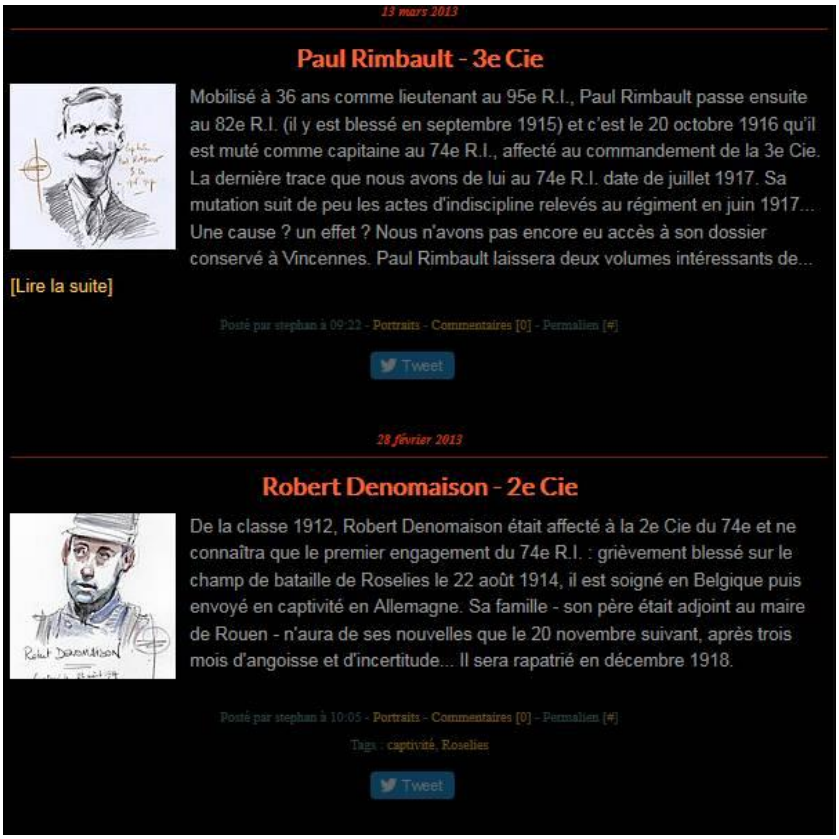


Figure 27. Portraits de soldats sur Le 74ème Régiment d'infanterie

Les liens pointent vers des ressources complémentaires liées au régiment, qu'elles soient publiques, comme l'historique du régiment sur l'Argonnaute de la BDIC ou le Canard du Boyau numérisé dans Gallica, ou privées, comme d'autres blogs reproduisant des carnets de soldats.

Ces sites ont des niveaux d'élaboration variables en fonction de l'ancienneté de l'activité, de la qualité de la collection personnelle du chercheur, de la richesse des fonds accessibles en ligne.

Le modèle du site régimentaire est le plus fréquent parmi les sites visibles sur la carte. En effet, l'échelon du régiment permet d'inscrire l'histoire d'un individu dans celle d'un groupe, dans une histoire sociale et d'intéresser un public plus large (les descendants des soldats du régiment en particulier).

Mais bien d'autres sites, qui n'ont pas atteint ce niveau de visibilité existent. Ils portent principalement sur un ancêtre, dont les documents sont resitués dans le contexte social et géographique. Ainsi abondent des sites constitués autour d'un document d'archive trouvé par un descendant. C'est en particulier le cas d'Aline Duchêne (http://1914-joseph-duchene.eklablog.com/) qui a entrepris de retranscrire tous les carnets de son arrière-grand-père sur un blog en les décryptant et en les illustrant avec des documents d'archives diverses. Transcription, commentaire et mise en forme éditoriale constituent ici le cœur de l'activité.

Figure 28. Les carnets de Joseph Duchêne, site d'Aline Duchêne

Inversement, d'autres sites ont une portée plus large qu'un seul régiment. Par exemple, le site Indre 14-18 tenu par Jérôme Charraud, modérateur du forum, retrace l'historique de tous les régiments de l'Indre, avec comme souci de reconstruire des données précises sur la contribution de l'Indre à l'effort de guerre.

Figure 29. Indre 14-18, site de Jérôme Charraud

De même, le site Chtimiste a une vocation plus généraliste. Ce site décrit les parcours de plus de 500 régiments, identifiant les combats et batailles qui bénéficient de description précises. Par ailleurs, l'auteur du site recueille les journaux de soldats et photos qui proviennent de tous les régiments. Il retranscrit peu à peu les journaux qui lui sont confiés et les met en ligne.

Si les sites de l'armée de terre sont les plus nombreux, certains portent sur les autres armes. Dans le domaine de l'armée de l'air, Albin Denis essaye de reconstituer l'histoire de toutes les escadrilles pendant la guerre, en sollicitant l'aide de volontaires.

Des sites transversaux traitent d'aspects spécifiques. Sur Cartographie 1914-1918, Jean Michel Martin dresse la carte du front jour après jour et nous permet de vivre 100 ans plus tard l'évolution du conflit. Sur Vestiges 1914-1918, André Fribourg photographie sur toutes les zones du front les vestiges de la guerre.

Le site combattant.14-18 animé par Arnaud Carobbi, également modérateur du forum, est quant à lui peu à peu devenu un site de référence sur les parcours des soldats. Ce site a une dimension pédagogique (à l'attention des autres amateurs) puisqu'il décrit la démarche de recherche qu'il faut mener pour reconstruire le parcours d'un combattant, en proposant une analyse critique des sources. Il y poste également des analyses pointues de faits liés à la guerre comme l'usage du nom de soldats morts pour désigner les tranchées, ou les origines géographiques des membres d'un régiment. Dans une autre section, il analyse finement des photographies montrant aussi les limites de l'exercice quand la mémoire s'étiole. Cette réflexion sur la qualité des sources et de l'interprétation est au cœur des préoccupations des membres du forum.

Figure 30. Combattant 14-18, site d'Arnaud Carobbi

Le périmètre de ces sites est donc variable, un soldat, un régiment, un ensemble de régiments… Les plus visibles sur la carte sont les sites régimentaires ou les sites portant sur une thématique plus large. Qu'ils soient circonscrits à un soldat, un régiment ou un ensemble de régiments, ces sites très interconnectés partagent un même positionnement. Ils ont une manière commune de présenter la raison d'être de leur site : ils se disent guidés par un devoir de mémoire et cherchent à éviter à ces combattants une seconde mort, celle de l'oubli.

> « En mémoire des 3500 hommes ayant perdu la vie dans les rangs du 74ème RI » (Stéphan Agosto, © BLEU HORIZON - 74e R.I. ©)

> « Le temps est venu de rendre hommage aux aviateurs de la Grande Guerre. » (Albin DENIS, sur son site dédié aux escadrilles)

Chtimiste, dans sa rubrique « Pourquoi ? » dialogue avec le passé :

> « " On oubliera. / Les voiles de deuil, comme des feuilles mortes, tomberont. / L'image du soldat disparu s'effacera lentement dans le cœur consolé de ceux qu'il aimait tant. / Et tous les morts mourront pour la deuxième fois." Ces lignes de Roland DORGELES résument pourquoi, vous, moi, et d'autres travaillons à leur mémoire pour les générations futures. / Pour ne pas oublier… (Chtimiste) »

Denis Delavois sur http://amphitrite33.canalblog.com/ dédié au 149ème Régime d'Infanterie écrit :

> « Ce blog est dédié aux humbles qui auraient certainement préféré continuer de pousser la charrue et cultiver leurs champs pour nourrir les leurs. Aux artistes qui ne purent jamais achever leurs œuvres. Aux savants qui gardèrent leurs inventions dans un petit coin de leur tête et qui ne purent jamais les réaliser, et à tous les autres… »

Le Web est ainsi mobilisé pour faire revivre dans le présent le passé de ces soldats, pour retracer leurs parcours, pour les rendre présents à notre mémoire, pour éviter

qu'ils ne sombrent définitivement dans l'oubli. Ce travail de mémoire s'appuie sur des recherches historiques et mobilise des savoirs pointus. Il s'agit d'une mémoire externe, reconstituée à partir de sources primaires : les traces laissées par les soldats (photos, carnets, lettres…). Bien que les auteurs de ces sites n'aient pas eu une expérience directe du conflit, ils y sont subjectivement reliés au travers de ce travail de lecture, conservation et publication des traces laissées par ces soldats. Cette mémoire externe est reliée à une mémoire interne. Les échanges et les partages sur le forum montrent comment ce réseau d'amateurs tend à construire une mémoire collective de ce passé en articulant travail individuel et échanges collectifs.

Alors que tous les témoins ont disparu, ces mémoriaux numériques constituent des lieux d'un genre nouveau qui tentent de reconstituer au plus près, en s'appuyant sur des sources ce qu'a pu être l'expérience vécue. Il n'y a donc pas antinomie entre mémoire et histoire : le travail historiographique, documenté, sourcé étant mis au service de la mémoire. Cette motivation se retrouve également chez les participants du forum. Ils proposent une histoire de la guerre centrée sur les hommes qui y ont participé. La dimension humaine et la compassion pour l'expérience vécue sont des valeurs partagées par beaucoup. La Première Guerre mondiale a été la première guerre « démocratique », en ce sens qu'elle a concerné toute une génération, qui a partagé la même expérience et s'est construite autour de cette mémoire collective. Cette génération a entièrement disparu et ces historiens cherchent à reconstituer cette expérience à travers les médiations documentaires. Il s'agit d'une mémoire externalisée qui se reconstitue à travers des « lieux de mémoire parce qu'il n'y a plus de milieu de mémoire » (Nora, 1984). Dans l'espace numérique, il s'agit d'un réseau de lieux de mémoire, activé par le réseau de relations entre ces gardiens de la mémoire. En ce sens, au travers de leurs échanges et de leurs sources publiées et partagées, ces acteurs reconstituent en ligne un milieu pour faire revivre cette mémoire : il s'agit donc de lieux de mémoire inscrits dans un milieu de mémoire.

Ainsi le forum et ce réseau de sites personnels très connectés, proches dans leur manière d'aborder la mémoire de la guerre et de mener un travail historiographique constituent un collectif d'amateurs, une communauté de pratique (Wenger, 1998) visible par les proximités dans la manière de mener des recherches et d'écrire sur la guerre. Le forum constitue pour eux le lieu de formation d'une culture commune et partagée.

5.3.3 Des sites personnels ou associatifs hors du collectif des amateurs

Comme nous l'avions vu au début de cette section, certains sites personnels ont une position à l'écart de ce collectif d'amateurs. Ils figurent sur la carte dans la zone institutionnelle.

Prenons le site Association 1914-1918. Cette association rassemble des passionnés, publie une revue, met en relation différentes associations… Il s'agit sans doute d'un collectif dont les contours sont différents de celui du forum, avec sans doute moins de

visibilité en ligne. Il est positionné au nord-est de la carte car il renvoie à tous les sites institutionnels mais aussi à toutes les associations dont on avait vu qu'elles figuraient plutôt au nord-est de la carte. Le site Picardie 14-18, porté par un amateur, constitue un exemple d'une autre nature : son objectif est de recenser tous les lieux de mémoire de la guerre en Picardie, il est donc fortement connecté à tous les sites institutionnels liés aux mémoriaux mais en revanche il ne participe pas au collectif des amateurs que nous avons identifié. Il en est de même pour le site Marsouins et Bigors (troupesdemarine.org) consacré aux troupes de marine de l'armée de terre. Il s'agit d'un site non officiel d'une association d'anciens militaires, qui se situe sur la carte à proximité de tous les sites du ministère de la défense : là encore ils sont très éloignés du collectif.

Dernier exemple, le site Sources de la Grande Guerre, animé par un enseignant-chercheur et un archiviste. Il s'est spécialisé dans la présentation didactique de ressources numérisées sur la guerre. Il s'agit d'un site non institutionnel qui se trouve à proximité des sites institutionnels puisqu'il les cite abondamment. Du fait de leur positionnement sur les sources, les auteurs sont peu impliqués dans le forum.

Figure 31. Sources de la Grande Guerre

Ainsi un certain nombre de sites personnels et associatifs ne participent pas au collectif d'amateurs autour du forum et constituent des espaces alternatifs de production de discours sur la guerre. Bien qu'ils ne soient pas élaborés dans un cadre professionnel, ces sites sont proches de la sphère institutionnelle en raison de leurs objectifs et de leur positionnement. Il n'y a donc pas de coïncidence entre les sites personnels et le collectif des amateurs que nous avons identifié, même si la grande majorité des sites se trouve bien dans la zone amateurs. La seule qualification des auteurs d'un site ne suffit pas à identifier une communauté amateur, seul le réseau de sociabilité définit les contours du collectif.

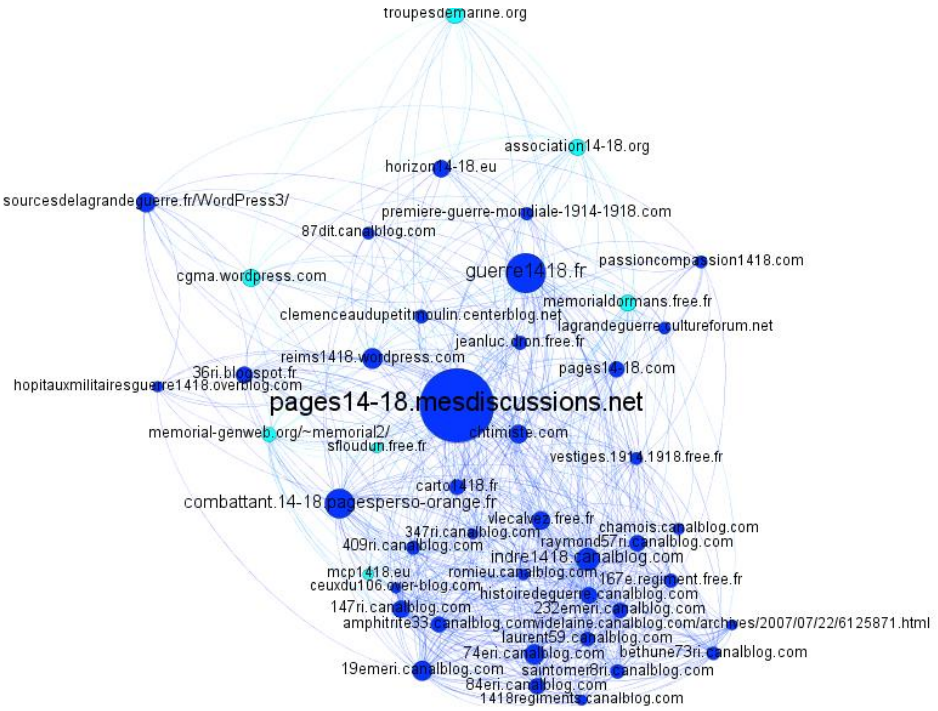


Figure 32. Sites personnels et d'associations

Enfin, un dernier site occupe une place particulière : un internaute a ainsi élaboré un site qui fait office de portail de ressources sur la guerre (Figure 33). Comme il l'indique dans l'espace contact » de son site : « Guerre1418.fr est une initiative personnelle, n'est pas une association et n'a aucun rapport avec une quelconque institution ». Ce répertoire de grande ampleur qui répertorie de très nombreux sites est relativement peu cité par les autres sites.

Figure 33. Portail personnel de ressources sur la Guerre 14-18 (mars 2015)

5.3.4 Les sites de généalogie

L'étude des profils des personnes inscrites au forum, l'analyse des échanges dans le forum (cf. Volume II) et les entretiens nous ont montré que les participants étaient motivés par deux centres d'intérêt majeurs : la généalogie et l'histoire. Les sites de généalogie sont donc d'une grande importance et pourtant ils sont invisibles ou peu visibles sur la carte. Cela tient à des difficultés techniques. En effet, les sites commerciaux ne peuvent pas être archivés car ils ne peuvent être « crawlés » sans inscription, c'est ainsi que Geneanet ne figure pas alors qu'il est souvent mentionné dans les entretiens. Par ailleurs, le site MemorialGenWeb, site collaboratif et amateur de généalogie militaire, a connu pendant la période de la collecte des changements d'hébergement ce qui fait qu'il est particulièrement mal représenté : il s'est dissocié de FranceGenWeb, site de généalogie qui l'accueillait, et s'est autonomisé en 2015. Geneanet et MemorialGenWeb devraient être beaucoup plus visibles sur la carte.

Le portail MémorialGenWeb, qui se situe clairement du côté du pôle « amateur », constitue une ressource fondamentale pour les recherches généalogiques sur la Grande Guerre. Ce site, créé en 2000 par Eric Blanchais, visait à construire une base de données des Morts pour la France « au travers de relevés de monuments, plaques et cimetières militaires ». Comme nous l'explique Alain Girod qui en a été un des administrateurs, des bénévoles parcourent les villages et villes de France et font le relevé des lieux commémoratifs portant des noms (monuments aux morts, plaques commémoratives, vitraux dans les églises, plaques dans les établissements publics) et des tombes dans les cimetières : ils répertorient ainsi tous les morts pour la France, fiches qui sont ensuite reportées en ligne et complétées par les participants. En cela, ils réitèrent une partie du travail de recension fait à l'issue de la guerre. 32 bénévoles travaillent au fonctionnement du site (vérification des données avant publication) et ce dernier indique que depuis les origines 5000 contributeurs ont participé à l'élaboration des 84 621 relevés et 3 165 991 fiches que contient le site. Un formulaire permet de retrouver facilement par nom ou par lieu géographique toutes les données disponibles.

Figure 34. Page d'accueil de Memorial GenWeb (Janvier 2016)

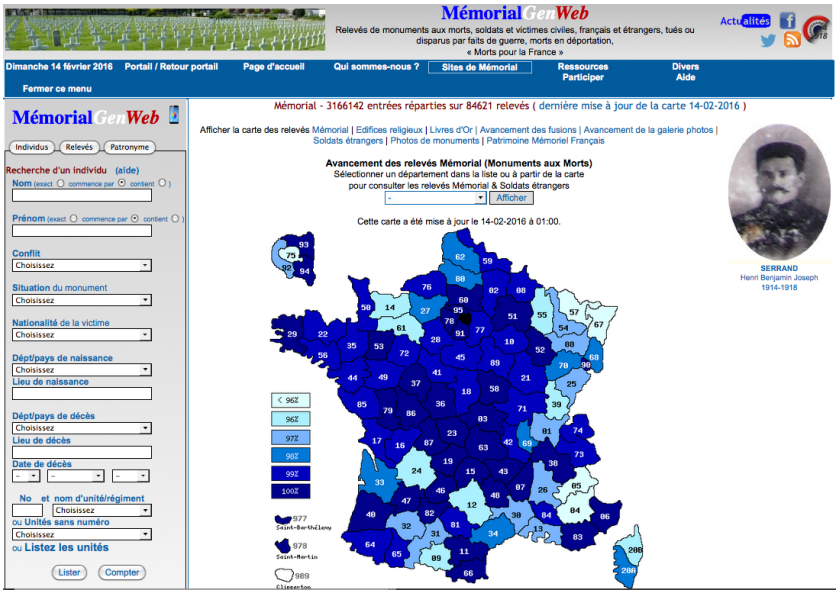


Figure 35. MémorialGenWeb, formulaire de recherche

Comme le rappelle Alain Girod, ce site a été précurseur en étant le premier site des « Morts pour la France », dès 2000, avant Mémoire des hommes. Initié par un réseau d'amateurs, ce projet de numérisation a connu une phase d'institutionnalisation avec la mise en place de Mémoire des hommes par le ministère de la Défense, qui n'a cependant pas remplacé MémorialGenWeb. En se tournant vers les amateurs pour la transcription des fiches, le site institutionnel crée une nouvelle forme d'association avec les publics, qui ont répondu au travers du projet Un jour un poilu. La trajectoire bricolage-institutionnalisation souvent considérée comme une trajectoire « naturelle » des processus de numérisation est en fait plus complexe car elle est marquée par des formes de dépendance mutuelle entre institutions et amateurs. Des formes d'entraide se mettent en place entre l'Etat et ses citoyens sur ces questions mémorielles.

Un second site de généalogie, privé et payant, joue aussi un rôle pour les recherches autour de la Grande Guerre, il s'agit de Geneanet qui propose une section dédiée à 14-18 : Nos ancêtres dans la Grande Guerre. Il est absent de la carte car, comme tous les sites commerciaux, il ne peut être archivé.

En revanche, le site du Centre de Généalogie de Maison Alfort (CGMA) fait partie des nœuds importants de la carte, parce qu'il référence beaucoup de sites. En revanche, il n'est quasiment jamais cité et il n'a d'ailleurs pas été mentionné dans les entretiens.

Figure 36. Geneanet, Nos ancêtres dans la Grande Guerre

5.4 Les sites patrimoniaux comme entrepots de documents

Jusqu'à présent, nous avons examiné la structure de la carte en intégrant les liens sortants et entrants. En changeant de point de vue et en se limitant aux liens entrants (par combien de sites, le nœud est cité), la configuration de la carte change (cf. Figure 37). Deux sites y acquièrent une grande visibilité: Mémoire des hommes et Gallica, qui se situent à la zone frontière entre la zone institutionnelle et amateur.

Les sites d'institutions patrimoniales mettent à disposition du public des archives numérisées : Mémoire des Hommes, Gallica, l'ECPAD, L'Argonnaute etc. Pour certaines de ces institutions, l'approche du Centenaire a conduit à une accélération de la mise en ligne des documents. Par exemple, dans la perspective du Centenaire, la BDIC a mis en place en novembre 2014 une nouvelle bibliothèque numérique, l'Argonnaute, et elle a entrepris le travail de numérisation et de mise en ligne de tous les albums Valois (cf. rapport phase 3 de Lionel Maurel). L'accès aux archives se fait par des interfaces dédiées, certaines se présentant comme de véritables « bibliothèques numériques », qui ne pointent pas vers d'autres sites mais qui sont très cités tant par les sites officiels que personnels.

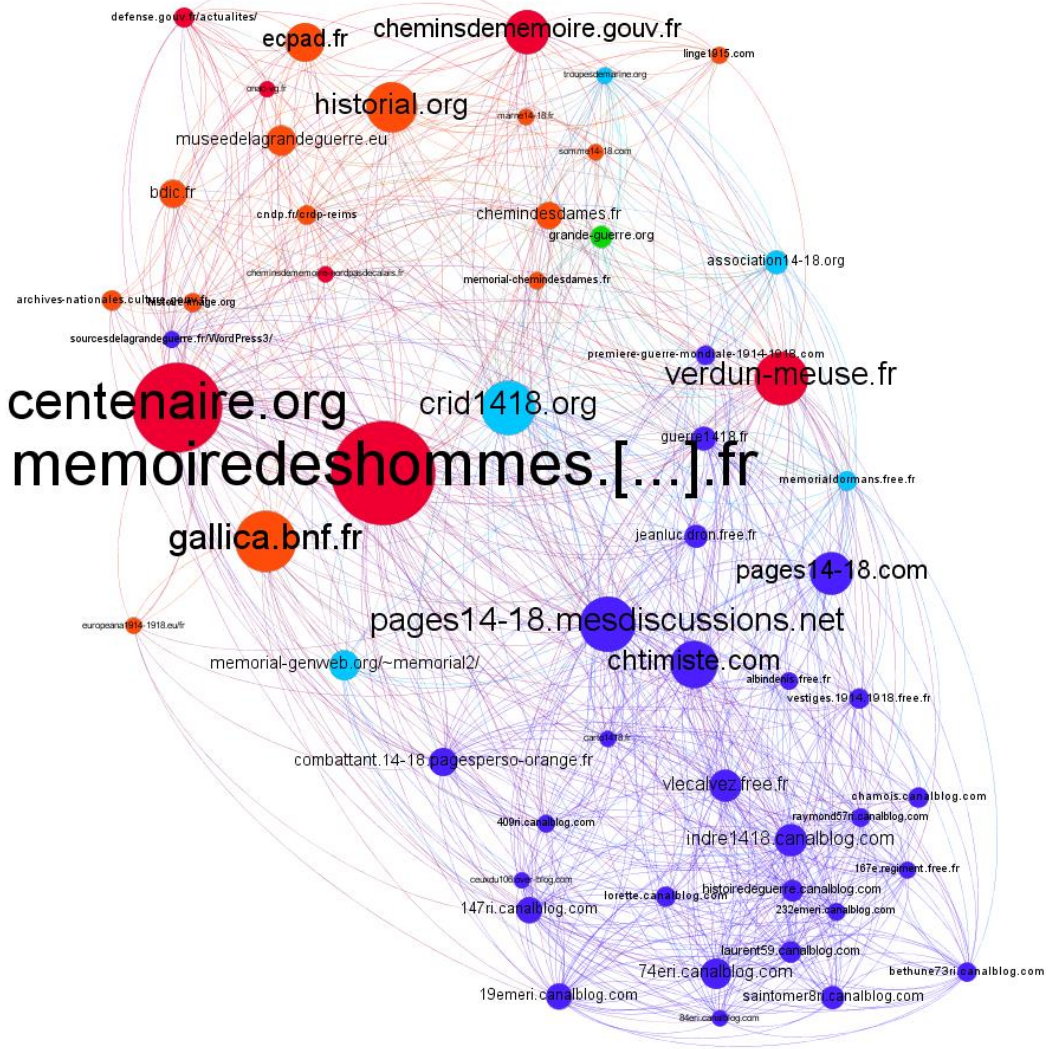


Figure 37. Carte des liens entrants du Web Grande Guerre (collecte de novembre 2014, sites avec degrés entrants > 20)

Mémoire des hommes, site du ministère de la Défense, lancé en 2003, met à « disposition du public des documents numérisés et des informations issues des fonds d'archives et des collections conservés par le ministère de la Défense ». Pour le premier conflit mondial, il donne accès, entre autres, aux fiches individuelles de tous les soldats « Morts pour la France », à la base des fusillés et à tous les Journaux des Marches et des Opérations (JMO) des différents régiments. *Mémoire des hommes* constitue une ressource de première importance pour les recherches généalogiques et pour les recherches sur les régiments et le déroulement des combats, fort prisée par les chercheurs amateurs, ce qui explique sa position centrale. Il s'agit évidemment

d'un site institutionnel mais comme il est très utilisé par les amateurs de généalogie et d'histoire, il est « attiré » par la zone amateur.

Figure 38. Page d'accueil du site Mémoire des hommes (Août 2015)

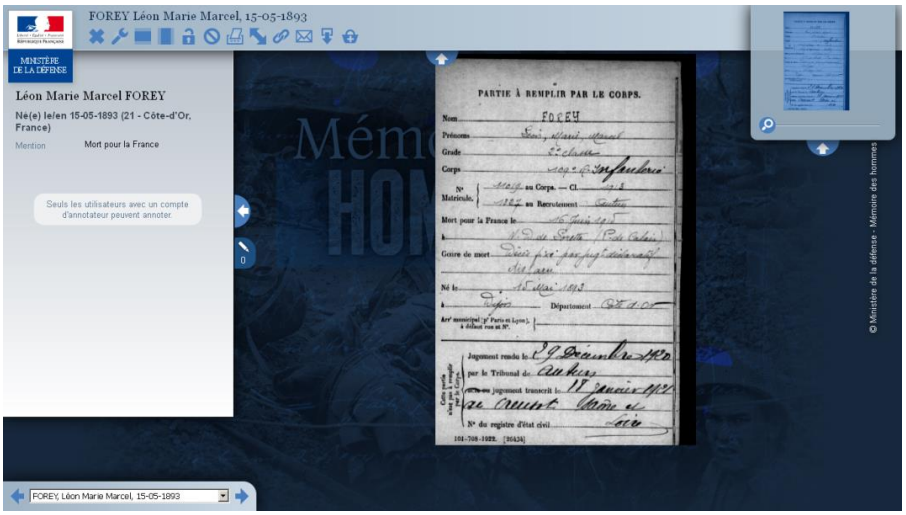


Figure 39. Fiche Mort pour la France, Mémoire des hommes (août 2015)

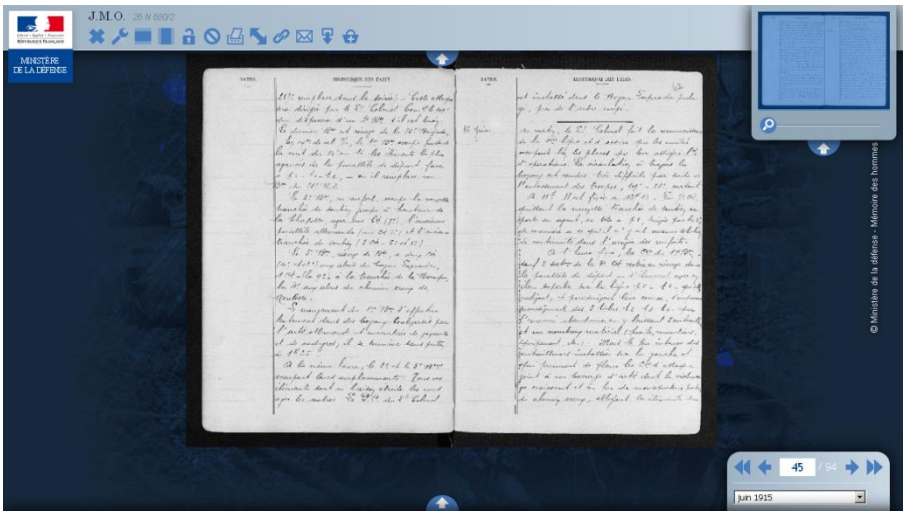


Figure 40. Journal des Marches et des opérations, Mémoire des hommes (Août 2015)

Gallica, la bibliothèque numérique de la BnF (Bibliothèque nationale de France), comprend des fonds essentiels pour l'histoire de la guerre : les journaux officiels, les journaux de tranchée, les historiques régimentaires, des fonds photographiques d'agences… La BnF a également organisé avec le ministère de la Défense une exposition à visée pédagogique sur la Grande Guerre dans la perspective du Centenaire.

La BDIC *(*Bibliothèque de documentation internationale contemporaine), avec les historiques régimentaires, les collections de photographies (albums Valois) est également une source importante pour les chercheurs. L'ECPAD, Etablissement de Communication et de Production Audiovisuelle de la Défense, met en ligne ses fonds audiovisuels sur la Grande Guerre. Leur position sur la carte montre cependant que ces sources sont moins utilisées par les amateurs.

Les sites patrimoniaux qui proposent des ressources documentaires occupent une place tout à fait unique dans notre carte à l'interface entre les sites institutionnels et les sites amateurs. Les ressources proposées sont principalement mobilisées par des amateurs qui deviennent des médiateurs et diffuseurs de ces documents du passé.

5.5 Proprietes des nœuds

Après ce parcours de la carte, qui nous a permis d'en explorer les différentes régions, revenons à quelques indicateurs de graphes pour situer les sites les uns par rapport aux autres. La répartition des liens entrants et sortants pour les principaux nœuds permet de voir la spécificité des sites « entrepôts de documents », comme Mémoire des hommes et Gallica, qui sont la cible de beaucoup de liens mais ne pointent pas vers d'autres sites. A l'autre extrême, certains sites peu cités pointent vers les ressources en ligne, comme guerre1418.fr ou reims1418, ou même Sources de la

Grande Guerre. Entre ces deux catégories, on a des sites qui jouent le rôle de médiateurs, et qui sont des lieux de passage.

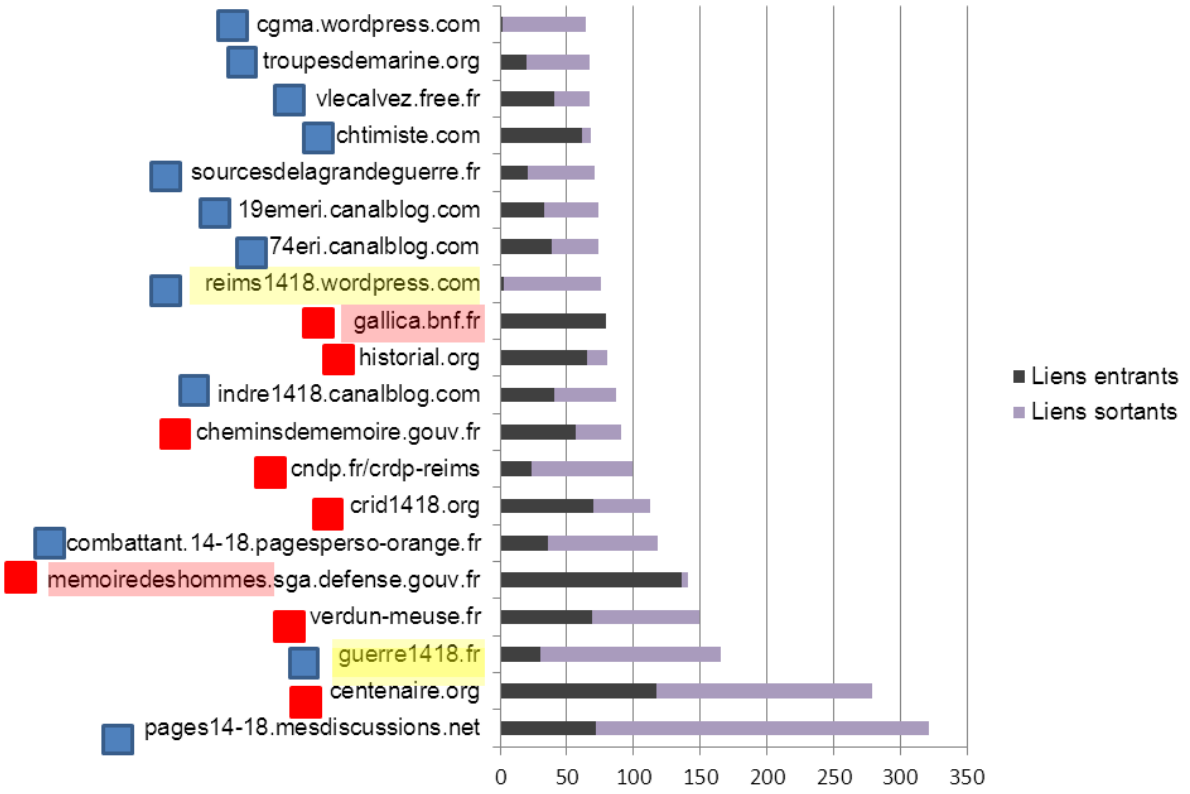
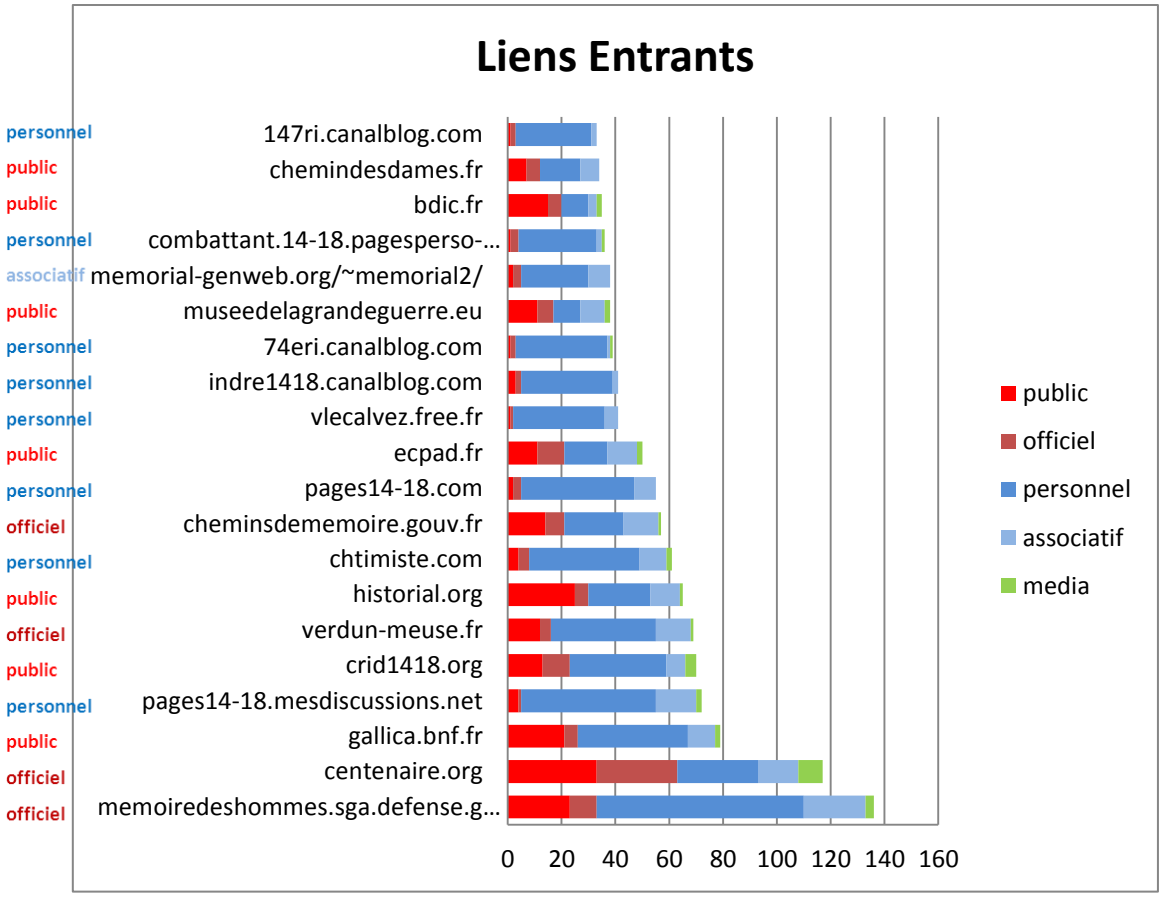


Figure 41. Répartition liens entrants/sortants pour les sites les plus reliés du graphe

La répartition des liens entrants selon l'origine (Figure 42) montre que les sites personnels et associatifs sont les plus gros pourvoyeurs de liens, ils jouent clairement le rôle de médiateurs vers l'ensemble des ressources liées à la Guerre, que ce soit les sites d'archives, les sites de référence ou les autres sites personnels et associatifs. Ils contribuent beaucoup plus fortement que les sites institutionnels à tisser le réseau des associations entre les sites portant sur la Grande Guerre.



Sur Mémoire des hommes, 23 liens proviennent de sites publics, 10 de sites officiels, 70 de sites personnels…

Figure 42. Répartition des liens entrants selon la source

Dans les travaux sur la structure en réseau du web, il est classique de distinguer des sites qui sont des « hubs » (sites qui pointent vers beaucoup de sites, comme les portails) et des autorités, des sites qui sont cités par de très nombreux sites et qui jouent le rôle de site de référence. Les tableaux suivants montrent les résultats de l'algorithme HITS (cf. Encadré : Propriétés des graphes) et listent les autorités (par ordre décroissant) et les « hubs ».

Noeud	Authorities
memoiredeshommes.sga.defense.gouv.fr	0,019990014
centenaire.org	0,013534165
crid1418.org	0,012751409
chtimiste.com	0,012584487
gallica.bnf.fr	0,012441366
verdun-meuse.fr	0,011960576
pages14-18.mesdiscussions.net	0,011619414
pages14-18.com	0,011009076
indre1418.canalblog.com	0,010503854
cheminsdememoire.gouv.fr	0,010253901
historial.org	0,010152366
vlecalvez.free.fr	0,009665257

Noeuds	Hub
pages14-18.mesdiscussions.net	0,037974206
centenaire.org	0,027616889
guerre1418.fr	0,025174395
combattant.14-18.pagesperso-orange.fr	0,017520824
verdun-meuse.fr	0,016331408
reims1418.wordpress.com	0,015186535
cgma.wordpress.com	0,014578688
cndp.fr/crdp-reims	0,014462776
36ri.blogspot.fr	0,013947398
indre1418.canalblog.com	0,013880982
crid1418.org	0,012807842
sourcesdelagrandeguerre.fr/WordPress3/	0,012229963

Tableau 4. Autorités et 'Hubs' : Top 12

Quelques sites ont le double statut de site de référence (autorité) et de site portail (hub) comme le forum, le site du Centenaire, le site régional officiel Verdun-Meuse, le site de recherche du Crid 14-18 et le site personnel dédié aux soldats de l'Indre.

Enfin, la comparaison des indicateurs de densité entre les différentes zones de la carte montre que la zone des amateurs (degré moyen de 13 et densité de 0,036) est beaucoup plus dense que la zone institutionnelle (degré moyen de 7 et densité de 0,019). Les sites des associations ont une densité encore plus faible (degré moyen de 3 et densité de 0,18) ce qui explique qu'ils disparaissent de la carte quand le filtre se fait par le degré.

5.6 Le point de vue des acteurs sur la carte

La carte (Figure 19) représente à une date donnée (novembre 2014) les sites web consacrés à la Grande Guerre à travers leur réseau de liens : seuls sont visibles les sites inscrits dans un réseau de relations. Les sites isolés, même si leur audience peut être importante, comme les dossiers de presse, en sont absents. Cette carte rend compte du réseau des liens existant entre les sites et en même temps elle dessine à l'adresse du visiteur les chemins potentiels de visite.

Proposer aux participants de lire et de commenter la carte est un moyen de donner sens à cette cartographie. Cette représentation est-elle interprétable ? La différenciation des territoires institutionnel et amateur est-elle pertinente ? Les hiérarchies entre les sites sont-elles conformes à leurs représentations ? Ce point de vue des acteurs a déjà été intégré dans les sections précédentes, par exemple en redonnant une place aux sites de généalogie mal traités dans la collecte.

Face à la carte, les participants sont nombreux à souligner le caractère enchevêtré des points et des lignes et à la trouver difficile à lire. « C'est marrant parce que je ne sais pas trop comment on lit, mais avec le gribouillis que c'est et avec des interactions partout, c'est curieux. » (SA). L'un face à la carte rappelle une anecdote passée : « C'est un général américain en Afghanistan, je ne sais plus comment il s'appelle, je crois que c'est Petraeus. Et on lui présente un truc comme cela et il dit : 'je ne suis pas venu manger des spaghettis' » (IM).

5.6.1 La polarité de la carte

Certains réagissent avec des « Je ne suis pas étonné » (JC) comme si la représentation révélait des choses évidentes. Le clivage de la carte est clairement identifié par les participants : « Le bleu et le rouge je le vois bien, il y a quand même cette séparation-là » JC ; « C'est deux mondes, c'est-à-dire qu'il y a peu de liens entre les deux, c'est cela » SA. EP synthétise le point de vue, en y lisant une opposition entre monde professionnel et amateur :

> « On a une espèce de web professionnel et on aurait un web amateur, le CRID est bien placé, il est pile à l'interface des deux. [...] Mémoire des hommes c'est pareil, c'est les usages qui peuvent être des usages enseignants de Mémoire des hommes et Page 14 18, effectivement il y a pas une reconnaissance de la part des sites plus académiques de l'importance de ce type de pratiques, et après il y a les sites régimentaires. » EP, Crid 14-18

En examinant la zone « amateur », les participants au forum reconnaissent un environnement familier de sites et de personnes qu'ils fréquentent, de ressources qu'ils utilisent :

> « Oui, c'est tous les blogs dont je parlais, il y en a qui me sont familiers. [...] Ne serait-ce parce qu'ils interviennent sur le forum et parce qu'ils ont l'adresse de leur blog en signature, donc cela apparait à chaque fois. » (YV)

« Il y a des sites que je connais, il y a des copains. [...] c'est pareil, c'est des affinités, c'est normal parce que eux ils ont un peu la même approche régimentaire on va dire, même si après chacun le fait à sa sauce, [...] oui je les connais tous, là [...] Oui, j'ai identifié des copains, les copains de la tranchée ils sont là. (SA)

On notera que la proximité intellectuelle « une même approche régimentaire » se double d'une proximité affinitaire « les copains de la tranchée », connaissance et amitié se trouvent ainsi liées dans ce type de collectif en ligne. Les entretiens montrent d'ailleurs que les relations en ligne sont couplées à des rencontres, à des visites des lieux.

« Il faut savoir que dans les premières années de page 14 18, on avait des voyages, on s'est retrouvé à Notre Dame de Lorette en 2006, on s'est tous retrouvé, on a bouffé ensemble, on a visité le champ de bataille, c'était organisé par les gars du cru, donc des mecs du pas de calais qui nous ont fait visiter des champs de bataille et on est au delà de la liaison internet pure et dure, c'est des amis. » (JC)

Certains sites peuvent être reconnus en tant que ressource même si leur auteur ne participe pas au forum :

« le Chtimiste qui est un site important. Ah oui, c'est encore un type qui a fait un travail de folie [...] il a quand même fait plusieurs centaines de régiments, pas loin de 500 en comptant l'artillerie, parce qu'il a fait l'artillerie. [...] c'est un site de référence, si vous voulez savoir à quel endroit était ce régiment à telle date, moi je vais sur le chtimiste. (...) Il est très souvent cité, mais non il n'intervient pas. » (IM)

En revanche en examinant la zone nord de la carte, le sentiment de familiarité et de connaissance intime s'amoindrit.

« Donc là c'est beaucoup plus institutionnel, là c'est Historial, que j'adore, cela je ne connais pas, Mémoire oui, Chemin des dames oui, là je connais moins, et là c'était un ancien forum, l'association 14 18, il y avait un forum avant là-dessus qui existe toujours d'ailleurs. » (SA)

Les sites sont reconnus, mais pour la plupart sont peu utilisés. Ainsi, à propos des sites institutionnels sur le Centenaire :

« C'est vrai que je ne suis pas tellement allé voir [...] Je pense que c'est assez destiné au grand public. [...] Oui, j'ai déjà vu des sites comme cela, qui répertorient, et personnellement je n'en ai pas trop besoin (YV) »

Les sites institutionnels situés à la frontière, Mémoire des hommes et Gallica, sont clairement les plus utilisés par tous.

Pour les participants au forum, ce sont les sites qui sont proches du forum sur la carte qui sont les plus familiers ou les plus fréquentés. La représentation des liens entre les sites nous donne finalement une approximation des usages, du moins pour ceux qui participent à la construction de la mémoire de la Grande Guerre en ligne.

5.6.2 Les absents de la carte

Cependant, l'absence de certains sites a été soulignée par la plupart des personnes rencontrées : des sites qu'ils utilisent beaucoup pour leurs recherches n'apparaissent pas sur la carte.

Le recours à des sites étrangers est évident lorsque les recherches portent sur des armées étrangères étant intervenues sur le territoire français. C'est par exemple le cas pour ADC spécialisé dans les armées canadiennes en France.

Beaucoup insistent en général sur l'importance des sites étrangers pour mener leurs recherches.

> « Parce que vous avez des sites hollandais qui curieusement sont très forts sur 14 18, il faut comprendre le flamand, que je ne comprends pas, mais souvent quand je fais des recherches, je tombe chez eux, les anglais, ah les anglais, les américains pas du tout, très peu, quasiment pas, les allemands je ne sais pas, je suis tombé parfois sur des trucs mais je n'ai jamais été plus loin mais par contre il y a des étrangers qui viennent sur le forum, pas beaucoup mais il y en a, on a une partie internationale. » (IM)

Leur absence dans la collecte, qui est centrée sur le périmètre français, constitue une limite pour la recherche : les contraintes juridiques liées à l'archivage du web conduisent à traiter de manière nationale et locale un conflit dont la dimension est fondamentalement internationale. Une initiative portée par IIPC vise justement à croiser les archives du web de plusieurs pays.

Pour les recherches généalogiques, le site de la croix rouge (CICR, http://grandeguerre.icrc.org/fr) qui recense tous les prisonniers de guerre, quelle que soit leur nationalité, constitue une ressource de première importance dont l'absence sur la carte est déplorée. La localisation en Suisse de l'association empêche son archivage.

Parmi les sites très utilisés par les membres du forum et peu visibles sur la carte, figurent les sites d'archives départementales qui permettent d'accéder aux fiches matricules des soldats, documents essentiels pour reconstruire le parcours d'un combattant. Ces fiches matriculaires ont été numérisées et mises en ligne dans une grande partie des départements, et l'approche du centenaire a clairement accéléré le processus. L'absence des archives départementales sur la carte tient à leur éclatement par département et à la grande hétérogénéité des solutions techniques choisies pour la mise en ligne.

Enfin, comme nous l'avions indiqué, l'absence ou la faible visibilité des sites de généalogie comme MemorialGenWeb et Geneanet est souvent soulignée car ceux-ci sont des ressources importantes pour nombre de participants.

> « Mémorial Genweb, je suis étonné qu'il soit aussi peu gros.[…] je sais qu'il est beaucoup cité au même titre que le Chtimiste, parce que quand les gens cherchent une tombe ils vont sur Mémorial Genweb » (IM)
>
> « Je regarde aussi sur un site de généalogie qui s'appelle Geneanet, c'est gratuit, quelqu'un m'avait dit c'est pas mal, j'ai tapé les départements, j'essaye de croiser. […] C'est un site de généalogie donc j'ai mis un arbre généalogique, c'est gratuit, je crois qu'il y a des fonctions qui peuvent être payantes, […] et ils ont un espace, un mini site sur la guerre de 14 et ce qu'on peut faire c'est que les personnes qui sont sur notre arbre généalogique, on peut noter qu'elles ont participé à la guerre de 14, dans tel régiment ». (AD)

Finalement, les utilisateurs se font une représentation de l'espace du web qu'ils fréquentent : ils en reconnaissent les contours, distinguent les zones, identifient ce qu'il manque sur la carte.

5.6.3 Le Web Grande Guerre et la commémoration

Ce projet s'étant déroulé pendant la période du Centenaire de la Grande Guerre, il nous parait utile de spécifier l'incidence de cet anniversaire sur nos observations. Si le forum Pages 14-18 a une longue histoire et qu'il est devenu au fil des années un site de référence sur la Grande Guerre en France, le site du Centenaire, a quant à lui été mis en œuvre pour le Centenaire : la Mission étant circonscrite à la période 2014-2018, on peut faire l'hypothèse qu'il ne jouera plus le même rôle une fois la mission achevée au terme de la commémoration.

Les participants reconnaissent que le Centenaire semble avoir joué un rôle important pour dynamiser l'activité autour de la Grande Guerre tant dans l'effort de mise en ligne de documents :

> « effectivement il y a eu une impulsion politique quand même qui a été là vraiment pour dire qu'il serait bien que certains documents [comme les fiches matricules] soient mis en ligne de façon un peu anticipée » SA

que dans l'effort d'organisation d'événements au niveau local :

> « La Mission du Centenaire en accordant son patronage a boosté des opérations et même en n'accordant pas de subventions, avec cet accord de patronage, ils ont incité des communes qui n'avaient jamais rien fait à se réveiller et à faire des choses. » AG

En même temps, les amateurs insistent sur le fait que le Centenaire n'a rien changé pour eux, qu'ils s'y intéressaient avant et qu'ils continueront à s'y intéresser.

> « Moi personnellement cela n'a rien changé pour moi » SA

> « j'ai eu plus d'activités au niveau des conférences et au niveau des sollicitations pour des renseignements, plus depuis le centenaire, sinon je n'en ai pas fait ni plus ni moins qu'avant le centenaire. » AG

La question de l'après Centenaire se pose déjà. Les participants ne s'inquiètent pas pour la poursuite de leur engagement en tant que collectifs d'amateurs, mais s'interrogent sur le devenir des initiatives plus institutionnelles.

> « Oui, il y a énormément de choses, je pense que cela profitera peut-être un peu, une fois que 18 va être terminé, on en parlera beaucoup moins, mais j'espère quand même qu'il subsistera quelque chose, que cela ne va pas avoir été que des actions pendant 4 ans pour après ne plus rien faire » AG

> « centenaire.org on a le côté évènementiel, effet 2014, globalement qu'est ce qui va se passer en 2018, on va dissoudre et les gentils fonctionnaires qui y sont, vont retourner à d'autres activités ou d'autres missions. Pages 14-18, c'est le côté historique, on était là avant, on sera là après, j'espère ». JC

Tout se passe comme s'il y avait un temps long de l'engagement chez les passionnés non professionnels qui est parfois opposé au temps éphémère des commémorations.

6 CONCLUSION

Au terme de ce parcours, revenons sur les éléments principaux.

Grâce à la collecte Grande Guerre mise en place par la BnF, qui vise à archiver les sites web liés à cette thématique, nous avons pu cartographier le réseau des relations entre les sites liés à la Grande Guerre, tels qu'ils étaient en novembre 2014.

L'analyse de la collecte montre tout d'abord l'existence d'une intense activité numérique autour de ce premier conflit mondial : plus de 500 sites en France (sans tenir compte des réseaux sociaux, que nous n'avons pas inclus dans ce projet) ont été identifiés dont une grande majorité est produite de manière bénévole en dehors du champ des institutions, en dehors d'un cadre professionnel, témoignant d'un engagement fort d'une communauté d'amateurs.

La proximité relationnelle des sites sur la carte, qui dépend de la structure des liens, dessine le territoire numérique du web dédié à la Grande Guerre. Ce territoire n'est pas homogène, il est marqué par une partition et des mécanismes de polarisation. L'activité de mémoire autour de la Grande Guerre est structurée autour de deux pôles, l'un institutionnel et l'autre amateur. Il ne s'agit pas de deux mondes complètement séparés ou cloisonnés, mais bien de deux polarités avec une zone frontière particulièrement intéressante.

A côté de l'activité institutionnelle autour de la mémoire de la guerre (renforcée dans le cadre de la commémoration), on a pu mettre à jour la vivacité d'une activité de recherche qui implique des chercheurs, historiens et généalogistes pour la plupart non professionnels. Cet espace très actif et très interconnecté regroupe des sites web organisés autour d'un forum qui rassemble un collectif très dense de personnes qui échangent et apprennent ensemble. Les sites sont souvent guidés par un devoir de mémoire (garder le souvenir des soldats qui ont combattu, leur éviter une seconde mort). Ce travail de mémoire s'appuie sur des sources, des archives officielles et personnelles combinées pour construire des mémoriaux numériques d'un genre nouveau. Il est intéressant de noter que la construction de ces monuments virtuels dédiés aux combattants d'un régiment d'infanterie, ce qui constitue le cas le plus fréquent, est à l'articulation entre un travail individuel et un travail collectif : en effet, les échanges dans le forum sont le lieu de la construction d'une manière commune d'aborder le passé et surtout d'un savoir commun dans la méthode de recherche. En ce sens, il y a bel et bien une mémoire collective qui se construit entre le forum et les sites reliés. Bien que le lien organique à l'événement vécu ait totalement disparu, il s'agit pour les auteurs de ces sites de reconstruire une mémoire sur la base de toutes les traces laissées par les soldats : il n'y a plus de lien d'expérience direct avec l'événement, mais un lien subjectif et un attachement reconstruits grâce à pratique et la fréquentation des sources premières qu'elles soient physiques ou numériques.

Si ces sites amateurs sont rarement cités par les sites institutionnels, ils jouent à l'égard de certains de ces derniers un rôle central. En effet, ces amateurs assurent un rôle de médiation vers les ressources des sites institutionnels : ils font découvrir les fonds, aident à les explorer, pointent vers les ressources. Ils jouent le rôle de connecteurs en mettant en relation les individus avec les documents du patrimoine numérisé. Cette activité explique la position frontière de sites comme Mémoire des hommes ou Gallica, entrepôts de documents de première importance pour l'étude de la guerre.

En termes de méthode, ce travail de cartographie du web a été possible dans des conditions de sérieux et de reproductibilité attendues, grâce à l'utilisation des archives du web de la BnF. En effet, le Web Grande Guerre, comme tous les espaces du Web vivant, se transforme en permanence. Ce phénomène « naturel » est plus intense encore autour de la Première Guerre mondiale en raison du contexte de la commémoration : renouvellements et apparition de nouveaux sites transforment le paysage en profondeur… La constitution d'archives permet de travailler sur un corpus délimité à une date donnée et de pouvoir retourner aux sources, plus précisément aux sites tels qu'ils étaient au moment de la collecte, pour pourvoir vérifier ou approfondir les analyses. La recherche est reproductible et d'autres chercheurs pourront retravailler sur ce même corpus et sur les collectes suivantes pour évaluer les transformations de l'espace dédié à la Grande Guerre sur le Web.

En ce qui concerne l'analyse de l'évolution de la structure du graphe dans le temps, nous avons rencontré une difficulté épistémologique majeure que nous n'avions pas anticipée. L'existence de collectes menées à dates régulières devait permettre de représenter l'évolution du réseau de sites au fil du temps. Cependant, il est impossible de démêler si les transformations proviennent de l'évolution même du web ou des observateurs. Cette indécidabilité rend l'approche diachronique hasardeuse. Par ailleurs, si on peut comparer les principaux indicateurs de graphes (densité, modularité), les comparaisons de graphes ne sont pas aisées.

Travailler sur les archives du Web permet de construire le corpus à partir d'une collecte de qualité, bien délimitée et archivée. En dépit de ces conditions idéales de travail, la structure même du web fait que toute une série de choix ont dû être faits pendant la chaine de traitement, qui ont de toute évidence une incidence sur la représentation proposée : sélection, filtrage… En même temps, le choix s'est porté vers une représentation qui permettait de visualiser aisément des phénomènes qui avaient une certaine forme de stabilité et de rémanence dans toutes les analyses. Il nous semble que les options qui ont été prises dans ce travail de simplification de la représentation sont aujourd'hui reproductibles et pourront être utilisées dans d'autres projets de recherche. C'est pourquoi nous les avons présentées de manière détaillée dans la section 4.

Nous voudrions revenir sur la nécessité d'inscrire le travail de fouille de données du web dans le cadre d'une démarche de sciences sociales pour que celui-ci puisse avoir du sens. De toute évidence, des hypothèses provenant des sciences sociales guident la manière d'explorer les données : prétendre que la démarche est purement inductive est insatisfaisant. Les hypothèses sont nourries par les travaux menés dans le champ de la sociologie de l'internet et dans celui des travaux sur la mémoire.

Par ailleurs, il nous a semblé essentiel d'adopter une démarche qualitative pour confronter la visualisation du graphe à la représentation qu'en ont les parties prenantes. Dans ce contexte, les entretiens qualitatifs ont été une occasion de proposer à des personnes qui participent à la fabrique du Web Grande Guerre de commenter et de réagir à la carte en la confrontant à leur propre représentation. Des interventions dans des colloques et des rencontres institutionnelles (Historial de Péronne, Mémorial de Verdun…), organisées par Philippe Chevallier, ont permis d'échanger avec des acteurs institutionnels ce qui a permis de confirmer certains faits rendus visibles par la carte et d'en corriger d'autres. Enfin, il nous faut souligner que pour donner sens à la carte, une exploration ethnographique des sites principaux de la carte a été indispensable : qui en est l'auteur, à qui s'adresse-t-il, avec qui est-il en relation, quel type de ressources utilise-t-il, comment met-il en scène ses contenus ? Sans cette étape fondamentale de retour aux données initiales, la construction du sens et de l'interprétation est peu fiable. L'exploration des données est enrichie tout du long par d'autres formes de connaissance qui permettent de faire sens.

Pour finir, il parait utile de rappeler que ces approches impliquent de la pluridisciplinarité (la connaissance du domaine en sciences humaines – sociologie des usages et de la mémoire), la maitrise de l'outil informatique pour extraire et préparer les données, la maîtrise des méthodes de fouille de données, numériques, textuelles et relationnelles pour traiter les corpus et la maitrise des méthodes d'enquête et d'entretien) et par conséquent beaucoup de collaboration. Ce type de projet a pu voir le jour grâce à la complémentarité des équipes et des profils impliquées dans le projet.

7 BIBLIOGRAPHIE

AKRICH M., 2010, « From Comunities of Practice to Epistemic Communities Mobilizations on the Internet », *Sociological Research Online*, *15*, 2, p. 10.

AMAR M., CHEVALLIER P., 2013, « Rapport d' étude sur les usages des corpus numérisés de Gallica sur la Grande Guerre », Paris.

ARGUELLO J., BUTLER B., JOYCE E., KRAUT R., LING K.S., WANG X., 2006, « Talk to Me : Foundations for Successful Individual-Group Interactions in Online Communities », *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems.*, p. 959‑968.

AUDOIN-ROUZEAU S., 2008, *Combattre. Une anthropologie historisque de la guerre moderne (XIXe-XXIe siècle)*, Les livres, Paris, Seui.

AURAY N., HURAULT-PLANTET M., POUDAT C., JACQUEMIN C., 2009, « La négociation des points de vue : une cartographie sociale des querelles dans le Wikipedia francophone », *Réseaux*, *n°27*, 152-4, p. 15‑50.

BEAUDOUIN V., 2012, « Trajectoires et réseau des écrivains sur le Web. Construction de la notoriété et du marché », *Réseaux*, *vol. 30*, n° 175, p. p. 107‑144.

BEAUDOUIN V., VELKOVSKA J., 1999, « Constitution d'un espace de communication sur Internet (Forums, pages personnelles, courrier électronique…) », *Réseaux*, *vol. 17*, n° 97, p. p. 121‑177.

BIRNBAUM J., 2006, « 1914-1918, guerre de tranchées entre historiens », *Le Monde*, 2006, p. 1‑5.

BOULLIER D., LEVY J., 2016, « topographies/topologies. Langages spatiaux, spatialités, espaces », *Réseaux*, *34*, 195, p. 9‑162.

BRÜGGER N., 2013, « Historical Network Analysis of the Web », *Social Science Computer Review*, *31*, 3, p. 306‑321.

CARDON D., FOUETILLOU G., ROTH C., 2014, *Topographie de la renommée en ligne*, 85 p.

CARDON D., LEVREL J., 2009, « La vigilance participative. Une interprétation de la gouvernance de Wikipedia », *Réseaux*, *27*, 154, p. 51‑90.

CLAVERT F., 2016, « #ww1. The Great War on Twitter », *Digital Humanities 2016: Conference Abstracts.*, p. pp. 461‑462.

CONEIN B., LATAPY M., 2008, « Les usages épistémiques des réseaux de communication électronique : Le cas de l'Open-Source », *Sociologie du Travail*, *50*, 3, p. 331‑352.

DEPS, 2001, « Les publics des archives départementales et communales », *Développement culturel*, *137*, octobre 2001.

DONNAT O., 2011, « Pratiques culturelles, 1973-2008 Dynamiques générationnelles et pesanteurs sociales », *Culture études. Politiques publiques et régulation*, 7.

FLICHY P., 2010, *Le sacre de l'amateur*, Seuil La R, Paris.

FOLLIC A. LE, STIRLING P., WENDLAND B., 2012, « Putting it all together: creating a unified web harvesting workflow at the Bibliothèque nationale de France », *IIPC*.

FREEMAN L.C., 1977, « A Set of Measures of Centrality Based on Betweenness », *Sociometry*, *40*, 1, p. 35‑41.

GENSBURGER S., 2005, « Essai de sociologie de la mémoire : le cas du souvenir des camps annexes de Drancy dans Paris », *Genèses*, *4*, p. 47‑69.

GHITALLA F., DIEMERT E., MAUSSANG C., PFAENDER F., 2004, « Tarente: an experimental tool for extracting and exploring web aggregates », *Information and Communication Technologies: From Theory to Application*, p. 627‑628.

GIBSON D., KLEINBERG J.M., RAGHAVAN P., 1998, « Inferring web communities from link topology », *Proceedings of the ninth ACM conference on Hypertext and hypermedia: links, objects, time and space---structure in hypermedia systems: links, objects, time and space---structure in hypermedia systems*, p. 225‑234.

GUIGUENO B., 2015, « Qui sont les publics des archives? Enquêtes sur les lecteurs, les internautes et le public des activités culturelles dans les services publics d'archives (2013-2014). »

HALBWACHS M., 1994, *Les cadres sociaux de la mémoire.*, Paris, Albin Michel (1ère édition, 1925, Librairie Alcan).

HALBWACHS M., 1997, *La mémoire collective*, Paris, Albin Michel (1ère édition, 1950).

HARVEY F., 2005, « La généalogie et la transmission de la culture : une approche sociologique », *Les Cahiers des Dix*, 59, p. 285‑305.

HEATON L., MILLERAND F., CRESPEL É., PROULX S., 2011, « La réactualisation de la contribution des amateurs à la botanique », *Terrains & travaux*, *1*, 18, p. 155‑173.

HEYMAN S., GRAND B. LE, 2014, « Investigation visuelle d'événements dans un grand flot de liens », *EGC 2014*.

JACOMY M., VENTURINI T., HEYMANN S., BASTIAN M., 2012, « ForceAtlas2 , A Continuous Graph Layout Algorithm for Handy Network Visualization », *PLoS ONE*, *9*, 6, p. 1‑22.

KOTRAS B., 2013, « Des opinions qui comptent. Influence, visibilité et hiérarchisation des opinions sur le web », *Colloque international sur l'e-réputation : « e-réputation et traces numériques : dimensions instrumentales et enjeux de société »*.

KOTRAS B., 2015, « L'opinion autorisée », *Revue d'anthropologie des connaissances*, *9,3*, 3, p. 311.

LAVABRE M.-C., 2016, « La "mémoire collective" entre sociologie de la mémoire et sociologie des souvenirs ? », *<halshs-01337854>*.

LAVE J., WENGER E., 1991, *Situated learning : legitimate peripheral participation*, Cambridge, Cambridge University Press.

LEDOUX S., 2012, « Écrire une histoire du « devoir de mémoire » », *Le Débat*, *170*, 3, p. 175.

LICOPPE C., BEAUDOUIN V., 2002, « La construction électronique du social : les sites personnels. L'exemple de la musique. », *Réseaux*, *Vol. 20*, n° 116, p. p. 53‑96.

LOEZ A., 2008, « Le site web du Crid 14-18 », *Matériaux pour l'histoire de notre temps (BDIC)*, *3*, 91, p. 90‑91.

MICHEL J., 2015, « Mémoire publique et mémoire collective de l'esclavage. », 1994, p. 1‑23.

NORA P., 1984, « Entre mémoire et histoire: La problématique des lieux », dans *Les lieux de mémoire I. La République*, Paris, Gallimard, p. XVII - XLII.

OFFENSTADT N., 2010, *14-18 aujourd'hui - La Grande Guerre dans la France contemporain*, Paris, Odile Jacob.

OLICK J.K., ROBBINS J., 1998, « Social Memory Studies: From "Collective Memory" to the Historical Sociology of Mnemonic Practices », *Annual Review of Sociology*, *24*, 1, p. 105-140.

RICOEUR P., 2003, *La mémoire, l'histoire, l'oubli*, Paris, Seuil.

ROBIN R., 2003, *La mémoire saturée*, Stock.

ROUSSO H., 1987, *Le syndrome de Vichy, de 1944 à nos jours*, Paris, Seuil.

ROUSSO H., 2016, *Face au passé. Essais sur la mémoire contemporaine*, Paris, Belin.

SAEMMER A., 2015, *Rhétorique du texte numérique : figures de la lecture, anticipations de pratiques*, ENSSIB.

SAGNES S., 1995, « De terre et de sang : la passion généalogique », *Terrain*, *septembre*, n° 25, p. 125-146.

SAGNES S., 2004, « Cultiver ses racines : mémoire généalogique et sentiment d'autochtonie », *Ethnologie française*, *XXXIV*, n° 1, p. 31-40.

SCHNEIDER S.M., FOOT K.A., 2004, « Web Sphere Analysis: An Approach to Studying Online Action », *Virtual Methods: Issues in Social Science Research on the Internet*, March, p. 1-16.

SEVERO M., VENTURINI T., 2016, « Enjeux topologiques et topographiques de la cartographie du Web », *Réseaux*, *34*, 195, p. 84-105.

TRUC G., 2015, « Venir à Ground Zero, se souvenir du 11-Septembre.Peer review », *Espacestemps . net*, p. 1-20.

WENGER E., 1998, *Communities of Practice: Learning, Meaning, and Identity*, Cambridge, U.K., Cambridge University Press.