Corpus d'enquêtes sur les pratiques d'information scientifique des chercheurs. Constitution et exploitation des données

Florence Thiault, Marie-Laure Malingre

Florence Thiault et Marie-Laure Malingre, « Corpus d'enquêtes sur les pratiques d'information scientifique des chercheurs. Constitution et exploitation des données », *Revue française des sciences de l'information et de la communication* [En ligne], 24 | 2022, mis en ligne le 01 janvier 2022, consulté le 24 mars 2022. URL : http://journals.openedition.org/rfsic/12228

Corpus d'enquêtes sur les pratiques d'information scientifique des chercheurs Constitution et exploitation des données

Contexte du projet

Les pratiques (numériques, informationnelles, communicationnelles, …) à l'œuvre dans le secteur de la recherche et de la Science Ouverte sont extrêmement diverses selon les objets de recherche, les champs de savoir, les méthodologies, les activités, les modalités de valorisation… L'information scientifique et technique (IST) connaît des mutations importantes en lien avec la diffusion globale du numérique (techniques, outils, supports, usages) et du web dans les processus de la recherche, modifiant profondément les pratiques. Cette transformation se joue à un double niveau : dans la manière de faire de la recherche et dans les modalités de diffusion des résultats de la recherche. Pour interroger les activités d'information scientifique numérique, différents angles d'approche sont possibles : la numérisation et la question des corpus numériques, les outils et interfaces, plateformes et dispositifs web (archives ouvertes, plateformes de revues, réseaux sociaux académiques), les formats et langages, et enfin les pratiques de diffusion et de valorisation de la recherche, individuelles ou collaboratives. Ce renouvellement des questionnements est présent dans les enquêtes qui visent à caractériser les pratiques informationnelles des chercheurs à l'ère numérique. Stéphane Chaudiron et Madjid Ihadjadene (2008) définissent la notion de pratiques informationnelles en ces termes : « la manière dont l'ensemble des dispositifs, des sources, des outils, des compétences cognitives est effectivement mobilisé dans les différentes situations de production, de recherche, traitement de l'information ». Les auteurs englobent aussi bien dans le terme de pratiques : « les comportements, les représentations que les attitudes informationnelles de l'humain (individuel ou collectif) associés à ces situations » (2010). L'étude des pratiques informationnelles se centre donc sur les préoccupations et les compétences des chercheurs en considérant les interfaces, les systèmes et les dispositifs de d'information et de communication.

Il existe aujourd'hui une grande variété d'enquêtes sur les pratiques en information scientifique et technique (IST) des chercheurs portant sur des périmètres et des objets différents (supports, disciplines ou champs de savoir, type de public, …). Cependant, la plupart des études sur les pratiques informationnelles des chercheurs sont très localisées et de ce fait difficiles à généraliser et à conceptualiser comme le souligne Annaïg Mahé (2012) dans un panorama des enquêtes publiées lors de la décennie 2000-2010. Les modalités d'investigation y sont dans de nombreux cas hétérogènes et s'appuient sur des approches méthodologiques qui restent spécifiques. Ces enquêtes essaient de rendre compte de l'hétérogénéité de la pratique de recherche en fonction d'habitus disciplinaires et méthodologiques précis (Meunier et al., 2014).

Elles étudient les changements de granularité dans le traitement de l'information scientifique, transformations à la fois gouvernées par la logique propre de l'activité scientifique et par la généralisation du numérique. Quant aux enquêtes les plus récentes, elles cherchent à identifier des évolutions notables liées au contexte de la Science Ouverte (Couperin, 2020 et SOSP-FR, 2021).

Ces enquêtes la plupart du temps auto-administrées se fondent sur une approche déclarative individuelle de l'analyse de l'activité du chercheur. Notre projet, engagé en 2018, a été retenu dans le cadre de la mise en place du « GIS « Réseau URFIST ». Ce groupement d'intérêt scientifique a vocation à encourager le développement de la recherche au sein des différentes URFIST, acteurs de la formation à l'information scientifique. Le projet présenté dans ce data paper a pour ambition de recenser les principales enquêtes menées en France et à l'étranger sur les pratiques informationnelles des chercheurs en matière d'IST, afin d'en tirer divers enseignements notamment sur les méthodologies retenues, les acteurs et les thématiques abordées. Cette étude s'inscrit dans le prolongement d'actions engagées au sein de la Bibliothèque Scientifique Numérique (BSN 9 : Formation, compétences et usages) sur les formations à l'IST et des travaux menés aujourd'hui au sein du Comité pour la Science Ouverte (COSO, collège compétences et formation).

La description de l'approche expérimentale reviendra sur les choix et ajustements méthodologiques successifs de l'équipe projet.

2. Constitution du corpus, méthodologie et description des jeux de données

Nature du / des corpus et génération de données

Les enquêtes relatives aux pratiques d'Information Scientifique et Technique (IST), diverses et hétérogènes dans leurs modalités comme dans leurs approches méthodologiques et leurs visées, sont un objet qu'il est, de ce fait, souvent difficile d'appréhender et d'analyser, en particulier si l'on veut adopter une démarche comparative et globalisante. Dégager, à partir des enquêtes qui ont pu être menées en France et à l'international, un cadre conceptuel et méthodologique pour l'observation des pratiques des chercheurs relatives à l'IST - tel est l'objectif que s'est donné notre projet – suppose de procéder à une analyse fine de la structure de ces enquêtes en s'éclairant de leurs postulats, de leurs principes opérationnels, de leurs résultats, ainsi que de tous documents associés.

Pour ce faire, l'étude envisagée s'appuie sur différents dispositifs de collecte et de production de données : dans un premier temps, sur la constitution d'un corpus d'enquêtes qu'il s'agit de recenser, d'indexer et d'observer selon des critères définis en amont pour en tirer tous les enseignements utiles. L'état des lieux ainsi réalisé, couplé à un état de l'art de la littérature spécialisée et enrichi d'un ensemble de métadonnées et d'annotations, fait l'objet d'une bibliothèque Zotero partagée qui constitue un premier livrable et fichier de données *(cf. infra tableau des jeux de données, 1)*.

Cet état des lieux est également à la base d'une analyse statistique initiale portant sur le profil général et les caractéristiques principales du corpus. Les graphiques et tableaux qui en résultent (données produites par l'étude), incluant leur interprétation, ont vocation à fournir de nouveaux fichiers de données *(cf. infra tableau des jeux de données, 2)*.

Par ailleurs, il s'est vite avéré essentiel de disposer d'un espace dédié pour stocker et valoriser les documents associés à la description bibliographique des enquêtes : l'ensemble des questionnaires ainsi archivés constitue un corpus secondaire. Le choix des membres du projet[1] s'est porté sur un espace Nakalona (associant bibliothèque numérique Omeka[2] et plateforme d'exposition de données Nakala[3], portée par la TGIR Huma-num[4]). En effet, outre que l'accueil des questionnaires d'enquête y est facilité par un plugin d'import zotero, les fonctionnalités proposées permettent de bénéficier d'une mise en forme des collections numériques et de l'enrichissement des métadonnées. Il est à noter que le pack logiciel Nakalona a été remplacé fin 2020 par Nakala-Press, système de publication de contenus intégré à la nouvelle version de Nakala.

La propriété intellectuelle des fichiers concernés étant extérieure, ce nouveau corpus se révèle plus limité que le précédent, dans la mesure où il correspond aux seules enquêtes dont les autorisations de diffusion et de réutilisation ont été obtenues de leurs auteurs *(cf. infra tableau des jeux de données, 3).*

Globalement, la caractérisation et la description des données collectées et produites pour le projet peuvent être synthétisées comme suit dans le tableau 1 :

[1] Présentation du projet sur le site du GIS « Réseau URFIST » http://gis-reseau-urfist.fr/enquete-ist/. Les membres du projet : M.-L. Malingre, F. Thiault, A. Serres (URFIST de Bretagne et Pays de la Loire), C. Denecker (URFIST de Lyon), G. Gallezot (URFIST Méditerranée), A. Mahé (URFIST de Paris).
Trois stagiaires en master Humanités numériques (Université Rennes 2), L. Le Rolland-Raumer, A. Ferret, L. Bégasse ont contribué au traitement des données.

[2] Logiciel libre de création et gestion de bibliothèque numérique, développé par le Center for History and New Media (CHNM) de l'Université George Mason, Etats-Unis, https://omeka.org/

[3] Interface d'exposition, de valorisation et de partage des données, conçue par Huma-Num, l'infrastructure pour les Humanités numériques, https://nakala.fr/

[4] TGIR Huma-Num, Très Grande Infrastructure de Recherche des humanités numériques, ayant pour objectif de construire une infrastructure numériques de niveau international pour les SHS, https://www.huma-num.fr/

Jeux et fichiers de données (collectées et produites) : types, supports, formats			
Réalisé			
	Type de données	Support	Format
1	Corpus d'enquêtes : recensement bibliographique	Bibliothèque Zotero	Fichier au format RDF Fichier au format CSV URL
		Fichier texte	Fichier au format PDF (les notices à la norme ISO 690)
2	Matériels complémentaires (Données et graphiques d'analyse du corpus, modèle autorisation de diffusion/réutilisation)	Fichier texte	Fichier final PDF
3	Fichiers des questionnaires et scénarios d'entretiens avec droits de diffusion et réutilisation	Collection Omeka / Exposition des données sur Nakala	Fichier au format CSV et URL
Reste à réaliser			
4	*Données d'analyse issues de la fouille de texte sur les contenus relatifs à la méthodologie d'enquête*	*Fichier iramuteq, visualisation de données avec gephi*	*Format iramuteq (.ira) et format de visualisation gephi (GEXF)*
5	*Base de questions / base de connaissance*	*Fichier final : base MySQL*	*Fichier final : base MySQL*

Tableau 1 : présentation des jeux de données

Il est prévu de produire de nouveaux jeux de données à l'issue de la dernière phase de l'étude, non encore finalisée, et qui consiste, d'une part dans la fouille (TDM) et l'interprétation des contenus textuels relatifs aux méthodologies des enquêtes sélectionnées, et, d'autre part dans l'appariement des questions constitutives des questionnaires d'enquête au sein d'une base de connaissance et leur analyse comparative. L'objectif est de construire une réflexion méthodologique pour l'observation des pratiques informationnelles des chercheurs (pratiques IST) et d'en déduire des préconisations en la matière *(cf. tableau 1 des jeux de données, 4 et 5).*

Méthodologie de construction du corpus

Le périmètre

Les membres du projet ont tout d'abord travaillé sur l'élaboration d'une démarche méthodologique centrée sur le mode de constitution du corpus d'enquêtes. La première question était celle du périmètre. Il a été établi que le corpus viserait une recension d'enquêtes et de leur méthodologie en suivant différents critères : une couverture géographique privilégiant de la manière la plus exhaustive possible les enquêtes menées en France, à quoi s'ajoute un échantillonnage d'enquêtes menées à l'étranger en fonction de leur importance estimée (francophonie et essentiellement pays anglophones pour un *benchmarking* des enquêtes) ; une couverture linguistique centrée prioritairement sur le français et l'anglais ; des limites temporelles allant primitivement de 2005 à 2018, dans la mesure où la sélection s'appuyait d'une part sur un dépouillement de l' « Etat des enquêtes sur les pratiques numériques des acteurs de l'enseignement supérieur et de la recherche » réalisé dans le cadre de BSN 9 Lot 3[5] (enquêtes de 2005 à 2015), et d'autre part sur une recherche web exploratoire (pour sélectionner les enquêtes entre 2015 et 2018).

Phase 1 : Définition initiale du périmètre du corpus et établissement de critères de sélection	
Couverture géographique	France > couverture maximale Benchmarking des enquêtes selon leur importance sur la francophonie l'Europe / pays européens ; les autres pays
Couverture linguistique	Essentiellement français et anglais
Limites temporelles	2005-2018
Public des enquêtes	Prioritairement enquêtes ciblant la communauté de la recherche : chercheurs, enseignants-chercheurs, doctorants, ingénieurs de recherche. Sélection d'enquêtes s'adressant aux personnels de soutien / personnels IST et susceptibles de donner à un niveau méta des indications sur les pratiques des chercheurs
Thématique des enquêtes	Pratiques IST > informationnelles, documentaires, scientifiques (pratiques de publication, d'analyse textuelle…), pratiques numériques dérivées.
Profil des pratiques	Pratiques dans un contexte de recherche scientifique / intégration éventuelle de pratiques privées en fonction de l'intérêt de leur apport
Types d'enquêtes	Enquêtes quantitatives (questionnaires) Enquêtes qualitatives (entretiens) Autres, le cas échéant : focus groups, ateliers, études des logs de connexion… Articulation entre les différents types d'enquêtes

[5] Les données de la bibliothèque scientifique numérique ne sont plus disponibles en ligne. Nous nous reportons à un document de travail pour le recensement réalisé dans BSN9.
https://www.ouvrirlascience.fr/evolution-de-la-bsn-vers-le-comite-pour-la-science-ouverte-coso/

Phase 2 : Recensement et premier état du corpus	
Appui sur *« l'Etat des enquêtes sur les pratiques numériques des acteurs de l'enseignement supérieur et de la recherche »* réalisé dans le cadre de BSN9 (Lot 3) : enquêtes 2005-2015	Dépouillement du corpus BSN et sélection des enquêtes en adéquation avec les objectifs fixés
Recherche exploratoire complémentaire pour 2005-2015 et pour 2015-2018	Requêtage web (Isidore, Google Scholar, HAL, bases de données…)
Examen de l'état réalisé	Validation du recensement à partir du large spectre de départ : enquêtes de référence, décision de garder ou non les enquêtes trop spécifiques, ou avec un nombre de répondants trop limité
Phase 3 : extension du corpus primitif	
Mise à jour du corpus pour la période 2018-2021 (recrutement de stagiaire)	Requêtage web

Tableau 2 : étapes de construction du corpus d'enquêtes

Dans la suite du projet, on a procédé empiriquement à une extension du référencement pour prendre en compte la période 2018-2021 et opérer ainsi une mise à jour du corpus (dans l'optique d'une base de connaissance pérenne, il serait souhaitable de poursuivre cette mise à jour). Quant à la cible, le principe a été de retenir en priorité les enquêtes visant les pratiques des chercheurs, des enseignants-chercheurs ou des doctorants, dans un contexte de recherche scientifique (en intégrant le cas échéant les pratiques privées), mais aussi celles s'adressant aux personnels d'appui à la recherche, quand elles sont susceptibles de renseigner indirectement et à un niveau méta sur les pratiques des chercheurs. Si les enquêtes quantitatives par questionnaires ont dès le début constitué un matériau de référence du projet, dans la mesure où elles ont été valorisées et ont joué pendant longtemps un rôle central dans la connaissance des publics, on a souhaité également tirer parti d'autres formes d'investigation, en particulier des études qualitatives par entretiens, plus ouvertes et présentant moins de biais du fait de l'autonomie et de la diversité de parole des enquêtés. Par ailleurs, l'articulation, la combinaison des différentes approches, lorsqu'elles existent, revêtent un intérêt particulier d'un point de vue méthodologique.

La démarche documentaire

La démarche documentaire s'est faite en trois temps : l'idée était premièrement de reprendre et dépouiller le relevé d'enquêtes de BSN 9, pour examiner, dans la liste proposée, les documents qui pouvaient être pertinents pour notre projet et intégrés dans notre propre recensement. Un premier état a pu ainsi être constitué. C'est sur la base de cette sélection initiale que le corpus

a ensuite été complété par une démarche exploratoire sur le web, grâce à des stratégies de requêtage dans un ensemble prédéfini de ressources et d'outils de recherche (Google et Google Scholar, Isidore, HAL, BASE, bases de données spécialisées, bouquets de revues…). Par ailleurs a été engagée une exploration systématique de sites et structures de recherche en identifiant les enquêtes qui pourraient avoir été menées en leur sein. Pour mieux organiser et regrouper les enquêtes, un cadre d'indexation a été élaboré avec un dispositif de mots clés libres ou normalisés et d'annotations permettant de fournir des éléments d'information complémentaires homogènes : mots clés sur le champ disciplinaire concerné, sur le public visé, sur le périmètre, le pays et la langue, sur la méthodologie d'enquête, sur la thématique IST traitée, et précisions textuelles apportées sur le contexte et les modalités des enquêtes (nombre de répondants, taux de réponse…).

Système d'indexation (tags normalisés ou libres)	
Périmètre	périmètre_local périmètre-national périmètre_international
Public	public_chercheurs public_enseignants-chercheurs public_doctorants public_étudiants public_personnels de soutien
Méthodologie d'enquête	méthodo_quantitative (pour les questionnaires quantitatifs) méthodo_qualitative (pour les entretiens, l'observation participante)
Taille du questionnaire *(Au final, dans le fichier texte du corpus, ce type de tag a été abandonné au profit d'une indication du nombre de questions)*	longueur_long longueur_moyen longueur_court
Langue	langue_xxx
Pays	France, Xxx… (tag nom de pays)
Discipline	multidisciplinaire xxx (tag libre selon la discipline)

Types de pratiques	pratiques informationnelles, pratiques numériques, pratiques scientifiques
Thématique de l'enquête	thémaIST_xxx (recherche d'information, données de recherche, fouille de texte, archives ouvertes, réseaux sociaux, veille…)
Système d'annotation (informations complémentaires trouvées)	
Dates précises de l'enquête Objectifs de l'enquête Auteur / Commanditaire de l'enquête Effectifs visés et nombre de répondants (taille de l'échantillon enquêté) / Taux de réponse Type de pratiques (pratiques informationnelles, pratiques numériques, pratiques scientifiques) Autres éléments d'information connus jugés intéressants (par exemple le nombre de questions du formulaire d'enquête)	

Tableau 3 : Système d'indexation

Choix des outils et services de référencement

Très vite, la phase de construction du corpus s'est structurée autour de la constitution d'une bibliothèque Zotero, partagée entre les membres du projet et rendue publique[6].

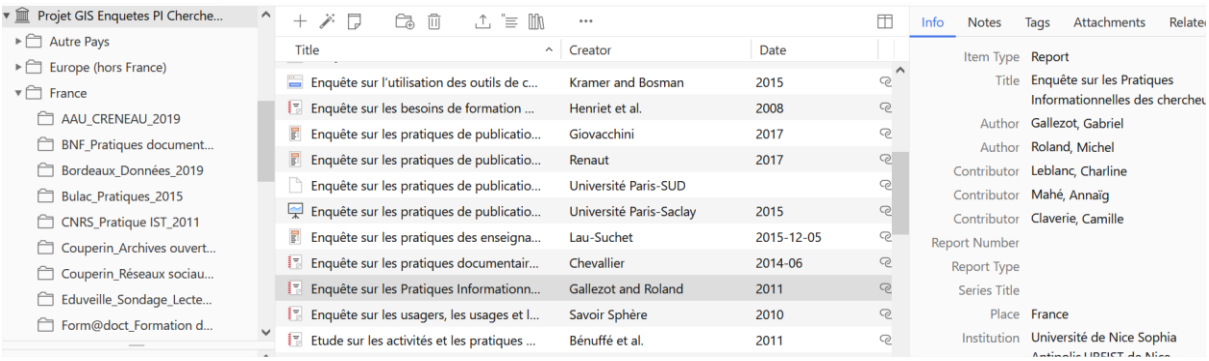

Figure 1 : Capture d'écran bibliothèque Zotero

Cette bibliothèque est le reflet du recensement transcrit primitivement sur un fichier tableur : outre une collection dédiée à l'état de l'art bibliographique, elle comporte les références des enquêtes, classées par commanditaires / structures les ayant conduites, au sein des quatre grandes catégories : France, Francophonie, Europe (hors France), Autres pays. Chaque enquête fait l'objet d'un dossier spécifique (sous-collection Zotero) dans l'arborescence créée.

[6] https://www.zotero.org/groups/2158779/enquetes_sur_les_pratiques_ist_des_chercheurs/library

Les documents source associés aux notices des enquêtes sont quant à eux déposés, lorsque cela est possible, dans la collection numérique créée en lien avec les services d'Huma-Num. A terme, l'ensemble des données générées par le projet sera accessible sur la plateforme Zenodo.

Droits sur les fichiers de données

L'étude projetée fait appel, comme on l'a vu, à plusieurs types et formats de fichiers de données, correspondant à chacune des étapes du processus d'analyse. La diversité et l'origine des données conduisent à distinguer deux situations au regard des droits qui leur sont associés en vue de leur diffusion et de leur possible réutilisation : les données extérieures collectées dans le cadre de la constitution du corpus (essentiellement questionnaires et scénarios d'enquêtes), et les données propres, produites dans le cadre des analyses menées. En ce qui concerne les données extérieures collectées, quatre cas se sont présentés : les fichiers de questionnaires déjà diffusés en ligne avec des licences spécifiant les droits de réutilisation (généralement licences *creative commons*), les fichiers accessibles sans spécification de droits, les enquêtes dont on pouvait totalement ou partiellement reconstruire les fichiers de questionnaires à partir des documents de restitution (synthèse, rapport, communication…), enfin les enquêtes pour lesquelles on ne disposait pas du fichier questionnaire et dont il était impossible d'obtenir une vision suffisamment précise du contenu. Dès lors, les auteurs ou organisateurs des enquêtes ont été sollicités d'une part pour une autorisation de diffusion et de réutilisation lorsque le fichier était disponible, mais qu'aucune possibilité de réutilisation n'était spécifiée (modèle de courrier du Matériel supplémentaire), d'autre part pour avoir communication du fichier et autorisation de réutilisation dans le cas où celui-ci n'était pas accessible initialement. Pour les données produites dans le déroulement du projet, leur régime juridique est régi à travers l'attribution de licences *creative commons*, en l'occurrence : paternité, libre diffusion, modification et réutilisation, pas d'utilisation commerciale, partage dans les mêmes conditions.

Limites dans la constitution du corpus

L'évolution du projet montre à la fois les atouts et les écueils de la démarche adoptée. Si un certain nombre de principes ont été adoptés en amont pour la constitution du corpus d'enquêtes (établissement d'un périmètre, de critères de sélection), ainsi que pour sa description et son indexation, ils sont le résultat de choix (par exemple l'échelle de couverture selon l'origine géographique), et donnent forcément à cette étude sur les enquêtes une orientation où la légitimité du corpus ne relève pas d'une exigence de représentativité, mais d'un regard différencié. Il faut donc tenir compte, pour une future réutilisation, du fait que le corpus ne compose pas une image représentative de l'ensemble des enquêtes réalisées, mais plutôt un outil pour l'étude de la méthodologie d'enquête sur les pratiques IST des chercheurs, démarche qui correspond à l'objectif initial affirmé. Par ailleurs, le corpus a évolué au fil du temps et en fonction des aléas du projet, afin d'une part de rester au plus près de l'actualité du sujet (élargissement des limites temporelles), et d'autre part de procéder à des ajustements successifs (prise en compte de la spécificité ou de l'importance de l'enquête, de la taille de l'échantillon enquêté dans la phase de validation). L'on peut y voir un mode agile comme une fragilité

méthodologique. Enfin, le corpus témoigne sans doute d'une certaine hétérogénéité, qu'il s'agisse de l'envergure des enquêtes, de leurs modalités, ou encore des communautés enquêtées auxquelles il renvoie. Et comparer ce qui n'est pas totalement comparable pourrait facilement apparaître comme une limite à notre d'étude. Cependant, l'état des lieux qu'elle propose vise d'abord, tout en s'éclairant des résultats des enquêtes eux-mêmes, à proposer un cadre conceptuel et méthodologique pour l'étude des pratiques IST des chercheurs. L'analyse de l'hétérogénéité s'avère un atout intéressant pour la compréhension globale des dispositifs mis en œuvre.

3. Analyse du corpus

Pour débuter cette section, nous présentons quelques éléments d'information et d'analyse au sujet du corpus constitué. Ce corpus a été constitué de façon évolutive par étapes successives de validation des entrées, de suppression des doublons et de retrait de références pour différentes raisons. En particulier, certaines enquêtes n'ont finalement pas été retenues quand elles apportaient peu d'informations (faiblesse du nombre de répondants, du questionnement) ou étaient trop spécifiques à un domaine de l'IST.

3.1 Périmètre du corpus

Nous précisons en premier lieu quelques éléments sur la zone géographique concernée par les enquêtes du corpus, puis sur l'échelle des enquêtes (locale, nationale, internationale). La répartition du public enquêté est ensuite interrogée ainsi que l'organisateur commanditaire de l'enquête de manière à les identifier. Le corpus final est constitué de 114 enquêtes. 90 enquêtes, soit 78% du corpus, concernent la France et 25 autres enquêtes des pays à l'international, soit presque 22% (dont 19 pays francophones et 6 pays européens). La figure 2 représente la répartition des enquêtés en fonction du périmètre géographique visé par les enquêteurs.

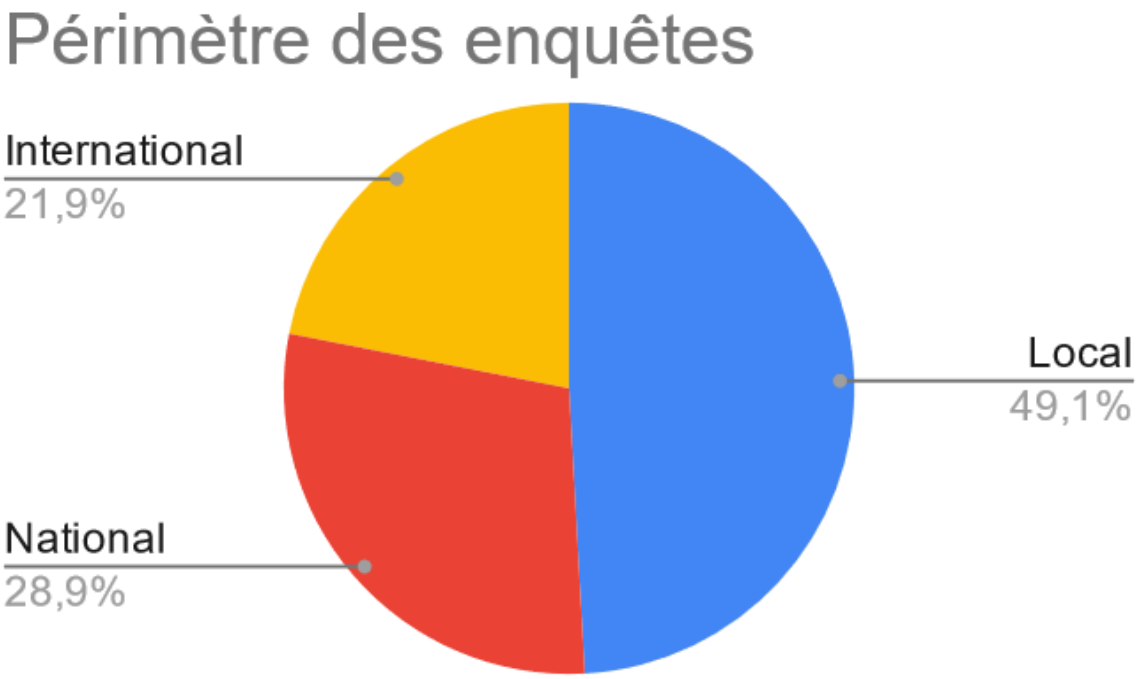

Figure 2 : Le périmètre géographique des enquêtes

Les enquêtes n'ont pas toutes le même niveau d'échelle et ne sont pas comparables. Les enquêtes locales se rapportent à une ou plusieurs institutions d'une région donnée (dont des enquêtes sur un laboratoire). Dans cette catégorie (56 items), on compte 45 études menées sur une seule institution, et 11 études qui portent sur des regroupements d'institutions régionales. L'échelle d'un établissement induit une limitation pour se projeter sur une échelle plus large. La catégorie « Nationale » correspond à une enquête portant sur un pays particulier, ainsi 33 enquêtes se concentrent sur un seul pays (France [24], Etats-Unis [3], Canada [3], Algérie [1], Cameroun [1], Suisse [1]). La catégorie « Internationale » (25 items) intègre les enquêtes portant sur plusieurs pays. L'échelle internationale accentue la difficulté de comparaison des pratiques alors que les contextes nationaux de recherches sont différents.

Quel est le public visé par les enquêtes du corpus ?

Le public enquêté correspond aux individus que les enquêteurs ont ciblés prioritairement en le précisant généralement dans le message d'accompagnement et de présentation de l'enquête. Une enquête peut cibler plusieurs types de publics.

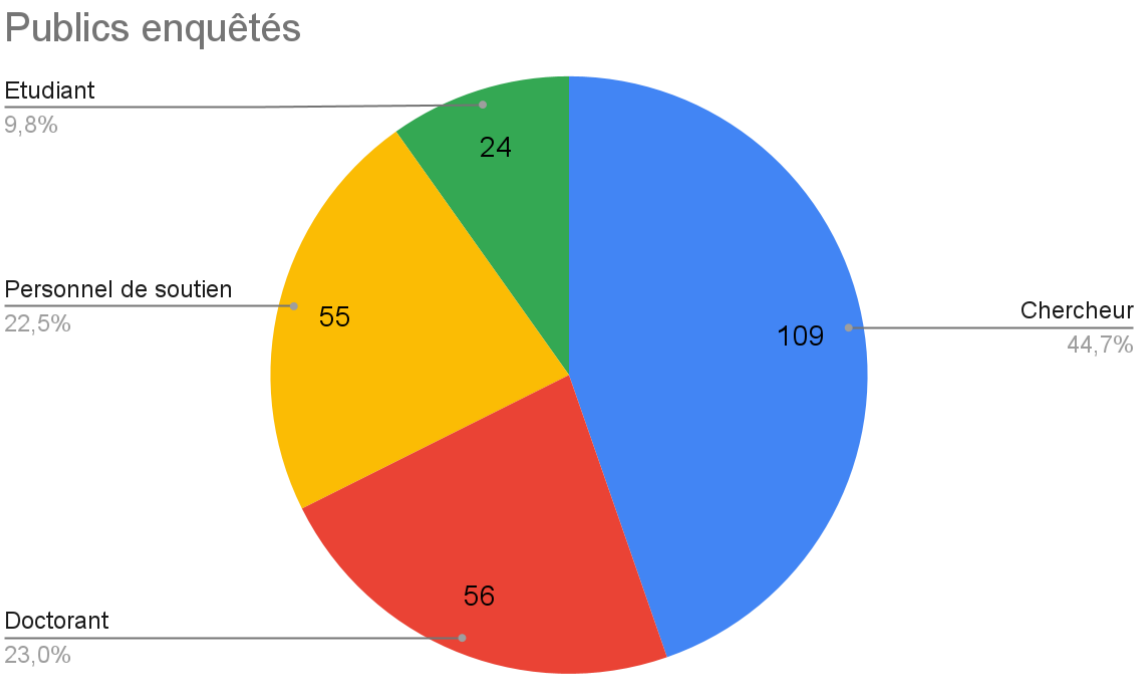

Figure 3 : Catégorie de publics enquêtés

Sans surprise, la catégorie des chercheurs est la plus représentée dans notre corpus (109 enquêtes), suivie par la catégorie des jeunes chercheurs doctorants (56 questionnaires). La catégorie « Personnel de soutien à la recherche » intègre plusieurs types de personnel. Ainsi, 18 enquêtes s'adressent aux bibliothécaires, 15 aux administratifs, et 12 aux ingénieurs. A noter que 28 enquêtes restent vagues sur le personnel interrogé en indiquant « autre profession », elles ont été intégrées dans la catégorie « Personnel de soutien ». En ce qui concerne les étudiants, les enquêtes ne précisent pas généralement le niveau d'étude, cependant 15 enquêtes s'adressent spécifiquement aux étudiants en master. Parmi ces 15 études, 8 visent également les étudiants en licence.

Qui sont les responsables des enquêtes ?

Nous avons cherché à identifier le contexte de la recherche, à savoir qui était responsable du projet d'enquête quand cet élément est précisé. Ce sont majoritairement des chercheurs qui apparaissent comme responsables des études (41 enquêtes), ou des professionnels de l'information-documentation (bibliothécaire et documentaliste, 34 enquêtes).

Le cadre de la recherche est afférent aux différentes institutions liées au projet de l'enquête, que ce soit le commanditaire ou bien l'institution d'appartenance des enquêteurs. Une enquête peut être menée par plusieurs institutions et éventuellement renouvelée dans le temps. Cet élément d'identification n'est pas toujours précisé clairement dans les enquêtes retenues.

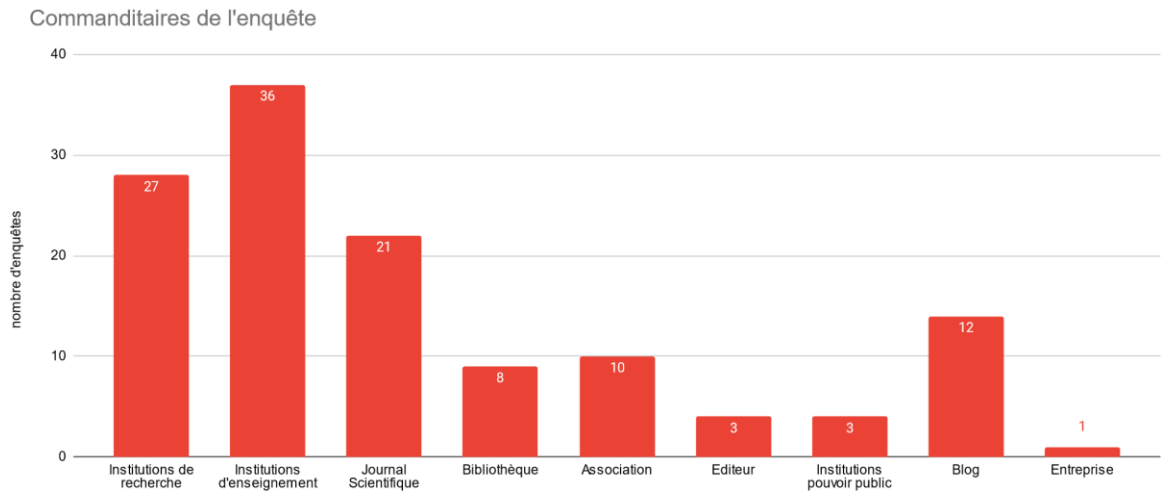

Figure 4 : Commanditaire de l'enquête

On notera que la BNF, en association avec Télécom Paristech, a mené une enquête en deux temps concernant les usages de Gallica (2014 et 2016, projet porté par Valérie Beaudouin). Le consortium Couperin a réalisé deux enquêtes sur les archives ouvertes, une en 2012 et une autre en 2017. Pour le laboratoire GERiiCO (Université Lille 3) qui travaille depuis 2013 sur les données en SHS, les deux enquêtes de 2015 et 2017 (coordination Joachim Schöpfel), les plus importantes sur le sujet, ont été retenues.

Quelles sont les disciplines universitaires concernées par les enquêtes ?

La répartition des enquêtes par discipline est très hétérogène, de nombreuses disciplines sont interrogées. Nous avons choisi de les regrouper en grands champs disciplinaires : STM pour les chercheurs en sciences exactes (14 enquêtes, soit 12% du corpus) ou SHS pour les chercheurs en Art, Lettres et Langues, Sciences humaines et sociales (28 enquêtes, 24,5 %). Cette distribution nous indique le domaine disciplinaire qui est le plus souvent interrogé. Quand une discipline précise n'est pas indiquée dans l'étude, cette dernière est alors considérée comme multidisciplinaire. Les enquêtes multidisciplinaires composent la majorité du corpus (72 enquêtes, 63 %).

3.2 Quelques constantes dans l'organisation des enquêtes

Nous avons repéré quelques constantes dans les modalités de passation des enquêtes en interrogeant les choix méthodologiques des auteurs, le nombre moyen de questions et enfin les espaces de diffusion des enquêtes et de leurs résultats.

Identifier le nombre de répondants à chaque questionnaire constitue un critère de sélection pour retenir ou écarter une enquête en fonction de sa dimension représentative ou non. Pour la

présentation des résultats cumulés, nous avons choisi une échelle variable en série (100, 500, 5000) pour visualiser le nombre de répondants aux questionnaires (Figure 5).

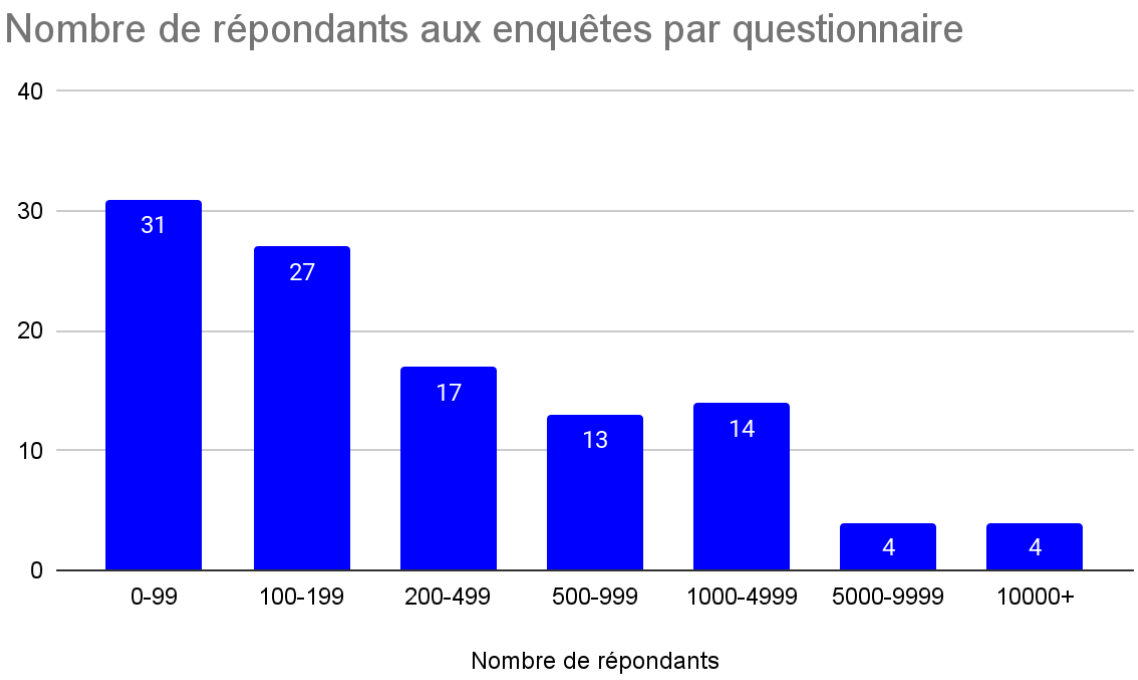

Figure 5 : Nombre de répondants par questionnaire

Tout d'abord, 4 enquêtes n'indiquent pas le nombre de répondants aux questionnaires et ne sont donc pas comptabilisées. La majorité des enquêtes (58) ont entre un nombre de répondants entre 1 et 200. En effet, la médiane s'établit à 210,5 répondants alors que la moyenne du nombre de répondants est de 1427. Pour rappel, la médiane constitue le point milieu d'un jeu de données, de sorte que 50 % des unités ont une valeur inférieure ou égale à la médiane et 50 % des unités ont une valeur supérieure ou égale. 22 enquêtes ont plus de 1000 répondants.

Quelles sont les méthodes d'enquête de collecte des données ?

Les informations sur la méthode principale de récolte et de traitement de données mise en œuvre lors des enquêtes, nous permettent de distinguer les enquêtes réalisées via des questionnaires auto-administrés, des entretiens qualitatifs avec grilles d'entretien. Notre objectif étant d'analyser les types de questions présentes dans les enquêtes portant sur les pratiques IST des chercheurs, nous avons prioritairement intégré dans notre corpus des enquêtes fondées sur des questionnaires disponibles (Figure 6). Les entretiens ont été retenus lorsque nous avions accès au guide d'entretien ce qui rend possible l'identification des questions posées.

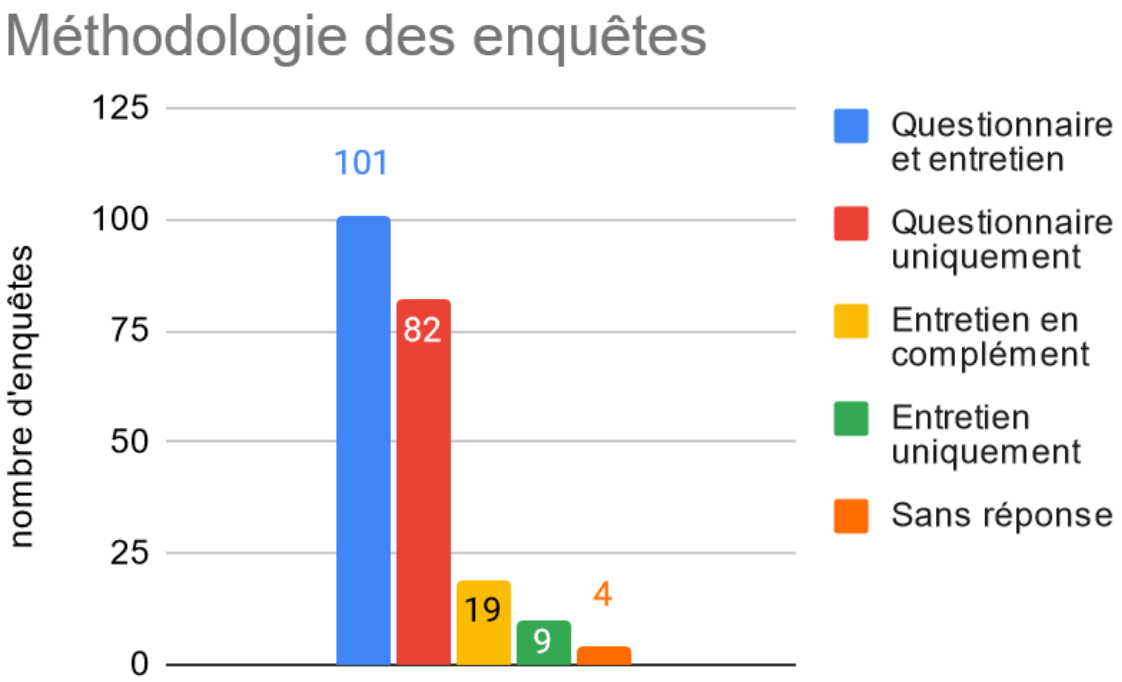

Figure 6 : Méthodologie principale de l'enquête

101 enquêtes mixent une démarche d'investigation par questionnaire avec une série d'entretiens. De fait, dans notre corpus seulement 9 enquêtes sont menées uniquement à partir d'entretiens qualitatifs. 19 entretiens complètent des questionnaires. Pour 4 enquêtes la méthodologie n'est pas évoquée, nous avons accès uniquement aux résultats.

Combien de questions sont posées par questionnaires ?

Nous souhaitions connaître le nombre moyen de questions pour ce type d'enquête. Quelle est la pratique en ce domaine des institutions qui se lancent dans cette démarche ? Dans l'objectif d'obtenir une visualisation globale du nombre de questions par questionnaire, nous avons réparti le nombre de questions par séries de 5 ou 10 (Figure 7).

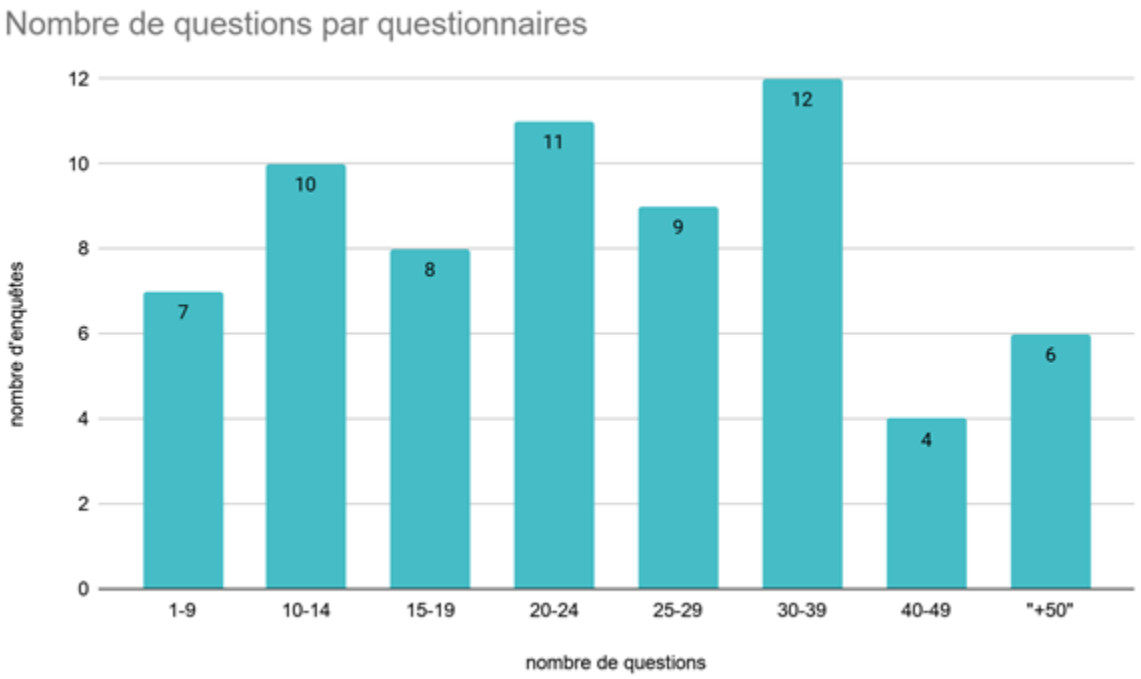

Figure 7 : Nombre de questions par questionnaires

Nous avons volontairement choisi d'affiner en séries de 5 les questionnaires comprenant entre 10 et 30 questions qui sont les plus représentés dans le corpus afin d'éviter d'avoir deux colonnes centrales trop importantes. La moyenne du nombre de questions est de 28,41 questions par questionnaires, et la médiane se situe à 24 questions.

Où sont publiées et déposées en archivage les enquêtes et leurs résultats associés quand ils sont disponibles ?

Nous avons qualifié sous le vocable « type de dépôt » l'espace de publication où nous avons recueilli des éléments d'information sur l'enquête et quand ils étaient disponibles les questionnaires originaux.

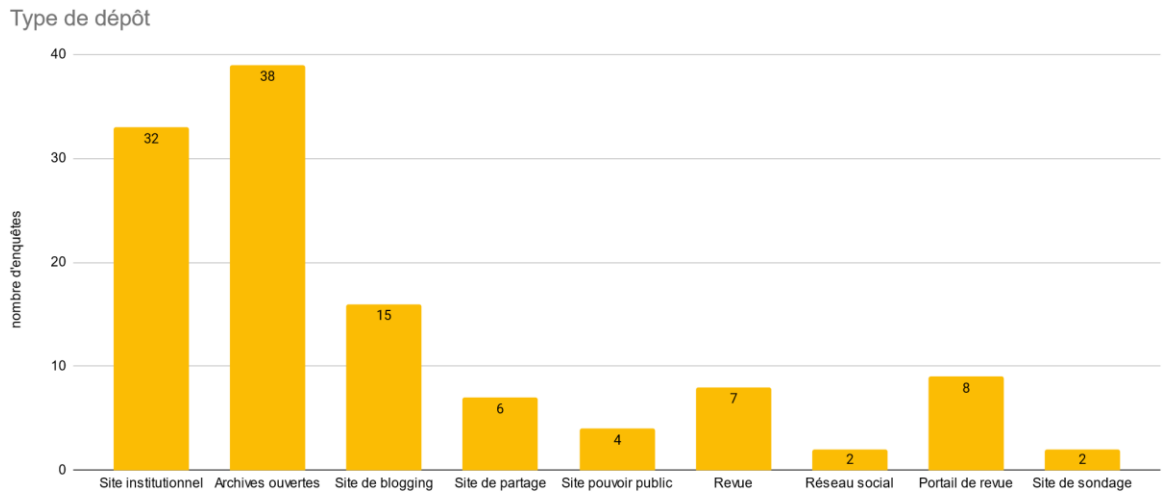

Figure 8 : Types de dépôts des enquêtes

La catégorie « site institutionnel » regroupe principalement les établissements qui présente l'étude sur leur site public (32 items). Le plus souvent les enquêtes sont disponibles via des dépôts institutionnels dans les archives ouvertes d'un établissement identifié (38 items). Les chercheurs choisissent également de donner accès aux enquêtes par l'intermédiaire de blogs (15 items), en particulier sur la plateforme de blogging scientifique en SHS, Hypothèses. La catégorie « sites de partage » se rapporte aux plateformes web qui ont vocation à partager des publications (Slideshare, Sciencesconf…). La catégorie « revue » concerne les sites de revues scientifiques, qui ne sont pas intégrés dans un portail de revue (revue d'association scientifique, revue étrangère …). La catégorie « réseau social » renvoie spécifiquement aux enquêtes trouvées sur le réseau social académique ResearchGate.

3.3 Analyse thématique

Un de notre objectif principal dans cette étude est de déterminer les questionnements présents dans l'ensemble des enquêtes pour appréhender les pratiques IST des chercheurs dans leur complexité.

Sur quel sujet portent les questions posées dans les enquêtes ?

Pour ce faire, nous avons identifié les grandes thématiques récurrentes des questionnaires. Ceux-ci sont structurés par grandes sections, ce qui permet de regrouper des ensembles de questions. La caractérisation des thématiques des questions s'est appuyée sur l'analyse des différentes questions posées aux enquêtés. Lorsque la thématique définie est représentée dans une enquête, elle est codée 1. Les résultats sont exposés par ordre décroissant en fonction du nombre d'enquêtes qui présentent les différents thèmes dans les questions posées.

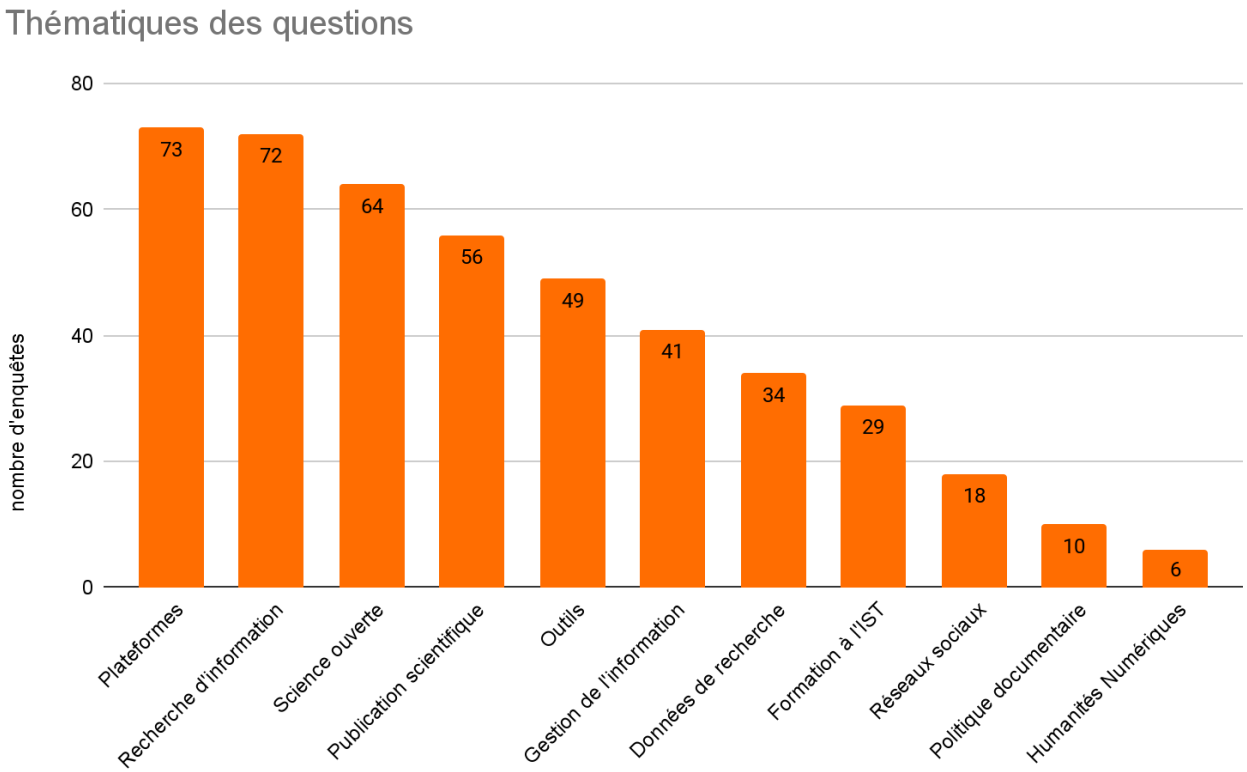

Figure 9 : Thématique des questions

Nous décrivons brièvement les différentes thématiques identifiées dans les questions par ordre décroissant de présence dans le corpus. Pour illustrer, ainsi 73 questionnaires proposent une ou plusieurs questions sur le sujet des « plateformes ».

- « plateformes » : sites web offrant une fonction particulière pour la recherche (HAL, Moodle…),
- « recherche d'information » : processus de recherche d'information, ce sujet apparaît fréquemment dans les enquêtes,
- « Science ouverte » : mouvement dont l'objectif est de rendre universellement accessibles les résultats de la recherche scientifique (publications et données de recherche, notamment),
- « publication scientifique » : questions qui cherchent à comprendre les logiques de publication des chercheurs,
- « outils » : logiciels utilisés dans les travaux de recherche des chercheurs.
- « gestion d'information » : politique de sauvegarde et de mise à disposition en ligne des données,
- « données de recherche » : pratiques de production ou de modification de données scientifiques,
- « formation à l'IST » : moyens de mieux diffuser la connaissance en Information Scientifique et Technique,

- « réseaux sociaux » : réseaux sociaux de la recherche et également les réseaux sociaux grand public dans les logiques de diffusion du savoir.
- « politique documentaire » : questions qui investiguent sur les logiques d'acquisition de documentation pour la recherche, que ce soit des ouvrages ou bien des abonnements à des bouquets en ligne.
- « humanités numériques » : rapport des chercheurs en SHS avec l'outil informatique, et les nouveaux résultats que cela produit pour la recherche.

Sur les 59 enquêtes pour lesquelles le questionnaire était disponible, une analyse complémentaire du nombre de questions par thématique a été menée. En effet, dans notre corpus, outre les enquêtes menées uniquement à partir d'entretien (3), nous avons choisi d'intégrer les présentations de résultats d'enquête sous forme de rapports, mémoires ou articles. La présentation des résultats permet d'identifier le thème des questions posées mais reprend rarement en intégralité les questions posées. Il est alors difficile de comptabiliser le nombre de questions par sujets.

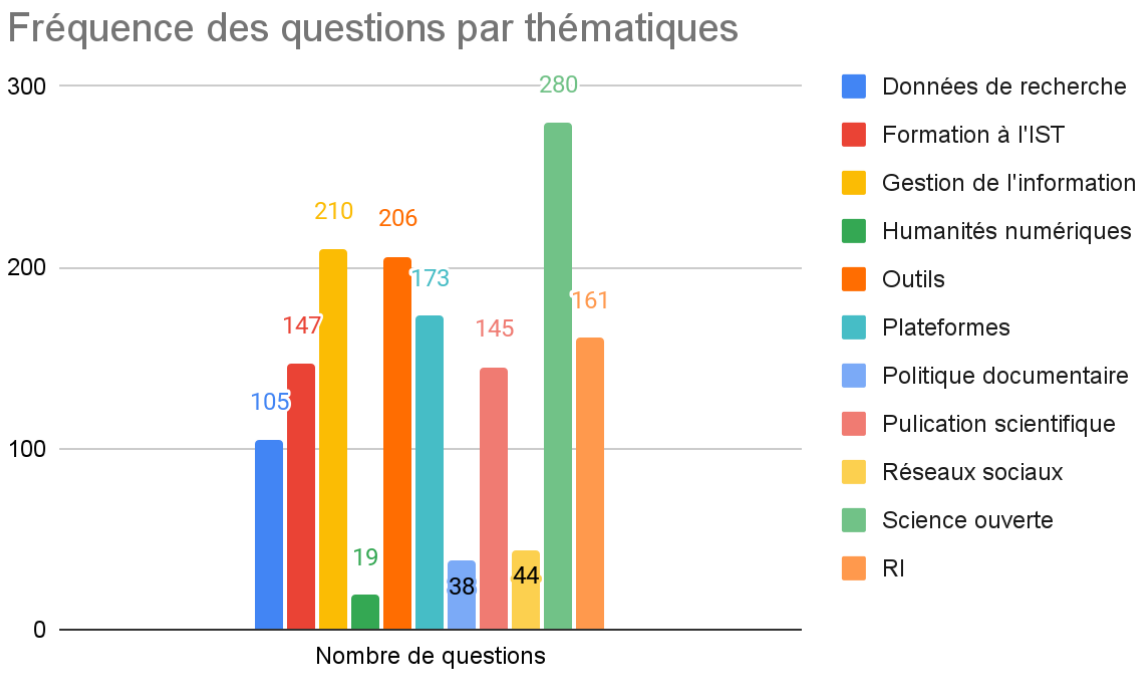

Figure 10 : Fréquence des questions les plus posées selon la thématique

Les questions les plus nombreuses portent sur l'identification du profil des interrogés. La 2° série de questions la plus posée concerne la science ouverte, puis par ordre décroissant la gestion de l'information, les outils, la recherche d'information (RI), etc…

4. Que peut-on faire avec les jeux de données ?

Notre but principal dans ce data paper est de décrire les jeux de données collectées et produites pendant le projet de recherche afin de favoriser leur réutilisation. L'analyse des enjeux propres au partage des données scientifiques met en perspective l'intérêt des métadonnées pour le partage des données de recherche (Edwards et al., 2011). Cette activité indispensable reste très souvent invisibilisée dans le processus de production de la recherche scientifique (Millerand, 2012).

Dans le contexte de la science ouverte, les principes FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*) fournissent des lignes directrices pour améliorer la facilité de repérage, l'accessibilité, l'interopérabilité et la réutilisation des ressources numériques. En ce qui concerne les données du projet, le principe « *Findable* » est appliqué du fait de l'ajout de métadonnées au moment du dépôt des données sur l'entrepôt Zenodo[7]. Cet entrepôt attribue également un DOI aux jeux de données, ce qui permet de les retrouver de manière pérenne. Le principe « *accessible* » est atteint par la définition des conditions d'accès aux données, du choix de la licence *creative commons* apposée qui précise les droits de réutilisation des données. Le principe « *interoperable* » est traité par le choix de formats ouverts (rdf) et fréquemment utilisés (pdf). L'accès via des URL publiques aux différentes collections (Huma-Num, Zotero) permet un accès généralisé quel que soit le navigateur employé. La sélection de liens additionnels vers les rapports d'enquête et articles associés disponibles dans la collection Zotero complète cette interopérabilité. Le principe « *Reusable* » concerne le choix de la licence de diffusion retenue pour la réutilisation des données par d'autres chercheurs.

Plusieurs pistes de réutilisation des données produites sont envisageables pour élaborer de nouvelles analyses sur le corpus mis à disposition. Ce dernier, limité par une date butoir, peut être complété par l'ajout de nouvelles enquêtes et a également vocation à être élargi à l'international. Il est possible de traiter le corpus avec d'autres logiciels d'analyse, dans le cadre de nouveaux contextes à définir, et éventuellement d'interroger une thématique particulière des questions (exemple : la science ouverte). En outre, analyser les pratiques des chercheurs nécessite de prendre en compte celles des professionnels de l'IST qui participent au déploiement des dispositifs informationnels. Une étude spécifique pourrait être ainsi menée sur les questionnaires renseignés par les acteurs de la documentation, en lien avec les évolutions de leur fonction dans un contexte organisationnel en pleine mutation. Les transformations induites par la disponibilité des données numériques dans le domaine scientifique ont un effet en particulier sur la collaboration entre chercheurs.

De même, cette mise à disposition fait ressortir les enjeux du travail collaboratif nécessaire à la production de données partagées. Le rapport qu'entretiennent les chercheurs aux données qu'ils produisent (Bowker, 2000), la valeur qu'ils leur accordent impactent la motivation et la décision de leur partage.

[7] Les données complémentaires du projet sont disponibles à l'adresse : https://zenodo.org/record/6065229 (DOI 10.5281/zenodo.6065229)

Conclusion

Les pratiques scientifiques et informationnelles des chercheurs, au-delà du lien évident avec les dispositifs socio-techniques et les ressources documentaires, combinent des pratiques individuelles et des pratiques collectives fondées sur un ensemble de normes et de valeurs partagées. Ce constat de disparité fait que nous sommes face à un ensemble ne pouvant véritablement être saisi dans sa globalité.

Des travaux de synthèse fondés sur l'analyse de ces résultats disparates manquent actuellement pour susciter une vision globale des pratiques informationnelles des chercheurs. Face à ces constats, le travail que nous avons engagé devrait, à terme, constituer un matériau de base pour des méta-analyses. Il se trouve toutefois confronté à un certain nombre de difficultés, dont une en particulier : approcher les pratiques informationnelles des chercheurs à partir des enquêtes sur ce sujet signifie que nous nous sommes intéressés principalement aux données quantitatives. Bien que certaines enquêtes qualitatives par entretien aient été intégrées de facto à notre champ de recherche, la centration de notre étude sur les enquêtes par questionnaire correspond à l'approche méthodologique la plus fréquente dans ce domaine, ce qui laisse de côté en partie d'autres études avec des méthodes composites (observations, entretiens qualitatifs, analyse de traces d'activité…). Ainsi, la dimension de compréhension du contexte présente dans les études ethnographiques reste difficilement accessible à partir de l'analyse de données de questionnaires. Recourir au contexte (Paganelli, 2016) permet de décrire le cadre dans lequel se passent les activités et pratiques informationnelles étudiées afin de faire émerger des facteurs explicatifs. Prouxl et Rueff (2018) s'interrogent sur la validité des méthodes employées en sciences humaines et sociales pour traiter des univers numériques. Ils distinguent cinq grandes approches : les méthodes conventionnelles ; les ethnographies virtuelles (ou ethnographies « en ligne ») ; les méthodes computationnelles tirant profit des *big data* ; les méthodes numériques (*digital methods*) ; et finalement, les méthodes numériques quali-quantitatives. Les auteurs suggèrent une hybridation des approches qualitatives et computationnelles pour appréhender les pratiques informationnelles contemporaines.

De ce fait, malgré leur richesse, comparer les résultats des enquêtes demeure une entreprise délicate. Il s'avère par conséquent indispensable de réfléchir à la meilleure manière de prendre en compte les éléments qualitatifs dont nous pouvons disposer. Pour autant, le corpus déjà constitué et ses métadonnées associées offrent des perspectives pour de nouvelles exploitations. Les tableurs de recueil de données (documents de travail) sont susceptibles d'être réexploités dans de nouvelles recherches afin de générer des statistiques plus abouties à partir d'autres croisements que le premier tri à plat présenté dans ce data paper. Enfin, les orientations actuelles du projet : base de questions permettant des comparaisons fines, analyses par Iramuteq sur les intitulés des enquêtes et sur les contenus textuels relatifs aux méthodologies employées (dans les rapports, les analyses de résultats), permettront, nous l'espérons, de compléter, corriger l'approche et d'enrichir la réflexion menée dans le cadre de ce projet. Elles généreront à terme de nouveaux jeux de données.

Bibliographie :

Geoffrey C. Bowker, "Biodiversity Datadiversity". *Social Studies of Science*, Vol. 5, n°30, 2000, p. 643-683.

Paul N. Edwards, Matthew S. Mayernik, M., Archer L. Batcheller, Geoffrey C. Bowker, Christine L. Borgman, "Science friction: Data, metadata, and collaboration". *Social Studies of Science*, Vol. 5, n° 41, 2011, p. 667-690.

Stéphane Chaudiron, Madjid Ihadjadène, « De la recherche de l'information aux pratiques informationnelles », *Etudes de communication*, n° 35, Décembre, 2010, p. 13-30.

Madjid Ihadjadène, Stéphane Chaudiron, « L'étude des dispositifs d'accès à l'information électronique : approches croisées », dans Fabrice Papy (dir.), *Problématiques émergentes dans les sciences de l'information*, Paris, Hermès, Lavoisier, 2008, p.183-207.

Annaïg Mahé, « Les pratiques informationnelles des chercheurs dans l'enseignement supérieur et la recherche : regards sur la décennie 2000-2010 », dans Ghislaine Chartron, Benoît Epron, Annaïg Mahé (dir.), *Pratiques Documentaires Numériques à l'Université*, Presses universitaires de l'Enssib, 2012, p.11-41.

Dominique Meunier, François Lambotte, Sarah Choukah, « Du bricolage au rhizome : comment rendre compte de l'hétérogénéité de la pratique de recherche scientifique en sciences sociales ? », *Questions de communication*, Vol. 23, n°1, 2014, p. 345-366.

Florence Millerand, « La science en réseau. Les gestionnaires d'information invisibles dans la production d'une base de données scientifiques », *Revue d'anthropologie des connaissances*, Vol. 6, n° 1, 2012, p. 163-190.

Céline Paganelli, « Réflexions sur la pertinence de la notion de contexte dans les études relatives aux activités informationnelles. » *Études de communication*, Vol. 46, n°1, 2016, p. 165-188.

Proulx Serge, Julien Rueff. *Actualité des méthodes de recherche en sciences sociales sur les pratiques informationnelles*. Centre d'études sur les médias, 2018.