Open, or not Open, Research Data ? Jeux d'acteurs et questions d'accès à l'IFSTTAR

Cécile Delay-Artous

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

Ecole Management et Société-Département CITS

INTD

MEMOIRE pour obtenir le

Titre professionnel "Chef de projet en ingénierie documentaire" INTD

RNCP niveau I

Présenté et soutenu par

Cécile Delay-Artous

le 10 décembre 2014

Open, or not Open, Research Data ?

Jeux d'acteurs et questions d'accès à l'IFSTTAR

Jury :
Mme Ghislaine Chartron, Professeure titulaire de la chaire d'ingénierie documentaire du CNAM, directrice de l'INTD
M. Jean-Paul Hubert, Directeur de recherche au laboratoire Dynamiques Economiques et Sociales des Transports-IFSTTAR

Promotion 44

Remerciements

En premier lieu, je tiens à remercier Ghislaine Chartron, qui a su m'encourager et m'a permis de traiter ce sujet passionnant, en restant pédagogique face à mes nombreuses interrogations. Une pensée chaleureuse, également, pour toute l'équipe de l'INTD, attentionnée et encourageante, et les auditeurs de la promotion 213/2014, formidables compagnons d'apprentissage.

Un grand merci à Alain Drouet pour son accueil, son soutien quotidien, et nos conversations qui m'ont permis d'avancer parmi les méandres de l'IFSTTAR, de la recherche française et des problématiques liées aux publications scientifiques. A toute l'équipe du service documentation de l'IFSTTAR de Marne la Vallée, merci : vous vous êtes tous rendu disponible avec gentillesse et patience. Avec une mention spéciale pour Hélène Le Bot, archiviste, qui a été, en quelque sorte, ma « veilleuse » sur les archives de la recherche, et à Malika Kahal, pour m'avoir si bien accueillie dans son bureau et donné de l'énergie en chanson.

Merci à Jean-Paul Hubert et Olivier Bonnin, pour l'opportunité qu'ils m'ont offerte de traiter des données de la recherche dans le cadre d'un beau projet qu'ils portent depuis plus de cinq ans, et pour leur disponibilité.

Un remerciement aussi pour les équipes du CDSP –et notamment Emilie Groshens, qui m'a proposé de les rencontrer- et de l'ADISP, pour m'avoir reçue si gentiment, m'avoir montré leur impressionnant travail et initiée à l'utilisation de DDI.

Je dois un grand merci également à Joël et à sa technique dite « de l'abri bus », qui m'a permis d'avancer pas à pas, en essayant de ne pas regarder la ligne d'arrivée tant que je n'en approchais pas sérieusement…

Et un immense remerciement à Solen Berhuet, pour sa très belle thèse en sociologie, qui m'a donné l'envie de retourner à l'école, et pour ses relectures et nos longues discussions, sans lesquelles j'aurai eu bien du mal à réaliser ce mémoire.

Notice

Résumé :

Participant de la révolution internet et du mouvement *Open data*, les données de la recherche font partie des préoccupations des professionnels de l'Information Scientifique et Technique. En s'appuyant sur le terrain constitué par un stage effectué durant trois mois au sein d'un établissement public de recherche, l'IFSTTAR, ce mémoire propose d'observer quelques-unes des nombreuses questions que le principe enthousiasmant de l'*Open access* soulève, en axant les recherches sur l'ouverture de données des Sciences Humaines et Sociales et la problématique de l'accès, plus ou moins large, contenue dans la notion d'ouverture.
Afin de comprendre quel type de description nécessite l'accessibilité des *Research data* l'auteure interroge les définitions de ces données de la recherche et des données dites brutes. La localisation conditionnant l'accès, la seconde partie porte sur les lieux où se trouvent ces données et dresse pour la France une cartographie de cet espace.
Pour éclairer ces réflexions à la lumière d'un cas pratique, la troisième partie décrit quelques aspects d'un projet de partage et d'ouverture de données en cours à l'IFSTTAR. Cette partie est également l'occasion d'une interrogation sur le type d'accès et d'ouverture possible et souhaitable pour des données de recherche.

Descripteurs : Donnée ouverte ; Donnée de la recherche ; Information scientifique et technique ; Recherche scientifique ; SHS ; IFSTTAR ; Accès libre ; Accès contrôlé

Table des matières

Liste des tableaux et figures

Les termes suivis d'un astérisque représentent des entrées de la cartographie située en annexe.

Introduction

Lorsque l'on évoque les données de la recherche –ou données de recherche ou encore les *Research data* en anglais- au cours de discussions informelles, y compris au sein de la promotion 2013/2014 de l'INTD ou au restaurant d'entreprise de l'Institut français des sciences et technologies des transports, de l'aménagement et des réseaux (IFSTTAR), on se retrouve généralement confronté à des interlocuteurs surpris de l'emploi de l'expression, et interrogateurs quant à sa signification précise.

Pourtant, dans la communauté de la recherche en France et parmi les professionnels de l'information scientifique et technique (IST), le sujet semble crucial tant il est régulièrement au centre de publications et de colloques. L'année 2014 a pu être qualifiée d' « année des données de la recherche » [1]au vu du nombre impressionnant de documents, revues spécialisées et congrès qui leur ont été consacrés. A l'été 2014, les documents en français introduisant à la gestion de ces données fleurissent[2].

Cette apparente contradiction trouve sa source dans les caractéristiques de son objet. Participant de la révolution internet et du mouvement *Open data*, les données de la recherche ne peuvent qu'être au cœur des préoccupations professionnelles des chercheurs et des professionnels de l'IST. Mais comme elles sont également constituées d'objets si divers[3] qu'elles en paraissent insaisissables, elles sont presque inévitablement sujet d'étonnement, et de multiples tentatives de définition. Ce qui explique que les chercheurs en fabriquent, compulsent, et utilisent... sans le savoir, ou du moins sans explicitement se référer aux « données de la recherche ». En effet, sans une définition univoque et qui fasse consensus, comment communiquer ?

---

[1] GAILLARD, Rémi. « De l'Open data à l'Open research data : quelle(s) politique(s) pour les données de recherche ? », p.11. Mémoire, Enssib, 2014. [1, GAILLARD]<http://www.enssib.fr/bibliotheque-numerique/notices/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche> [En ligne] Consulté le 20/04/2014.[1]

[2] Notons par exemple le dossier suivant : CIRAD. « S'initier en ligne aux données de la recherche et à leur gestion ». *Coopérer en information scientifique et technique*, 21 juillet 2014. <http://coop-ist.cirad.fr/gestion-de-l-information/gestion-des-donnees-de-la-recherche/s-initier-en-ligne-aux-donnees-de-la-recherche-et-a-leur-gestion/1-familiarisez-vous-avec-le-concept-de-donnees-de-la-recherche> [En ligne] Consulté le 01/09/2014. Ou encore celui-ci : CONTAT, Odile. « Gestion des données de la recherche : quelques pistes pour démarrer, comprendre et se former ». *IST SHS correspondants*, 18 août 2014. <http://corist-shs.cnrs.fr/gestion_donnees_recehrche_guideetformation> [En ligne] Consulté le 1/09/2014.
Ces deux dossiers renvoient vers de nombreux documents, parfois en anglais, pédagogiques et introductifs, sur la nature des données de recherche comme sur leur gestion.

[3] FAYET, Sylvie. « "Données" de la recherche, les mal-nommées ». *URFIST Info*, 15 novembre 2013. <http://urfistinfo.hypotheses.org/2581> [En ligne] Consulté le 4/06/2014.[2]

Dans un contexte international de mouvements comme ceux de l'*Open data* et de l'*Open access*, liés au développement de l'Internet, les données ou *data* paraissent indissociables de l'idée d'ouverture, de partage. Pourtant, les données existaient avant internet…tout comme le partage au sein de la communauté scientifique pré-éxistait à l'informatique.

Au sein des sciences de l'information et de la documentation, et plus particulièrement en Information scientifique et technique, les professionnels ont une grande habitude de la gestion et de la valorisation des publications scientifiques. Toute la question est de savoir s'il leur sera donné de faire de même avec les données de la recherche, et s'ils en auront les compétences et les moyens.

Le contexte national est, en France et en SHS, marqué par un retard, comme le notait déjà Roxanne Silberman[4], auteure en 1999 d'un rapport qui marque le début de l'intérêt national pour les données scientifiques en SHS sur le territoire français :

> « *La France est en retard sur ce point de plus de vingt ans sur plusieurs de ses voisins européens qui ont construit, à l'image de ce qui avait été entamé aux États-Unis, de puissants Data Archives pour la recherche en sciences sociales. Elle est également absente des grandes enquêtes européennes et internationales universitaires.*»

Deux ans plus tard, alerté par ce rapport, le Ministère crée le Comité de concertation pour les données en sciences humaines et sociales (CCDSHS*). Mais aujourd'hui, soit treize ans plus tard, tandis que les britanniques en sont à évaluer (quantitativement et qualitativement) l'impact[5] de *l'Open Research data*, le site de Sciences Po conclu encore au retard français, au moins en SHS[6].

En nous appuyant sur le terrain constitué par le stage effectué durant trois mois au sein d'un établissement public de recherche, l'IFSTTAR, et sur des lectures, nous allons observer quelques-unes des nombreuses questions que le principe enthousiasmant de l'*Open access* soulève, en axant nos recherches sur l'ouverture de données de SHS et la problématique de l'accès, plus ou moins large, contenue dans la notion d'ouverture.

Afin de comprendre si l'accessibilité des *Research data* nécessite une description riche et fournie, et des modèles de métadonnées internationaux, ou si elle peut se contenter d'une mise en ligne des données elles-mêmes, accompagnée d'une sorte de notice minimale, nous commencerons par chercher une définition de ces données de la recherche et de ce que l'on nomme les données brutes.

---

[4] SILBERMAN Roxanne. *Rapport Silberman : Les sciences sociales et leurs données*, Ministère de l'Education Nationale, de la Recherche et de la Technologie, 1999, p.3. [3] Plus récemment, sur le site de Sciences Po, au sein des pages consacrées à l'équipement DIME-SHS : « La recherche française souffre d'un grand retard en matière d'équipements dédiés à la production, la diffusion et la réexploitation des données pour les sciences humaines et sociales » <http://www.sciencespo.fr/dime-shs/content/aux-origines> [En ligne] Consulté le 25/09/2014.

[5] De nombreux articles de recherche sur le sujet précis de l'évaluation paraissent au Royaume Uni, commandité et/ou réalisé entre autre par le Jisc, organisme public britannique. Par exemple, cette synthèse de trois études évaluant des *Research data centres* britanniques : BEAGRIE Neil, HOUGHTON John. The Value and Impact of Data Sharing and Curation. Jisc. 2014.

[6] <http://www.reseau-quetelet.cnrs.fr/spip/article.php3?id_article=220> [En ligne] Consulté le 05/08/2014.

La localisation conditionnant l'accès, nous nous interrogerons ensuite sur le - ou les - lieux où il est possible, en France, de trouver ou de déposer ces données de recherche. Afin de contribuer à une simplification de l'orientation dans le monde des *Research data*, nous essayerons de dresser une cartographie de cet espace.
Enfin, pour éclairer ces réflexions à la lumière d'un cas pratique, nous présenteront quelques aspects d'un projet de partage et d'ouverture de données actuellement en cours à l'IFSTTAR, et auquel nous avons participé. Cette troisième partie sera également le lieu d'une interrogation sur le type d'accès et d'ouverture possible et souhaitable pour des données de recherche. Est-ce nécessairement un *Open access* total et sans restriction ? Ou bien peut-on partager des données de façon contrôlée sans perdre les bénéfices de l'ouverture ?

PREMIERE PARTIE

I Que sont les données de la recherche ?

1.1 A la recherche d'une définition consensuelle et opérante

Mais que sont donc ces fameuses données de la recherche[7] ? Concrètement, Sylvie Fayet liste quelques exemples qui expriment bien la diversité des matériaux désignés en fonction des disciplines scientifiques :

> *« […]quand on évoque « les données de la recherche », on désigne des chiffres, relevés, mesures, résultats d'expérience, réponses à des enquêtes, statistiques, comptages, et autres données quantitatives sur la base desquels va s'élaborer une hypothèse, et/ou qui serviront à infirmer ou valider cette hypothèse[…]. »* [8]

Cet aspect quantitatif pourrait constituer un appui pour construire une définition…Mais cela nous couperait d'une part très importante de la multiplicité de ce que recouvrent les données de recherche, et notamment en SHS :

> « […] les « données » d'un médiéviste sont des sources archivistiques, archéologiques, épigraphiques, iconographiques, littéraires ; les « données » d'un géologue rassemblent des coupes et observations de terrain consignées sur un carnet, des résultats de carottage, des analyses d'échantillons, des données sismographiques… »[9].

Le point commun, selon certaines définitions, se trouverait dans ce que les données ne sont pas, à savoir : des publications, des résultats de recherche, dont elles ne sont que la source.

Les universités américaines, à la suite d'une circulaire fédérale de 1993, définissent ainsi l'expression :

> « La donnée de recherche est définie comme l'enregistrement factuel couramment considéré dans la communauté scientifique comme nécessaire à la validation des résultats de la recherche, à l'exclusion des documents suivants : analyses préliminaires, projets d'articles scientifiques, plans pour de futures recherches, évaluations par les pairs, ou communications avec les collègues »[10]

---

[7] Au sein des professions archivistiques et documentaires et en France, le site d'information sur les collectivités territoriales Territorial.fr relève déjà trois définitions différentes, en partant de la notion d'archives de la recherche. Il serait intéressant de se pencher sur les utilisations des deux termes, archives d'une part, et données d'autre part, parmi la communauté de la recherche. « Les professions archivistiques et documentaires donnent des définitions différentes des archives imprimées et numériques de la recherche. » Territorial.fr. Interview de Charlotte MADAY, mis en ligne le 28/10/2014. <http://www.territorial.fr/PAR_TPL_IDENTIFIANT/69669/TPL_CODE/TPL_ACTURES_FICHE/PAG_TITLE/Archives+de+la+recherche+%281%29++%3A+une+probl%E9matique+qui+monte+en+puissance+++/302-actu.htm> [En ligne] Consulté le 15/11/2014.

[8] FAYET, Sylvie, Op. cit.

[9] Ibid.

[10] « *Research data* is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses,

Dans la même ligne, l'organisation de coopération et de développement économiques (OCDE ou OECD en anglais), dans un guide publié en 2007, propose une définition excluant les carnets de laboratoires et les communications entre pairs :

> « […] les « données de la recherche » sont définies comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche. […] Ce terme ne s'applique pas aux éléments suivants : carnets de laboratoire, analyses préliminaires et projets de documents scientifiques, programmes de travaux futurs, examens par les pairs, communications personnelles avec des collègues et objets matériels (par exemple, les échantillons de laboratoire, les souches bactériennes et les animaux de laboratoire tels que les souris). »[11]

Mais d'une part, ces éléments que l'on pourrait classer du côté des archives, font pourtant partie des données de recherche pour certaines universités britanniques et australiennes[12].

Et d'autre part, avec l'essor du mouvement de *l'Open access*, de nombreuses données peuvent se retrouver « publiées » et commentées sur Internet, au sein de réseaux sociaux de chercheurs ou sur leurs blogs, ou encore dans des publications sans la validation d'un comité de lecture.

Enfin, la notion même de résultat est délicate : entre des résultats d'essai et les résultats de l'analyse de cet essai, quel objet est le plus légitimement nommé « résultat » ?

Du point de vue de la gestion de ces données, et dans le cadre d'initiatives de partage et d'ouverture, ces désaccords basés sur des typologies sont peu opérants. Les définitions récentes élaborées par des professionnels de l'information scientifique avec un objectif utilitaire (la réutilisation des données, par exemple), permettent de surmonter (ou du moins de reporter à plus tard, lorsque l'on voudra décrire des données) quelques difficultés.

Voici par exemple la définition très large que l'on trouve en introduction au dossier publié par le site CoopIST :

---

drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. »
Office of management and budget, 1993. Circulaire A-110 amendée le 30/09/99.
<http://www.whitehouse.gov/omb/circulars_a110#36> [En ligne] Consulté le 3/10/2014.

[11] Organisation de Coopération et de Développement Economiques (OCDE), *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics*, Paris, 2007.
<http://www.oecd.org/fr/science/sci-tech/38500823.pdf> [En ligne] Consulté le 4/07/2014.

[12] GAILLARD Rémi, op. cit., p. 16.

> « Les données de la recherche […] peuvent être définies comme l'ensemble des informations collectées, observées ou créées sous une forme numérique dans le cadre d'un projet de recherche. »[13]

Voici un périmètre défini pour partie par son format numérique ; mais ce format n'est pas obligatoirement natif, et les données peuvent avoir été recueillies sous une autre forme puis être numérisées. Il n'y a donc qu'une semi-restriction par le support. Et, du point de vue de la gestion des données, qui inclut la question de la diffusion (voir du partage avec la notion de *Data sharing,* littéralement partage de données) de ces données, cette semi-restriction est parfaitement opérationnelle. Elle s'adapte au contexte qui fait d'Internet, du Web et donc du numérique, l'outil principal des échanges d'information.

Surtout, cette description présente l'avantage d'inclure aussi bien les données d'observation, les données expérimentales, les modèles et simulations, que les données dérivées ou compilées, comme les bases de données issues de la compilation d'un ensemble de données collectées et/ou créées, les résultats de fouille de texte (*text mining*) ou de fouille de données (*data mining*). En somme, tout ce qui pourrait être réutilisé pour des projets de recherche, ou étudié pour vérifier des résultats. Mais aussi tout ce qui demande du travail, généralement invisible[14], et pourrait être valorisé. Il y manque toutefois une précision : il s'agit d'informations qui participent de l'élaboration des résultats[15], même de façon dérivée. Sans cela, des documents de communication ou d'administration qui précèdent ou suivent un projet viendraient encombrer notre périmètre. Mais exclure, comme l'OCDE ou Thierry Fournier tout ce qui ressemble à un carnet de laboratoire revient à postuler à priori qu'un type de données doit être exclu pour sa forme, ce qui n'est pas opérant dans une perspective de management des données.

La *Queensland University of Technology* propose une définition ouverte et efficace, plus subtile et complète, qui résout nos questionnements de « manager de l'info » :

> *« Les données de la recherche désignent des données sous forme de faits, observations, images, programmes informatiques, enregistrements, mesures ou expériences sur la base desquelles un argument, une hypothèse, une théorie ou tout autre produit d'une recherche s'appuie. Ces données peuvent être numériques, descriptives, visuelles ou tactiles. Elles peuvent être brutes, nettoyées ou traitées, et conservées sous tout format ou support. »*[16]

---

[13] CIRAD, 2014. Op. cit.
[14] DENIS Jérôme et GOËTA Samuel, « La fabrique des données brutes. Le travail en coulisses de l'open data ». Paris, 2013. <http://halshs.archives-ouvertes.fr/halshs-00990771> [En ligne] Consulté le 21/07/2014.
[15] FOURNIER Thierry, « Les données de la recherche : définition et enjeux. » In *Arabesques* n°73, p. 4
[16] "Research data means data in the form of facts, observations, images, computer program results, recordings, measurements or experiences on which an argument, theory, test or hypothesis, or another research output is based. Data may be numerical, descriptive, visual or tactile. It may be raw, cleaned or processed, and may be held in any format or media." Citée par *l'Australian National Data Service*, <http://ands.org.au/guides/what-is-research-data.html> [En ligne] Consulté le 03/08/2014

Parmi l'ensemble des sources d'information de la recherche se trouve les bibliographies des projets, constituée de publications. L'information scientifique et technique les connait depuis toujours et s'en occupe fort bien : c'est une partie de sa définition.

Du point de vue de l'organisation de l'information et de la documentation, on pourrait définir les données de la recherche par soustraction : ce sont toutes ces sources concourant aux résultats scientifiques, moins celles que nous connaissons le mieux.

1.2 L'incontournable question des données brutes : Raw data is (really ?) an oxymoron.

«*Raw data now*!» Nous voulons des données brutes ! Cette phrase que Tim Berners-Lee, le « père du web », fait reprendre par le public lors d'une conférence TED en 2009[17], sert de point de départ à un débat assez fondamental pour la question de la définition des données de recherche et plus généralement, des données.

En effet, une partie des acteurs des data –scientifiques[18], historiens des sciences, informaticiens, professionnels de l'information…- réserve, avec Tim Berners-Lee, le terme de « donnée » ou « data » à ce qu'ils nomment les données brutes (ou *raw data* en anglais). Il n'y aurait de « véritable » donnée que brute. Pour T. Berners-Lee, « brute » est dans ce cas synonyme de « non altérée »[19].

D'autres, au contraire, vont jusqu'à se demander si le terme de « brut » appliqué aux données a encore du sens. Aucune donnée ne pourrait être réellement brute[20].

Mais chacun s'accorde sur la difficulté à définir une frontière entre données brutes et données traitées ou dérivées.

Et si « brute » signifie « objective » ou « non travaillée », il est tentant de rejeter l'idée de « données brutes ». Car toute donnée a un contexte, et a été, à tout le moins, recueillie, et donc « orientée » et … travaillée. Pour prendre un exemple issu des sciences sociales, examinons un entretien avec un chômeur, réalisé dans le cadre d'une recherche sur les Maisons de l'Emploi par une sociologue[21] : les informations recueillies auprès de lui dépendent des questions qu'on lui pose. Il peut être interrogé sur les conséquences du chômage sur sa vie sociale, ses activités culturelles ou au

[17] BERNERS-LEE Tim. *The next web*. TED2009 conference. <http://www.ted.com/talks/tim_berners_lee_on_the_next_web> [En ligne] Consulté le 17/10/2014
[18] Voir par exemple THESSEN Anne E., PATTERSON David J., « Data Issues in the Life Sciences », *ZooKeys*, novembre 2011, n° 150, p. 15-51, p. 17, <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3234430/> [En ligne] Consulté le 17/10/2014
[19] « […]*give us the unadulterated data, we want the data. We want unadulterated data. OK, we have to ask for raw data now*. » BERNERS-LEE Tim, op. cit. Notons que *unadulterated* peut être traduit comme « non-frelaté »…
[20] GITELMAN Lisa (dir.), *Raw Data is An Oxymoron,* Cambridge, MIT Press, 2013, 182p.
[21] BERHUET, Solen. *Les chômeurs et les intermédiaires de l'emploi : Une sociologie dynamique de leurs trajectoires au sein d'une Maison de l'Emploi*. Thèse de sociologie, dir. M. Lallement et P. Nivolle, Cnam, 2013.

contraire sur ses démarches de recherche d'emploi. Il y a donc en ce sens une « construction » des données.

Dans la définition même des données par les tenant du « brut » comme objectivité, le processus de construction, c'est-à-dire le travail, des données, semble oublié. Et dans cette optique, le quantitatif a la part belle : dans une confusion entre « donnée », « information » et « connaissance »[22], T. Berners-Lee nous conte l'histoire d'un monde rendu meilleur grâce à plus de données-connaissances.

Ce raisonnement ressemble à une fiction en forme d'équation. Si « plus » signifie « mieux », et si « donnée » signifie « connaissance », *Big data* devient synonyme de « davantage de connaissances ». Nous serions donc devant une masse de connaissances « brutes » et pures, un trésor, du pétrole qu'il n'y aurait qu'à raffiner à l'aide d'ordinateurs : le Data Mining (fouille de données) pourrait changer le monde.

Mais cette « injonction à l'ouverture opère une certaine mise en invisibilité. »[23], celle des conditions de production des données comme de leurs contextes d'utilisations premiers. Il ne suffit pas de mettre à disposition des données qui seraient « déjà là ». Leur ouverture n'est pas immédiate. Elle demande un long travail d'identification, d'extraction et de « brutification », travail qui permettra que les données soient trouvées, comprises et éventuellement réutilisées. La « brutification » peut être vue comme le processus qui va les rendre interopérables.

Pour filer le jeu de mot de Sylvie Fayet qui titre un article sur les « « Données » de la recherche, les mal nommées », on pourrait dire que les données brutes ne sont pas données. Avec Jérôme Denis et Samuel Goëta, nous poursuivrons l'idée que ces données « brutifiées » peuvent faire l'objet d'un don à la communauté. En somme, les données brutes existent bien, mais il faut les fabriquer.

1.3 Pourquoi et depuis quand parle-t-on de Research data ?

Au croisement de l'*Open Data* et de l'*Open Access*, cette notion parait encore un peu en marge des discours sur les data. Mais est-ce très étonnant, si nous nous rappelons qu'elle n'a émergé que récemment ?

1.3.1 Un contexte et un bref rappel historique : l'Open Access et l'Open Data

L'Open Access (OA) ou accès libre, fait une première apparition avec la naissance en 1991 de l'archive ouverte ArXiv, destinée aux prépublications pour la communauté des physiciens théoriciens[24]. Dans les années 2000, le mouvement passe des chercheurs et

[22] JEANNERET Yves, *Y a-t-il -vraiment- des technologies de l'information*, Presses Univ. Septentrion, 2000, p.58.
[23] DENIS Jérôme et GOËTA Samuel, op. cit.
[24] CHARTRON Ghislaine, « Open access et SHS : Controverses », *Revue européenne des sciences sociales* 1/ 2014 (52-1), p. 37-63 <www.cairn.info/revue-europeenne-des-sciences-sociales-2014-1-page-37.htm> [En ligne] Consulté le 30/07/2014.

des bibliothèques aux déclarations internationales qui le popularisent : à Budapest puis à Berlin, les définitions et recommandations se formalisent.

Ce mouvement vise à permettre une diffusion la plus large possible de l'information scientifique, sans aucune barrière. Il s'agit de l'élargissement d'une pratique de diffusion quasi traditionnelle entre chercheurs : en s'appuyant sur l'Internet, l'accès peut, virtuellement, concerner tout le monde. L'objectif posé est de faciliter la diffusion et le développement du savoir. En somme, l'OA voudrait formaliser « l'idée de pouvoir accéder librement à toutes les connaissances disponibles »[25] :

> *« Par "accès libre" à [la littérature de recherche validée par les pairs], nous entendons sa mise à disposition gratuite sur l'Internet public, permettant à tout un chacun de lire, télécharger, copier, transmettre, imprimer, chercher ou créer un lien vers le texte intégral de ces articles, les analyser automatiquement pour les indexer, s'en servir comme données pour un logiciel, ou s'en servir à toute autre fin légale, sans barrière financière, légale ou technique autre que celles indissociables de l'accès et de l'utilisation d'Internet. La seule contrainte sur la reproduction et la distribution et le seul rôle du droit d'auteur dans ce contexte devrait être de garantir aux auteurs un contrôle sur l'intégrité de leurs travaux et le droit à être correctement reconnus et cités. »*

C'est la fameuse définition de l'*Open Access* par la « *Budapest Open Access Initiative* » (BOAI, 2002), réaffirmée en 2012 (BOAI10).[26]

Cette définition de l'OA ne concerne pas -encore- les données de la recherche : il s'agit dans cette déclaration de donner accès à la littérature scientifique validée par les pairs, c'est-à-dire à ce que la communauté de la recherche nomme les publications.

Mais dès 2003, un texte dans la droite ligne de la déclaration de Budapest, mentionne explicitement les données comme devant relever de l'Open Access. C'est la déclaration de Berlin :

> *« Conformément à l'esprit de la Déclaration de Budapest pour l'accès ouvert[…] nous avons élaboré la déclaration de Berlin pour promouvoir un Internet qui soit un instrument fonctionnel au service d'une base de connaissance globale et de la pensée humaine, et de définir des mesures qui sont à envisager par les responsables politiques en charge de la science, les institutions de recherche, les agences de financement, les bibliothèques, les archives et les musées.[…] Les contributions au libre accès se composent de résultats originaux de recherches scientifiques, de données brutes et de métadonnées, de documents sources, de représentations numériques de documents picturaux et graphiques, de documents scientifiques multimédia. »*

Voici donc notre contexte : l'OA concerne tout type d'accès par Internet, pour tout type de réutilisation, de tout type de production scientifique.

---

[25] Ibid.

[26] Budapest Open Access Initiative. *Dix ans après l'Initiative de Budapest : ce sera le libre accès par défaut*. Budapest, 2012. <http://www.budapestopenaccessinitiative.org/boai-10-translations/french> [En ligne] Consulté le 17/09/2014.

Plus récemment, les acteurs politiques, notamment américains et européens, ont repris ces définitions et transformé ces recommandations en injonctions[27] aux orientations libérales, ce qui les éloigne sérieusement de l'esprit des pionniers de l'OA[28].

Ces orientations sont également visibles à travers les politiques d'*Open Data*[29] mises en place par les gouvernements. Prenons l'exemple français de la plateforme de données ouvertes data.gouv.fr, mise en place par la Mission Etalab pour le gouvernement. Sur la page d'accueil, ce sont les réutilisations des jeux de données qui sont mises en avant, sous forme de bandeau d'images défilant, au centre de l'écran. Outre la notion de transparence nécessairement liée à celle d'ouverture, c'est l'idée d'un retour sur l'investissement public qu'on retrouve.

Au vu de ces définitions, la question se pose de savoir si l'*Open Research data* est une composante de l'*Open Data* ou de l'*Open access*, ou si l'on peut l'imaginer comme une intersection de ces deux mouvements.

1.3.2 Des points de vue, des discours et des actions

Des acteurs politiques et économiques en faveur du Libre accès

> « *L'information scientifique est un bien commun, qui doit être disponible pour tous. Elle a un coût : il nous revient que ce coût ne soit pas une barrière, afin que la notion de bien commun de l'information scientifique devienne une réalité. C'est aussi un fondement de notre compétitivité par la qualité, […] qu'il nous faut valoriser au niveau national, européen, international, dans des modèles économiques innovants et diversifiés.*» [30]

Ainsi Geneviève Fioraso, secrétaire d'Etat à l'Enseignement Supérieur et à la Recherche, concluait elle son discours lors des cinquièmes journées *Open Access,* le 24 janvier 2013. Le positionnement du gouvernement français est donc clairement dans la lignée de celui de l'OCDE[31] et des recommandations de la Commission Européenne[32]. L'ouverture doit devenir la règle, la culture commune du monde de la recherche, pour les publications comme pour les données. Les justifications de ces acteurs politiques

---

[27] On en voit un exemple européen avec le programme cadre Horizon 2020, qui comporte un volet rendant obligatoire la diffusion en OA d'une partie de la production scientifique, y compris les données.
[28] CHARTRON Ghislaine, op.cit.
[29] L'*Open Data* peut concerner tout type de données numériques. Mais la notion de politique *Open Data* renvoie en général aux données publiques, à caractère administratif ou de gestion. Pour aller plus loin, CHIGNARD Simon. Open data: comprendre l'ouverture des données publiques. Paris : Éditions fyp, 2012. 191 p.
[30] Discours de la secrétaire d'Etat Geneviève Fioraso lors des 5 e journées Open Access. Disponible en ligne sur le site du MESR < http://www.enseignementsup-recherche.gouv.fr/cid66992/discours-de-genevieve-fioraso-lors-des-5e-journees-open-access.html> Consulté le 02/11/2014.
[31] Ce point de vue s'exprime par exemple dans les *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics.* Paris : OCDE, 2007. Op. cit.
[32] Un résumé de ces recommandations se trouve sur l'un des sites de l'Inist, dédié au Libre accès : <http://openaccess.inist.fr/?Resume-en-francais-du-rapport> [En ligne] Consulté le 02/11/2014.

relèvent d'une injonction normative autour de la notion de bien commun : il s'agit d'un principe de partage, de redistribution du bien commun que doit représenter la connaissance scientifique. La plupart des acteurs politiques et économiques rappellent également que le travail des chercheurs repose sur ce partage, cet échange permanent et les confrontations qui peuvent en découler. Mais le registre le plus commenté par ce type d'acteur reste celui de l'innovation et de l'économie. Ainsi, l'OCDE note dans un chapitre intitulé « Augmenter le rendement des investissements publics dans la recherche scientifique » :

> « *Un accès efficace aux données de la recherche, utilisé de façon responsable et efficiente, est nécessaire pour tirer pleinement parti des nouvelles possibilités et retombées offertes par les TIC. L'accessibilité aux données de la recherche est devenue une importante condition pour : Une gestion avisée de l'investissement public dans l'information factuelle ; Le développement de chaînes d'innovation à forte valeur ; L'accroissement de la valeur procurée par la coopération internationale.* »[33]

Le MESR traduit donc, logiquement, cette volonté politique forte, notamment par le biais d'actions portées par le CNRS, entité chargée de l'exécution de sa politique scientifique. Par exemple, c'est le CNRS qui porte, à la demande du ministère, le site internet de veille sur les données de la recherche à destination des chercheurs et des professionnels de l'IST : donneesdelarecherche.fr. C'est également cet organisme qui porte –souvent en partenariat avec d'autres- les TGIR, la Bibliothèque Scientifique Numérique (BSN), le CCDSHS, l'archive ouverte française HAL… et qui met à disposition au travers de ses nombreux sites internet, l'information tant grand public que technique, sur le libre accès aux productions scientifiques en France. Tous ces sites reflètent le volontarisme gouvernemental, et mettent l'accent sur des aides méthodologiques, des actualités, des réalisations. Bien que cela ne soit en général pas précisé, le champ lexical et la plupart des exemples utilisés renvoient plutôt aux sciences dites « dures ».

Des acteurs de l'IST aux points de vue encore en construction

Le débat sur l'*Open access* tournait jusqu'ici beaucoup autour de la question des publications scientifiques : fallait-il privilégier la voie dorée[34] (ou auteur-payeur, *Golden road* en anglais) ou la voie verte[35] (ou auto-archivage, *Green road* en anglais) ? Quelles durées d'embargo devaient être négociées avec les éditeurs, et fallait il les adapter en fonction des disciplines ? Le gouvernement français ayant opté pour une

[33] OCDE, op. cit.
[34] Selon le glossaire de l'Inist : « la voie dorée s'applique à la publication d'articles dans des revues en libre accès, quel que soit leur mode de financement. » <http://openaccess.inist.fr/?+-Voie-doree-+>[En ligne] Consulté le 17/10/2014.
[35] Toujours selon l'Inist : « La voie verte qualifie l'auto-archivage par les chercheurs ou l'archivage par une tierce personne des articles dans des archives ouvertes. » <http://openaccess.inist.fr/?+-Voie-verte-+>[En ligne] Consulté le 17/10/2014.

articulation des deux voies (et d'une voie dite de Platine ou *Platinium road*) et des durées d'embargo différenciées en fonction des disciplines, ce sont maintenant les questions concernant les données de la recherche qui font l'actualité. Mais le débat est encore neuf, et les acteurs n'ont pas, pour la plupart, arrêté de position nette. Les questions techniques sont au cœur des interrogations des professionnels de l'IST, qui se demandent comment ils peuvent et doivent participer à ce nouveau champ de l'information scientifique. Les publications qui s'adressent à ces acteurs sont, pour le moment, principalement orientées vers l'analyse d'exemples de réalisations *Open research data*. Avec, comme question plus ou moins clairement posée, celle de la place d'un métier vis-à-vis de l'objet *Research data* : qui va donc s'en occuper ? Les bibliothécaires, les documentalistes, les archivistes ? Comment s'imposer sur un terrain occupé « naturellement » par les informaticiens (puisque ces *data* sont ou deviennent numériques) et les chercheurs (qui les produisent ou les agrègent en fonction de leurs disciplines) ? Tous les groupes de l'IST ne sont pas sur la même position ; beaucoup cherchent encore à en définir une.

Voici une brève revue des points de vue des différents professionnels de l'IST que nous avons rencontrés au fil de nos lectures, et que nous tâchons de résumer. Nous avons essayé d'en dégager des positions sur les données de la recherche ou, à défaut, sur l'*Open access*. La liste est bien sûr non exhaustive ; elle représente les groupes qui sont apparus les plus audibles lors de nos recherches.

- Les professionnels des BU et de la documentation scientifique publique, avec le numéro 73 de la revue de l'ABES « Arabesque »[36] dont le dossier est consacré aux données de la recherche (Roger Genet, Directeur général pour la recherche et l'innovation au MESR signe d'ailleurs l'un des articles), et le mémoire de R. Gaillard (commandé par Couperin). C'est en quelque sorte le point de vue officiel : les options gouvernementales et leur traduction pour et par l'IST. La diffusion des données de la recherche y est décrite comme une révolution à ne pas manquer[37] :

  > « *L'exposition des données de la recherche va bouleverser, plus profondément encore que le basculement de la documentation du papier vers l'électronique, les métiers des bibliothécaires et documentalistes. […]Que ce soit le G8, les financeurs de la recherche, la communauté scientifique internationale, tous s'accordent sur l'indispensable ouverture des données de la recherche. Elle s'inscrit dans un contexte plus large d'ouverture des données publiques en générale.* »

  Cet extrait de l'éditorial de Jérôme Kalfon, directeur de l'ABES, exprime clairement un engagement en faveur de l'ouverture des données de recherche, et place celle-ci au sein d'une politique plus globale d'*Open data*. Les *Research data* sont donc des données publiques. D'ailleurs, « Le principe de la libre circulation des idées est la règle, la propriété intellectuelle, […] l'exception. »

[36] ABES. « Semer, essaimer : La valorisation des données de la recherche ». *Arabesque*, n° 73 (janvier 2014). <http://www.abes.fr/Arabesques/Arabesques-n-73> [En ligne] Consulté le 03/06/2014.
[37] Ou encore comme un train à ne pas laisser passer. GAILLARD Rémi, op. cit. p.14.

L'ouverture sera facilitée par le fait « qu'il n'y a pas de modèle économique préexistant » pour les données, contrairement à ce qui a pu se passer pour les publications scientifiques. Et la justification est claire : « Sans ouverture, pas de possibilité d'assemblage, de réutilisation, de construction de nouvelles connaissances. »

On retrouve des nuances au sein de l'expression de cette position, jusque dans les différents articles du même dossier (tous les auteurs ne font pas référence à la même définition, par exemple). Mais, tout en reconnaissant sans les minimiser les difficultés, l'unanimité se fait quant à la nécessité d'une ouverture, et sur l'idée qu'il ne faut pas rater cette opportunité.

Notons que sur la problématique du modèle économique, cette position est à l'inverse de celle d'éditeurs scientifiques et de certains chercheurs.

- Les documentalistes se sont exprimés quant à l'Open access au travers de la position de l'ADBS. En mars 2013[38] et à propos du « Libre accès aux résultats de la recherche », l'association des documentalistes met en ligne un texte qui doit expliquer sa position :

  « *L'ADBS a signé une motion proposée par Cairn, le portail des sciences humaines et sociales ; elle a également signé, via l'Interassociation archives-bibliothèque-documentation (IABD) dont elle fait partie, la pétition présentée par des responsables d'universités, d'enseignants-chercheurs, d'éditeurs et de responsables de bibliothèques alors que ce texte critique les appréhensions des éditeurs en SHS.*»

  Cette position est donc délicate puisqu'elle elle tente de concilier deux textes qui s'opposent. Bien que tout ceci ne concerne pas explicitement les données de la recherche, puisqu'il s'agit de réactions à la directive européenne concernant les publications, l'esprit de la motion des éditeurs[39] comme du texte publié dans Le Monde[40] permet de comprendre deux positions sur l'ouverture et le partage des résultats scientifiques.

  Le premier texte signé par l'ADBS est une lettre ouverte d'éditeurs de revues SHS nommée « *Open access* : le travail scientifique en sciences humaines et sociales et le débat public fragilisés par les mesures préconisées par la Commission européenne ». Comme l'indique ce titre, ces signataires craignent que la promotion de l'édition en libre accès des résultats scientifiques « ne permette le maintien d'un grand nombre de revues académiques » et « qu'au-delà des revues académiques, les autres domaines de l'édition du savoir en sciences humaines et sociales, notamment les revues de débat ou d'opinion,

[38] Texte disponible sur le site de l'ADBS <http://www.adbs.fr/libre-acces-aux-resultats-de-la-recherche-l-adbs-dit-oui-a-la-concertation-127097.htm?RH=1245421882337> [En ligne] Consulté le 03/11/2014.[20]
[39] Motion rédigée le 11 février 2013 à la Maison de la Chimie de Paris, à l'occasion d'une journée de réflexion intitulée : Les revues de sciences humaines et sociales survivront-elles aux mesures préconisées par la Commission européenne en matière d'*Open Access* ?<http://www.openaccess-shs.info/motion/> [En ligne] Consulté le 05/11/2014.
[40] Texte paru dans Le Monde daté du 15/03/2013 sous le titre Qui a peur de l'open access ? < http://www.lemonde.fr/sciences/article/2013/03/15/qui-a-peur-de-l-open-acces_1848930_1650684.html> [En ligne] Consulté le 05/11/2014.

pourraient être, eux aussi, menacés dans la mesure où leurs auteurs sont très souvent rémunérés sur fonds publics ». Ces éditeurs disent soutenir « l'objectif d'améliorer les conditions d'accès à l'innovation scientifique », mais redoutent un bouleversement de modèle économique qui pourrait écraser les revues de langues françaises de SHS, notoirement plus petites et plus fragiles que des grands groupes anglo-saxons.

En réponse à ce texte, des responsables d'universités, chercheurs et professionnels de bibliothèques répondent en faisant paraitre dans Le Monde un article titré « Qui a peur de l'open access ? », également signé par l'ADBS. Ils y expriment leur refus de voir les SHS se couper d'un mouvement de démocratisation du savoir car « Sortir les savoirs des silos et des frontières des campus, c'est les ouvrir à tous, c'est reconnaitre à la connaissance un rôle moteur, c'est ouvrir des perspectives d'enrichissement collectif. »

A des craintes d'une nature économique par les éditeurs, cet article répond par des arguments de principe difficilement contestables. Mais les éditeurs ne réclamaient tout de même (du moins ouvertement) qu'une concertation sur les enjeux de l'Open access…

Reste à savoir quelles seront les positions adoptées par ces mêmes groupes sur le sujet précis de l'ouverture des données de la recherche : seront-elles aussi tranchées ? Le débat sera-t-il aussi vif ? Le Groupement Français des Industries de l'Information (GFII), qui sur la question des publications adopte un positionnement proche de celui des éditeurs rédacteurs de la motion de février 2013[41], n'a pas encore, en novembre 2014, publié de texte définissant son point de vue.

Mais pourquoi l'ADBS, qui représente les documentalistes, se positionne-t-elle en faveur de deux textes qui se répondent en s'opposant ? N'ayant pas trouvé de réponse claire et tranchée dans les textes de l'association, nous ne pouvons qu'émettre des hypothèses. Il pourrait par exemple s'agir du reflet d'une diversité de points de vue au sein de l'association.

Sur la question spécifique de l'ouverture des données de la recherche, il n'y a pas encore d'expression des documentalistes, au sens de prise de position. Le numéro récent (début 2014) de Doc Sci[42] est axé sur la donnée en général, et fait donc la part belle aux données administratives et de management de l'Etat et des collectivités territoriales. L'idée principale semble bien être, comme pour d'autres groupes de métiers de l'information, qu'il y a là une nouvelle matière à travailler, et qu'il s'agit d'adapter ou d'acquérir des compétences, comme de mettre en place des méthodes et des outils. Et cela afin de montrer que les documentalistes sont légitimes au sein de ce « nouveau monde » de l'information.

---

[41] On peut consulter sur le site du GFII les positions du groupement. En ce qui nous concerne, le groupe de travail sur l'*Open access* a rédigé des conclusions suite aux recommandations de la CE. <http://www.gfii.fr/fr/document/recommandations-de-la-commission-europeenne-en-matiere-d-open-access-premieres-observations-du-gfii> [En ligne] Consulté le 05/11/2014.

[42] « Les métiers de l'information et la « donnée » : analyse d'un monde en mutation ». *Documentaliste-Sciences de l'Information* 50, n° 3 (2013). < http://www.cairn.info/revue-documentaliste-sciences-de-l-information-2013-3-page-26.htm> [En ligne] Consulté le 16/06/2014.

- L'association des archivistes français (AAF) et sa section AURORE (réseau des archivistes des universités, rectorats, organismes de recherche et mouvements étudiants) a publié en janvier 2014 un numéro de la Gazette des Archives dédié aux archives de la recherche[43]. Charlotte Maday, présidente de la section AURORE, précise le point de vue des archivistes de la recherche dans une interview publiée sur territorial.fr en octobre 2014. En forçant le trait pour être synthétique, nous pourrions dire qu'il en ressort que les données de la recherche devraient être des archives comme les autres, et donc traitées par les archivistes, à l'aide des notions d'archivistique comme celle de « cycle de vie » peu usitée par les autres métiers de l'information. L'inquiétude est vive, chez ces professionnels, de la place que la communauté de la recherche voudra bien leur laisser. Cela s'explique aisément quand on sait que la notion d'archive de la recherche est récente, et que les archivistes sont très peu nombreux au sein des établissements de recherche.
  Les textes cités ont un autre intérêt que celui de nous éclairer sur cette volonté de participer au travail des données : celui de revenir à l'idée d'une définition, commune à tous les professionnels de l'information scientifique et aux chercheurs, des données de la recherche, et de partir de celle-ci pour redéfinir le rôle de chacun, tout au long du cycle de vie des données.

- Des professionnels de l'IST et des chercheurs sur des blogs, s'expriment à titre moins officiel, comme sur la plateforme hypotheses.org. Les points de vue sont plus individuels, et donc très divers. Pour exemple, attardons nous sur un long billet de Sylvie Fayet[44], conservatrice des bibliothèques à l'URFIST[45] de Toulouse.
  En partant d'un point de vue très concret de professionnelle de l'IST, ce texte rend compte des nombreuses questions que les données de recherche posent aux métiers de l'information : que sont exactement les données de la recherche ? Les données brutes existent-elles ? Quel est l'objectif du travail sur ces données ? Où se situe la frontière avec les publications ? Que faire quant aux droits d'auteur ? Et ces questions, comme d'ailleurs les conclusions de l'article, soulignent l'importance du travail sur les contenus avec les chercheurs et l'intérêt de penser l'ensemble des productions scientifiques. Ce texte rappelle également que deux sortes d'initiatives cohabitent : de grands réservoirs disciplinaires et internationaux, lieux « naturels » de dépôt pour toucher une communauté scientifique, d'une part, et des projets d'établissement, d'autre part, qui permettent de rendre visible la production scientifique d'une institution de recherche. Il faut penser dès maintenant l'interopérabilité entre ces deux types d'entrepôts, mais aussi une certaine mutualisation du travail de collecte

[43] « Les archives des établissements d'enseignement supérieur et de recherche. » *La Gazette des archives*, n°231 (2013-3).
[44] FAYET, Sylvie, op. cit.
[45] URFIST : Unité Régionale de Formation à l'Information Scientifique et Technique. Les 7 Urfist ont pour mission la formation des usagers aux nouvelles technologies de l'information et de la communication. Ils déploient également une activité de veille et de recherche. Le réseau des URFIST alimente un blog sur hypotheses.org, dont S. Fayet est une des contributrices. <http://urfistinfo.hypotheses.org/> [En ligne] Consulté le 04/06/2014.

des données et d'alimentation des réservoirs, si ce que nous voulons mettre en place est bien un système d'accès pérenne et mis à jour.
Et cela, de notre point de vue, présente l'intérêt de remettre la question du travail sur les données au centre des préoccupations.

Les professionnels de l'information dans leur ensemble consacrent de nombreux textes à la réflexion sur le rôle qu'ils peuvent et souhaitent jouer quant aux données en général et à leur gestion en particulier. Ils évoquent entre autre la question des compétences nécessaires, les difficultés rencontrées lors de la participation à des projets de partage ou d'ouverture de données (publiques ou de recherche), mais aussi des premiers succès et des méthodes concluantes. Mais ils évoquent finalement peu les débats sur l'ouverture elle-même et l'accès aux données de la recherche. Celui-ci semble acté, et les communautés professionnelles ne remettent pas le principe de l'ouverture en question.

1.3.3 *Research data* : du côté de *l'Open data* ou de l'*Open access* ?

Quel que soit le débat (s'agit-il *d'Open Data* ou d'*Open access*, d'une intersection des deux ou d'autre chose ? Pour ou contre l'ouverture des données de la recherche ? Quelle ouverture ? Qui doit s'en occuper ? Etc.), il faut en revenir à la question de la définition des données de la recherche. Si ce sont des données publiques, elles appartiennent à l'ensemble des données que l'Etat doit rendre accessibles. Si elles appartiennent d'abord à leur auteur-créateur-agrégateur, c'est à lui que revient la décision… du moins avant que l'ensemble du travail ne fasse partie des archives de la recherche, qui sont régies par le code du patrimoine. La définition porte en elle la question du statut juridique.

Mais, sous-jacente, la question des objectifs n'est pas moins centrale : le degré d'ouverture, le public ciblé se définiront en fonction du but poursuivi. S'il s'agit avant tout de partager entre chercheurs, pour enrichir la recherche, pour permettre des validations, pour valoriser le travail effectué…une ouverture partielle, restreinte à la communauté scientifique, suffira. Mais s'il s'agit d'abord de favoriser l'innovation au sein d'une partie du secteur privé, les initiatives individuelles, alors il est nécessaire que l'ouverture soit totale.

Au fond, les concepts aux fondements de *l'Open access* n'ont pas attendu le numérique : la révolution internet les favorise, les simplifie et les amplifie. Mais les astronomes ont pensé aux formats de partage de leurs données dès 1976[46]. Et n'ont jamais cessé depuis d'échanger leurs données. La science, par nature, se nourrit d'elle-

[46] GENOVA Françoise. Du nécessaire partage des données scientifiques : l'exemple de l'astronomie. In Arabesque 73, Fév-Mars 2014, ABES, p. 12-13. <http://www.abes.fr/Arabesques/Arabesques-n-73>
[En ligne] Consulté le 3/06/2014.

même, et les chercheurs doivent partager leurs résultats, tant pour les vérifier et être évalués par leurs pairs, que pour avancer.

Quant à *l'Open data*, il était certes contenu en germe dans les lois sur l'accès aux informations administratives mais restait plus ou moins lettre morte dans les faits avant l'apparition d'Internet et les politiques récentes d'ouverture des données publiques.

Selon qu'on tire les *Research data* du côté de l'*Open data* ou du côté de *l'Open access*, elles tombent plus ou moins du côté du *Research* ou du côté de la data… Et ces choix conditionnent toute suite : selon le public et la réutilisation visés, le choix des licences comme celui des outils et des formats varie. Ainsi que les arguments/injonctions pour convaincre/pousser les chercheurs à partager leurs données[47].

Ces débats semblent parfois faire l'économie de la question de l'intérêt, pour le grand public, d'un accès aux données les plus brutes possibles de la recherche. Or, c'est une question qui devrait se poser : la vulgarisation scientifique est une tâche ardue mais nécessaire, si l'un des objectifs reste que chacun puisse se faire une idée des débats et orientations qui animent la recherche et la science. L'accès technique à des données primaires, même contextualisées, ne permet pas forcément et à lui seul un accès réel, c'est-à-dire une compréhension de ces données.

Le site de l'INSEE présente ainsi de très nombreuses données issues de la statistique publique. Tout internaute peut y accéder. Mais en complément, le site propose des documents d'éclairage, des synthèses, des textes explicatifs, et une rubrique composée de définitions des termes techniques très riche. Sans ces documents -plus ou moins- pédagogiques, seule une minorité aurait, potentiellement, accès à l'information. Cette solution ne résout pas l'ensemble des soucis d'accès : une dose de volontarisme – voir un peu d'acharnement - est nécessaire aux non-statisticiens pour appréhender ces informations. Mais du moins est-ce possible. Des outils méthodologiques sont à disposition. La transparence n'est certes pas immédiate, du moins est-elle potentiellement là.

Qu'il soit nécessaire de mettre le maximum d'informations à disposition du plus grand nombre ne peut être contesté. Mais il serait regrettable que cela soit considéré comme suffisant. L'aspect « participatif » des projets d'*Open data*, qui permet que quelques « *geeks* » réutilisent, sur leur temps libre et bénévolement, ces données et en tirent des data-visualisations, suffira-t-il à pallier les difficultés d'accès inhérentes à l'extrême spécialisation scientifique ? Et que faire au sujet des formats informatiques spécifiques utilisés pour créer, et donc lire, ces données spécifiques ?

[47] Les questions concernant le statut juridique des données et de droits d'auteurs sont très bien posées au chapitre sur « Les enjeux juridiques de l'ouverture » du mémoire de R. Gaillard, op. cit. p.44 et suivantes. Entre autre, l'auteur explique les raisons pour lesquelles les licences et *waivers* (que l'on pourrait traduire par renonciation) sont, en l'état actuel des lois, la solution la plus adaptée pour une ouverture de données de la recherche.

1.3.4 *Research data* : des objectifs, des promesses et des risques

Outre la question de principe qui fait des connaissances scientifiques une partie du bien commun, nous avons rencontré trois types d'arguments qui expliquent l'intérêt et les objectifs de l'ouverture des données de la recherche :

- la transparence et la reproductibilité des résultats, comme le montre par exemple la polémique, en mai 2014, autour des données en économie de Thomas Piketty, enseignant à l'EHESS. Celui-ci a pu défendre sa bonne foi et sa position contre le Financial Times grâce à la mise en ligne des données (les mêmes qui avaient servi à semer le doute !) sur lesquelles il appuyait son raisonnement et ses calculs.
- la réutilisation (tant par la science que par les entreprises et les citoyens) et les économies, voir l'économie (au sens d'un modèle) que cela pourrait générer,
- mais aussi la valorisation d'un travail scientifique sur les données, souvent peu visibles, et la possibilité de citer des jeux de données comme cela se fait avec les publications traditionnelles.

Mais il existe également quelques raisons de faire attention aux modes de diffusion et de partage : la dissémination des données, les données personnelles, les illusions/confusions qui font passer les données pour une « fontaine du savoir », l'éventuel déséquilibre d'un modèle économique pour certains pans de l'édition scientifique…ne sont pas des motifs futiles.

Une autre série de questions se pose, que l'on voit apparaitre dans certains textes cités plus haut : en fonction des disciplines, la quantité et le type de données sont très différents. Existe-t-il des disciplines qui créent ou diffusent plus de données que d'autres ? Nous avons interrogé quelque grands catalogues d'entrepôts de données pour établir une très rapide estimation. Et par exemple :

- OAD (Open access directory)[48]propose une liste d'entrepôt de données en *Open access* qui présente 11 entrepôts en SHS (sciences sociales et archéologie), 16 entrepôts multidisciplinaires et 79 en sciences « dures » (physique, chimie, énergie, géoscience, sciences marines…).
- Re3data*, catalogue d'entrepôts de données en libre accès, répond à une interrogation de son moteur de recherche par 244 *data repositories* en SHS pour 465 en sciences naturelles, et 98 en sciences de l'ingénieur, sur un total de 915 entrepôts présentés.

A regarder ces listes, les SHS paraissent bien, quantitativement, être les « parents pauvres » des *data*.[49]

[48] <http://oad.simmons.edu/oadwiki/Data_repositories> [En ligne] Consulté le 15/07/2014.
[49] Mais est-ce lié à la nature des disciplines, objets et données de recherche des SHS ? On peut en faire l'hypothèse.

En somme, la question n'est pas tant de savoir : faut-il partager les données de la recherche ? Mais plutôt : quelles sont les données qu'il faut partager ? Et, bien sûr, en corollaire : comment le faire ? Avec quels moyens ?

DEUXIEME PARTIE

II Où sont les données de la recherche, ou quelle accessibilité pour les *Research data* en France ?

2.1 Le mille-feuille institutionnel où se perdent les données de la recherche

Même en se cantonnant à la recherche publique, le système de recherche français peut être décrit comme difficile à appréhender[50]. D'après les pages du site du ministère de l'Enseignement Supérieur et de la Recherche (MESR) consacrées à l'organisation du système de recherche :

- la politique et l'orientation de la recherche, définie par le MESR, d'autres ministères et les régions, se traduit par
    - la Stratégie Nationale de Recherche (SNR) qui est révisée tous les cinq ans ;
    - mais aussi par l'agenda stratégique pour la recherche, le transfert et l'innovation, dit "France Europe 2020", élaboré par le Conseil Stratégique de la Recherche (CSR), mis en place en décembre 2013 et qui remplace le Haut Conseil de la Science et de la Technologie.
- La programmation est établie en partenariat avec les agences de financement (Agence Nationale de le Recherche, Bpifrance…), les pôles de compétitivité, OSEO, certaines fondations et en accord avec l'espace européen de la recherche, sans exclure un certain nombre d'alliances pour des domaines clés ;
- le Haut conseil de l'évaluation de la recherche et de l'enseignement supérieur (qui remplace l'Agence d'évaluation de la recherche et de l'enseignement supérieur, AERES) peut conduire directement des évaluations ou s'assurer de la qualité des évaluations réalisées par d'autres instances en validant les procédures retenues.
- Enfin, la recherche proprement dite est réalisée par :
    - 74 universités,
    - 26 pôles de recherche et d'enseignement supérieur (PRES) ;
    - plus d'une centaine de grandes écoles et établissements d'enseignement supérieur ;
    - une trentaine d'organismes de recherche à vocation multidisciplinaire (CNRS) ou finalisée (INSERM, INRA, INRIA, CEA, CNES, IFREMER, IFSTTAR, etc.), Institut Pasteur, Institut Curie.

  Cette activité à donc lieu parmi les établissements d'enseignement supérieur (EPSCP.), les établissements publics à caractère scientifique et technologique (EPST), certains établissements publics à caractère industriel et commercial (EPIC), les centres hospitaliers universitaires (CHU.), les associations et

[50] SOULE, Véronique. Regroupement des universités : le haussement des pôles, et interview de Pierre-Paul Zalio, président de l'ENS Cachan : « Le système français est complexe et illisible. » in Libération du 19/10/2014.

fondations principalement financées par l'État, et quelques établissements publics administratifs et services ministériels[51]. Soit une multiplicité d'acteurs, mais également de statuts.

Afin de « remettre de l'ordre dans le grand enchevêtrement de l'enseignement supérieur français »[52], la rentrée universitaire 2014 voit la mise en place de regroupements de facultés, organismes de recherche et écoles, concrétisant une mesure de la loi sur l'enseignement et la recherche du 22 juillet 2013. Mais la page du site du MESR[53] qui liste les établissements du supérieur actualisée le 29 septembre 2014 prévient :

> *« Attention : les statuts des communautés sont en cours de révision et feront l'objet pour chacun d'eux de la publication d'un décret d'approbation. Le nombre de communautés est donc susceptible d'évoluer. La liste ci-dessous sera donc mise à jour régulièrement. »*

La question émergente en France des données de la recherche se situe donc dans ce cadre institutionnel[54] foisonnant mais rigide, changeant au gré des réformes[55] ou des réorganisations[56] mais toujours délicat à « attraper », cerner.

De plus, la recherche est aussi financée et évaluée au niveau européen[57], au sein de l'Espace Européen de la Recherche et du Conseil Européen de la Recherche (CER, ou ERC pour *European Research Council*). Le cadre institutionnel et administratif de la recherche publique française se structure donc également à l'échelle de l'Europe.

Cette organisation nationale et européenne se double d'un aspect international inhérent à la recherche et aux échanges d'informations rapides et répétés, accélérés par le développement d'Internet. Le labyrinthe institutionnel franco-européen se superpose ainsi avec le millefeuille international et disciplinaire du monde de la recherche.

Il n'est donc pas aisé de comprendre les rôles de chacun des acteurs quant aux données de la recherche. Aux questions : qui finance leur gestion ? Quelles règles encadrent leurs usages ? Que doit-on faire des données déjà accumulées ? Que prévoir

---

[51] France. Ministère de l'enseignement supérieur et de la Recherche. L'état de l'Enseignement supérieur et de la Recherche, n°7, Mars 2014. http://cache.media.enseignementsup-recherche.gouv.fr/file/EESR_2014/60/7/EESR7_316607.pdf [En ligne] Consulté le 8/08/2014.

[52] SOULE Véronique, Op. cit.

[53] <http://www.enseignementsup-recherche.gouv.fr/cid49705/etablissements-d-enseignement-superieur-et-de-recherche.html> [En ligne] Consulté le 20/10/2014.

[54] Voir les descriptions sur le site du Ministère de l'enseignement supérieur et de la Recherche. <http://www.enseignementsup-recherche.gouv.fr/pid24888/recherche.html>[En ligne] Consulté le 6/09/2014.

[55] Dernière en date : la loi n°2013-660 du 22 juillet 2013 relative à l'enseignement supérieur et à la recherche.

[56] Par exemple, le TGE Adonis est devenu en juillet 2013 la TGIR Huma-Num, en fusionnant avec la TGIR Corpus-IR.

[57] Comme le montre l'existence de programmes européens pour la recherche, dont Horizon 2020, dit H20-20, entré en vigueur en janvier 2014, et qui prends en compte la gestion des données de la recherche.

pour les projets de recherche en cours ? Qui est censé s'en occuper ? Et bien sûr, comment y accéder ? La réponse – de directeur de laboratoire, de personnel administratif d'établissement de recherche ou de professionnels de la gestion de l'IST - est invariablement : tout dépend – du contexte, des besoins, du type de données, du statut de l'organisme ou du chercheur, de la discipline concernée, etc.

2.2 Les acteurs et initiatives en France pour les sciences humaines et sociales

Au sein de ce paysage complexe et sur la question des données de la recherche, quelques acteurs émergent et semblent incontournables, à lire la littérature tant scientifique qu'administrative ou juridique traitant des *Research data* « à la française ». Ce sont parfois des institutions anglo-saxonnes comme le JISC (Joint Information Systems Comittee)[58] et son *Digital Curation Center* (DCC), ou des associations internationales, comme DataCite.

2.2.1 Une vue d'ensemble

C'est à une clarification de ce paysage que nous avons essayé de participer en élaborant une cartographie des initiatives et acteurs visibles en France concernant les données des sciences humaines et sociales (SHS). Nous avons tracé ce périmètre parce qu'il est nécessaire de se limiter sous peine de ne jamais terminer, mais surtout parce que c'est là que le manque nous a paru le plus important au regard de ce que nous avons pu lire sur le sujet, comme nous le signalions en introduction.

Il est impossible de s'arrêter aux initiatives purement françaises car bon nombre de projets en France sont des émanations de projets européens ou internationaux. On rencontre le même problème pour dessiner les contours des initiatives par disciplines, ou par grands champs disciplinaires : si l'on se limite aux SHS, que faire des projets inter disciplinaires ? Les nouveaux projets se multiplient (à voir par exemple le site d'information sur les données de la recherche, onglet « initiatives et projets »), il semble hasardeux de se prétendre exhaustif.

Lors de l'élaboration de la cartographie présentée en annexe (voir annexe 1), nous avons retenu tous les projets, acteurs et initiatives concernant les données de recherche, ayant un impact en France, et concernant au moins une science dite humaine ou sociale, que nous avons rencontrés lors de lectures (voir bibliographie) ou de rencontres et conférences professionnelles (notamment Datacite* 2014 à Nancy). Nous souhaitions simplement répondre à la question introductive : qui fait quoi ? Ou encore : quels acteurs pour quelles initiatives ?

A partir de ce que nous avions tout d'abord envisagé comme un simple glossaire, nous avons établi une proposition de typologie des initiatives, et reliées chacune d'elle au(x) acteur(s) qui les portent ou s'y adossent.

[58] Organisme public financé par les *Research Councils* et chargé de promouvoir auprès des universités britanniques l'utilisation des technologies de l'information et de la communication.

La première difficulté rencontrée a été de classer les entrées de ce glossaire dans nos premières catégories : Acteurs et Initiatives. Par exemple, Datacite est à la fois un acteur, en tant que consortium international, et une initiative qui consiste à accompagner la mise en place d'identifiants pérennes de type DOI* pour les jeux de données scientifiques. Plus compliqué, encore : le cas d'Horizon 2020*. Il s'agit d'un programme cadre. C'est à la fois une initiative européenne bornée dans le temps, et un important acteur du financement de la recherche en Europe, porteur d'initiatives…

Nous avons donc posé comme critère premier pour effectuer un choix la question de la durée pour dissocier les acteurs des initiatives. Si un «actant » -pour reprendre la terminologie de Bruno Latour[59]- est borné dans le temps, et ce dès sa conception, avec donc une date de fin connue, nous l'identifions à une initiative. Au contraire, si l'actant ne se définit pas par une date de clôture, il sera en général catégorisé comme acteur. Mais dans certains cas, cela ne suffisait pas. Un second critère est le fait d'être posé comme un outil (comme un entrepôt, un modèle d'archivage ou un programme informatique d'identification pérenne…) et/ou un service (d'aide à la gestion des identifiants pérennes, d'archivage à long terme ou d'aide à la création de *Data Management Plan*[60]…) proposé par un acteur ; dans ce cas, nous l'identifions à une initiative. C'est en croisant ces deux critères que la liste d'actants qui forment les entrées de la cartographie s'est constituée.

Il s'agissait ensuite d'établir une liste d'exclusion : certains actants de la recherche et de l'IST pourraient légitimement figurer dans la cartographie, mais non sans alourdir terriblement ce qui se veut un outil de travail et de réflexion sur la question des données de recherche. Par exemple, le schéma de Dublincore, très utilisé et souvent mentionné lorsqu'on parle de données de la recherche, pré existe à notre problématique et a de multiples usages en dehors de ce cadre. Les acteurs incontournables de la recherche et de l'IST en France – le ministère de l'Enseignement Supérieur et de la Recherche (MESR) et le CNRS, notamment, et son service IST, l'INIST, les universités et grandes écoles - ont également été écartés pour les mêmes raisons. A contrario, nous avons retenu des actants qui œuvrent pour la recherche, bien au-delà des données, comme l'initiative nationale BSN, dont un seul segment concerne directement les *Research data* ; ou Couperin, consortium français qui se préoccupe de toutes les ressources numériques nécessaires à la science. Mais ceux-ci ne sont observés ici que du point de vue de leur impact sur les données de la recherche.

Par manque de temps, nous avons également fait le choix de ne pas mentionner les *Data journals* parmi les initiatives (ou acteurs ?)[61], qui représentent pourtant une intéressante façon de penser la valorisation des données de recherche, et mériteraient

[59] Bruno Latour. C*hanger de société. Refaire de la sociologie*, Paris, La Découverte, 2006. Ce terme permet de réconcilier acteurs, objets et organisations en une même entité, qui se définit par sa capacité à peser sur le déroulement d'une action.
[60] En français : Plan de gestion des données. Généralement abrégé en DMP ou PGD.
[61] Sur la place des *Data journals* quant à la valorisation des données de la recherche, on peut lire le chapitre *Solutions éditoriales et data journals* du mémoire déjà cité de R. Gaillard, p.53-54.

à eux seuls un chapitre. Les réseaux sociaux scientifiques ont également été mis de côté, pour les mêmes raisons.[62]

Malgré ces restrictions, le glossaire est devenu cartographie et comprend trente-cinq entrées. Ce foisonnement témoigne d'une grande activité et d'un dynamisme certain. Mais il est aussi une des causes du retard avec lequel la recherche française se saisit de la gestion des données. En effet, si plusieurs semaines sont nécessaires avant toute orientation au sein de ce paysage, quels chercheurs prendront le temps de mettre de côté leurs activités de recherche pour se repérer avant d'organiser une recherche ou un partage de données ? Il ne suffira pas de les convaincre de l'utilité de ce travail : si celui-ci s'avère long et compliqué, ils ne l'entameront tout simplement pas – du moins, pas seuls. Ils continueront de se contenter d'échanges informels au sein de leur réseau traditionnel de chercheurs.

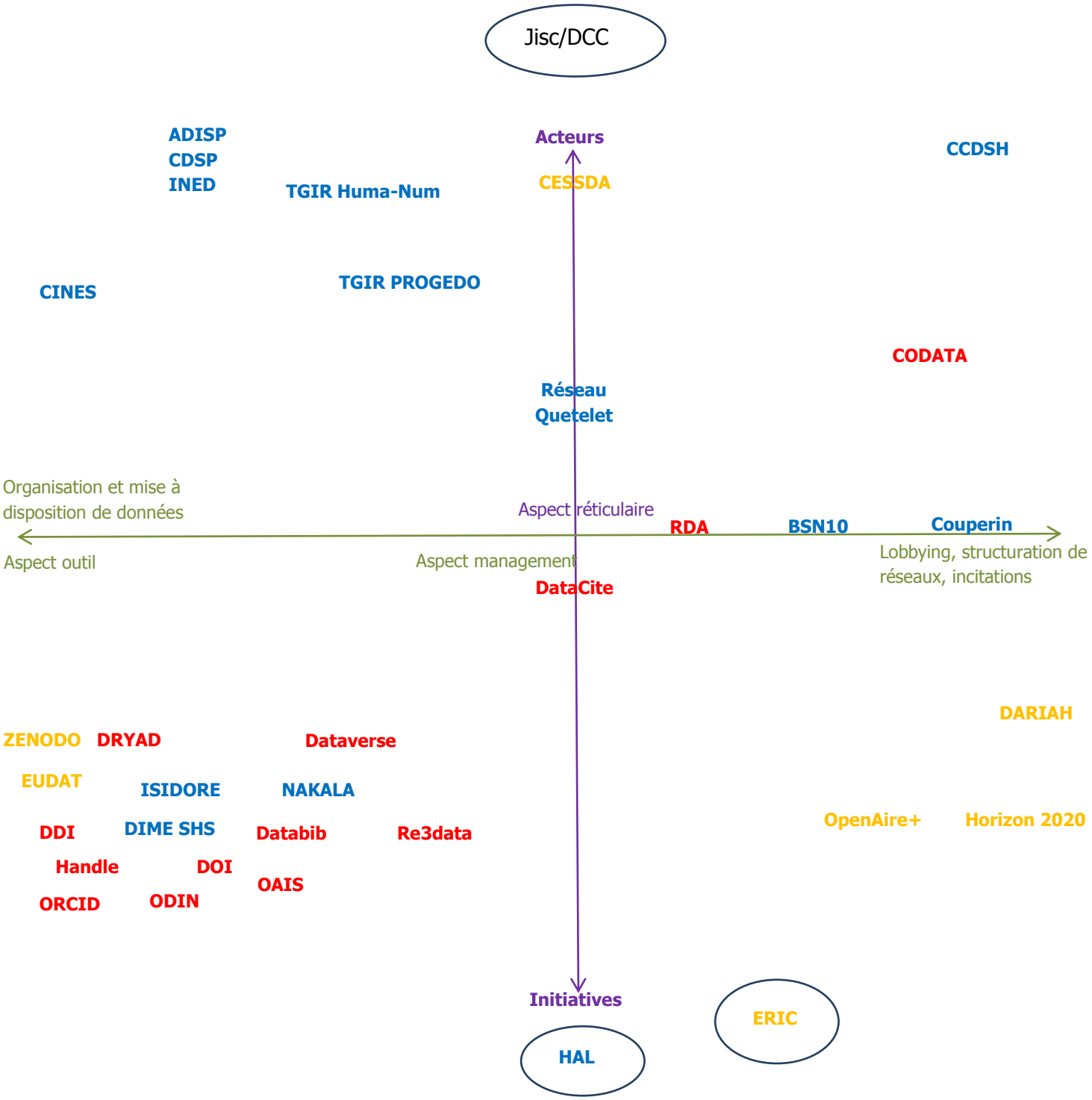
Pour éclairer le caractère éclaté des acteurs et initiatives, et afin de participer à une forme de « balisage » du travail de gestion des données, une visualisation nous a paru nécessaire.

En fonction de leur rôle d'acteur ou d'initiative (en ordonnée) et objectifs (en abscisse), nous avons positionné ces actants sur un plan cartésien. Les objets présentés hors cadre sont extérieurs, soit au territoire national (pas d'effet directs en France), soit aux données de la recherche proprement dites (qui concerne plutôt les publications, par exemple).

Cette catégorisation correspond à un « découpage de la réalité qui implique des choix dans le but de construire une typologie, c'est-à-dire un outil permettant de mieux comprendre les pratiques par une « stylisation de la réalité ». Ces modèles constituent des idéaux-types permettant de « simplifier » la réalité pour mieux la comprendre »[63].

[62] Sur ce sujet, on peut trouver de nombreux textes et études récents, comme cette enquête de 2013, menée par le service communication du CNRS : http://corist-shs.cnrs.fr/sites/default/files/evenements/brigitteperucca_reseauxsociaux.pdf ou cet article de Richard Van Noorden de Nature en 2014 : http://www.nature.com/news/online-collaboration-scientists-and-the-social-network-1.15711?WT.ec_id=NEWS-20140819 et, sur le blog Archeorient, par Christophe Benech en 2014 : http://archeorient.hypotheses.org/2554 [En ligne] Consultés le 2/10/2014.

[63] Schnapper Dominique, *La compréhension sociologique*. *Démarche de l'analyse typologique*. Paris, PUF, 1999, p.2, cité par BERHUET Solen. *Les chômeurs et les intermédiaires de l'emploi : Une sociologie dynamique de leurs trajectoires au sein d'une Maison de l'Emploi*. Thèse de sociologie, dir. M. Lallement et P. Nivolle, Cnam, 2013. P. 343.



Légende :
Item : français
Item : Européen
Item : International

Figure 1 : Acteurs et initiatives en SHS en France

Sur ce schéma, l'abscisse (ici, en vert) représente le type d'action ou de réalisation d'un actant. Plus il tend à créer des outils, plus il se trouve vers l'ouest. Au contraire, plus il œuvre pour la mise en place de politiques ou organise des actions de lobbying ou d'incitation, plus il sera placé à l'est. S'il est proche du centre, et donc au croisement avec l'ordonnée, cela indique son action managériale quant aux données.

En ordonnée (ici, en violet), c'est le type d'actant qui est caractérisé. Un actant situé vers le nord se comporte comme un acteur sur la question des données de la recherche. Un actant situé vers le sud ressemble à une initiative ou à un projet. Au croisement avec l'abscisse se situent les actants qui fonctionnent de façon très notable en réseau.

Notons que ces positions ne sont pas figées : les évolutions sont rapides, et bouleversent parfois la position d'un actant. Pour ne signaler qu'un changement prévu, fin 2015, les catalogues d'entrepôts Databib* et Re3data* devraient fusionner ; DataCite* prendrait alors la direction de l'outil. Ce nouveau service dont on ne connait pas encore le nom sera alors à positionner comme initiative, tandis que Databib et Re3data disparaitront de ce graphique.

Enfin, la présence directement sur l'abscisse ou sur l'ordonnée de six actants signale bien que nous sommes en présence de statuts hybrides, tant leurs rôles et activités sont difficiles à positionner.

L'aspect le plus frappant de ce schéma est de dessiner un paysage constitué en majorité :

- D'initiatives multidisciplinaires, souvent internationales (13 items dans le quadrant sud-ouest dont 9 internationales) ;
- D'acteurs français émanent des autorités compétentes pour la recherche (6 items dans le quadrant nord-ouest).

Parmi les projets de type initiatives, on trouve aussi bien des entrepôts (comme ZENODO* ou DRYAD*) et des catalogues d'entrepôts (Databib et Re3data), que des outils comme des schémas de métadonnées (comme DDI*, norme de description de données en SHS), ou des projets portants en germe plusieurs outils (DIME SHS*).

En comparaison, l'Europe de la recherche est plus présente au sein du quadrant sud-est ; elle se situe davantage dans une logique d'initiatives incitant à l'ouverture des données, notamment avec un huitième plan cadre, héritier du 7e PCRD (FP 7, en anglais), intitulé Horizon 2020* (abrégé en H2020) et l'outil que constitue OpenAire+* pour l'Open research data en Europe.

2.2.2 Quelques gros plans

Pour mieux cerner le contenu de ce « territoire », examinons quelques actants présents dans le graphique, choisis parce qu'ils représentent un quadrant ; mais aussi parce qu'ils sont particulièrement liés à d'autres actants, ce qui permet de décrire ces relations.

Le Réseau Quetelet, un acteur central et hybride.
Il s'agit du réseau français des centres de données pour les sciences humaines et sociales, créé en 2001[64]. A lui seul, cet actant va nous permettre d'en mentionner huit autres présents sur notre graphique.

Quatre unités partenaires constituent le noyau du Réseau Quetelet* : l'équipe des Archives de Données Issues de la Statistique Publique* du Centre Maurice Halbwachs (ADISP-CMH), le centre d'accès sécurisé distant aux données (CASD), le Centre de Données Socio-Politiques* de Sciences-Po (CDSP), et le Service des enquêtes de l'Institut National d'Etudes Démographiques* (INED). Le Réseau Quetelet s'adresse en priorité aux chercheurs français et étrangers. Il est la porte d'accès aux données pour les sciences humaines et sociales, via son portail et son catalogue de données. Il est donc par définition au centre des questions relatives aux données de la recherche des SHS en France.

En termes de statut et de fonctionnement, le Réseau Quetelet est tout à la fois :

- Une des composantes de l'infrastructure PROGEDO* (PROduction et GEstion des DOnnées), très grande infrastructure de recherche (TGIR) des sciences humaines et sociales qui assure la collecte, l'archivage, la documentation et l'accès des chercheurs à ces données. PROGEDO-Quetelet est une unité mixte de service (UMS) associant le CNRS et l'EHESS.
- Sous la responsabilité du Comité de concertation pour les données en sciences humaines et sociales* (CCDSHS, parfois abrégé en Comité des données); dans le cadre de la politique du Comité des données, le Réseau Quetelet coordonne les activités d'archivage, de documentation et de diffusion des données en sciences humaines et sociales de l'ADISP*, du CDSP* et du service des enquêtes de l'INED*. L'action du réseau est organisée par Roxane Silberman, secrétaire générale du CCDSHS. Ce comité établit un lien direct entre les missions fixées par le décret D2001-139 du 12 février 2001[65] pour les données en sciences humaines et sociales, les données archivées par les unités partenaires et le milieu de la recherche.
- Comme son nom l'indique, c'est un réseau : ce sont les quatre unités partenaires qui assument le fonctionnement de Quetelet. Trois plateformes universitaires de données (PUD) sont en charge d'actions de proximité pour la formation des usagers.

[64] Voir l'article en anglais <http://www.insee.fr/en/ppp/sommaire/cs11i.pdf>
[65] Décret no 2001-139 du 12 février 2001 portant création du comité de concertation pour les données en sciences humaines et sociales. [En ligne] < http://www.reseau-quetelet.cnrs.fr/spip/IMG/pdf/Decret_01-139-3.pdf>

- C'est aussi un groupement d'Intérêt Scientifique (GIS), fondé par le CNRS, l'Institut National d'Etudes Démographiques (INED), l'Ecole des Hautes Etudes en Sciences Sociales (EHESS), l'université de Caen-Basse Normandie, l'Université des sciences et technologies de Lille 1, la Fondation nationale des sciences politiques, l'Ecole d'Economie de Paris (EEP) et L'Institut de Recherche et Documentation en Economie de la Sante (IRDES). L'objectif du GIS Quetelet consiste à assurer la coordination des orientations proposées par le CCDSHS. Il a pour mission l'appui à la collecte, à la documentation, à la préservation et à la promotion d'un vaste ensemble de données françaises nécessaires aux disciplines des SHS dans un cadre européen et international. Le GIS gère notamment les modalités de coordination et de coopération entre les établissements partenaires pour les aspects scientifiques (archivage, documentation et diffusion des données), pour l'administration du portail national ainsi que pour les représentations internationales de ses activités dans les structures européennes ou internationales du domaine.

- Quetelet est également membre du *Council of European Social Science Data Archives** (CESSDA)[66], le réseau européen des banques de données pour la recherche en sciences sociales. Le CESSDA fournit un catalogue[67] des jeux de données de recherche mis à disposition en Europe. L'interface permet une recherche libre en neuf langues.

Réseau d'acteurs, GIS, composante d'une UMS et d'une TGIR, membre d'un réseau européen, partenaire pour un EQUIPEX… la multiplicité des types de réseau et de statuts de Quetelet montre bien qu'il se situe au cœur d'un nœud d'actants concernant les *Research data,* parmi les « incontournables » en SHS. Mais cette hybridation le rend difficile à définir.

En termes de réalisation, Quetelet est également multiple. Le site du réseau se présente comme : « le portail français d'accès aux données pour les sciences humaines et sociales ».

Dans le cadre de ses missions principales –archivage, diffusion et valorisation des données- Quetelet recense plus de 1100 jeux de données en sciences sociales. Un catalogue[68] en ligne permet, au choix, une interrogation libre sur quelques champs (dont le résumé) ou une recherche experte avec opérateurs booléens, mais également une recherche extrêmement fine parmi les questions et les variables[69] d'environ 20% des enquêtes diffusées par le réseau.

Pour documenter les données, le réseau utilise –entre autre- le standard international de métadonnées des données des sciences sociales DDI[70]. Un groupe de travail issu

[66] Voir l'entrée CESSDA de la cartographie en annexe.
[67] Ce catalogue est disponible en ligne : < http://www.cessda.net/catalogue/
[68] Disponible en ligne <http://www.reseau-quetelet.cnrs.fr/spip/article.php3?id_article=128>
[69] Cette recherche porte sur le texte des questions, les codes et étiquettes des modalités de réponse, les noms et les étiquettes de variables et concerne quelques 455 enquêtes, soit près de 182 118 questions ou variables, selon le site du réseau < http://bdq.reseau-quetelet.cnrs.fr/fr/Accueil >
[70] Voir l'entrée DDI de la cartographie en annexe.

des membres de Quetelet a traduit en 2004 -de l'anglais au français- le *codebook* de la version 1.2.2. Lorsque l'on cherche à identifier des interlocuteurs français utilisant ce standard international, c'est le réseau Quetelet que l'on trouve (sur le site ddialliance.org comme sur Wikipédia).

Le réseau Quetelet est également partenaire du projet d'équipement d'excellence DIME-SHS, en tant qu'infrastructure de recherche et au travers des unités partenaires qui le composent.

DIME-SHS (Données Infrastructures et Méthodes d'Enquête en Sciences Humaines et Sociales) est un EQUIPEX de la vague 2010 porté par Sciences Po. Il réunit des partenaires de différents pans de la recherche et de l'enseignement supérieur dans des disciplines variées (sociologie, démographie, sciences économiques, science politique) :

- Des Grandes écoles : Sciences Po, le Groupe des Ecoles Nationales d'Economie et de Statistique (GENES) et Télécom ParisTech ;
- Un Institut de recherche : l'Institut national d'études démographiques (INED) ;
- Une Université : l'Université Paris Descartes ;
- Un service de recherche privée : EDF R&D ;
- Une infrastructure de recherche : le GIS Réseau Quetelet.

DIME-SHS vise à doter la France d'une nouvelle structure en matière de collecte, d'enrichissement et de diffusion des données pour la recherche en sciences humaines et sociales. L'équipement propose des ressources aux chercheurs pour produire ou réutiliser des données dont la qualité repose sur une grande rigueur méthodologique.

Ce projet d'équipement se compose de trois instruments :

- DIME-SHS / Quanti : Un instrument pour les données quantitatives qui prend la forme d'un panel internet, ELIPSS[71] (Etude Longitudinale par Internet pour les sciences humaines et sociales) ; Au sein du consortium DIME-SHS, le CDSP de Sciences Po est responsable de la mise en place du panel, en partenariat avec le Service des enquêtes de l'INED. Soit deux partenaires de Quetelet présent pour cette partie de DIME-SHS*.

- DIME-SHS / Quali : Un instrument pour les données qualitatives qui prend la forme d'un site web, BeQuali (banque d'enquêtes qualitatives). Porté par l'équipe BeQuali du CDSP, unité partenaire de Quetelet, il comprend plusieurs projets et partenariats :

    - Un portail : beQuali[72], et une banque d'enquêtes qualitatives : enQuêtes[73].

[71] Alina Danciu, de l'équipe du CDSP, rencontrée le 11 septembre 2014, nous a présenté la mise en place de ce panel. Les premières diffusions de résultats devraient avoir lieu en 2015. Ce projet pourrait être sous-titré ainsi : « Un panel pour la recherche : des scientifiques au service des scientifiques ».
[72] Présentation et accueil du portail disponible en ligne <http://www.bequali.fr/bequali/>[En ligne] Consulté le 2/09/2014.
[73] Présentation et banque de données disponible en ligne <http://www.bequali.fr/app/>[En ligne] Consulté le 2/09/2014.

- Un consortium de la TGIR Huma-Num* qui fédère des laboratoires de sociologie et de sciences politiques pour constituer des Archives des sciences sociales du politique : archiPolis[74].
- Un projet ANR[75] : reAnalyse, projets d'expérimentation de l'analyse secondaire des données qualitatives.
- Un site ressource : Quali², propose des documents et débats sur les méthodes qualitatives[76].

- DIME-SHS / Web : Un instrument pour les données du web qui offrira des outils pour constituer des corpus et pour les analyser. L'instrument *DIME Web* comprend une série d'outils, proposés par une équipe – celle du médialab de Sciences Po - dont le savoir-faire porte sur les méthodes numériques dédiées aux sciences sociales. Cet outil est le seul du projet DIME-SHS a n'être pas réalisé par une unité partenaire de Quetelet.

On voit que, au sein même d'un projet, les initiatives se subdivisent et sont portées par de multiples acteurs. Au sein des équipes, on trouve d'ailleurs des compétences variées, comme en témoignent nos rencontres avec celles de l'ADISP ou du CDSP[77]. S'y retrouvent des documentalistes, archivistes et bibliothécaires – avec de solides notions de code - mais aussi des chercheurs en SHS - avec de bonnes connaissances en statistique - et des informaticiens qui maitrisent les formats et standards du Web de données et des bibliothèques[78].

L'accès aux données de Quetelet

L'accès aux jeux de données, réservé aux chercheurs et dans un but de recherche, demande une certaine ténacité : « Les conditions d'accès et les procédures varient selon les données. […] Certaines données sont librement consultables sur les sites des unités membres du réseau. D'autres peuvent être commandées en ligne. Vous aurez alors à vous identifier et à remplir une licence d'utilisation. »

Une fois identifié comme intéressant pour un projet, chacun des jeux de données doit faire l'objet d'une commande (à moins de faire partie des données en libre accès sur un site partenaire), c'est-à-dire d'une procédure propre. Le chercheur doit justifier de son besoin et décrire le projet de recherche dans le cadre duquel la commande est effectuée. Il s'engage également à rendre compte de l'utilisation faite des données, une fois le projet terminé. Sur le terrain, l'équipe ADISP et celle du CDSP[79] relèvent que seuls 2 à 3% des chercheurs leur font un retour sur le travail effectué avec les

[74] Présentation disponible en ligne <http://www.bequali.fr/archipolis/> [En ligne] Consulté le 5/10/2014.
[75] L'ANR finance la recherche sur projets. Pour en savoir plus, on peut trouver une description sur le site de l'Agence : <http://www.agence-nationale-recherche.fr/> [En ligne] Consulté le 5/10/2014.
[76] Présentation disponible en ligne <http://bequali.fr/quali2>
[77] Cela se voit également dans la présentation des équipes, comme par exemple celle de l'équipe beQuali du CDSP, disponible en ligne <http://cdsp.sciences-po.fr/enquetes.php?&idRubrique=enquetesQL&lang=FR>
[78] Voir en annexe 9 l'exemple d'une fiche de poste du CDSP pour un ingénieur d'étude.
[79] Nous avons rencontré l'ADISP le 20 juin et le CDSP le 9 septembre 2014.

données fournies par Quetelet. Du coté des utilisateurs, les sociologues et statisticiens de l'IFSTTAR[80], grands demandeurs des données du réseau Quetelet, confient également remettre à plus tard ces comptes rendus, tant le temps leur manque au quotidien. Mais aussi, les procédures d'accès aux *datasets* sont vécues comme longues et répétitives. D'autant qu'il arrive qu'un chercheur commande à plusieurs reprises le même *dataset*, pour différents projets, car chaque utilisation doit être justifiée. Le « poids » de la commande est ainsi comme allégé par le non-respect de l'engagement d'un retour.

Bien sûr, ces précautions sont largement justifiées par le type de données échangées, du point de vue de leur financement – public - comme de leur contenu –parfois sensible et concernant des populations et des individus. Mais en tout état de cause, nous voilà quelque peu éloigné de *l'Open access*. Les chercheurs connaissent certainement les sources et ressources de jeux de données les concernant directement : cet aspect complexe que nous relevons ne freinera que peu leurs accès à des *datasets* de leurs disciplines. Mais, en l'état actuel, des oncologues arriveraient-ils à trouver sans aide des jeux de données issus de la statistique publique ou d'enquêtes sur l'emploi ou le logement[81] ?

La TGIR Huma-Num, un acteur français et l'une des deux TGIR des SHS.

Huma-Num* se situe dans le quadrant nord-ouest du schéma, soit du côté des acteurs et des « fournisseurs d'outils ».

Statut, organisation et fonctionnement :

Née en mars 2013 de la fusion du TGE Adonis et de Corpus IR, Huma-Num est une très grande infrastructure (TGIR) visant à faciliter le tournant numérique de la recherche en sciences humaines et sociales.

- La TGIR Huma-Num est portée par une UMS associant le CNRS, l'Université d'Aix-Marseille et le Campus Condorcet.
- Pour remplir sa mission, « la TGIR Huma-Num est bâtie sur une organisation consistant à mettre en œuvre un dispositif humain (concertation collective) et technologique (services numériques pérennes) à l'échelle nationale et européenne en s'appuyant sur un important réseau de partenaires et d'opérateurs. »

Même si la description de l'organisation citée reste un peu floue, nous nous trouvons devant un actant plus simple à aborder que le réseau Quetelet. Un seul statut, une objectif, un périmètre : une difficulté pourrait résider dans la communication suite à la fusion de deux entités qui avaient acquis une certaine notoriété.

[80] Discussions informelles et entretiens menés entre juin et septembre 2014 dans le cadre du stage.
[81] On peut voir à ce sujet le court film d'animation pédagogique de la NYU Health Sciences Library, Data Sharing and Management Snafu in 3 Short Acts.
<http://www.youtube.com/watch?v=66oNv_DJuPc&feature=youtube_gdata_player>[En ligne] Consulté le 15/08/2014.

Reste le fait que peu de gens, y compris parmi le personnel de la recherche, semblent capables de définir précisément une TGIR, ou d'imaginer quelle aide elle peut apporter concrètement à une équipe de recherche, comme nous le verrons dans la troisième partie.

Missions et services :

- La TGIR Huma-Num favorise, par l'intermédiaire de consortiums regroupant des acteurs des communautés scientifiques, la coordination de la production raisonnée et collective de corpus de sources (recommandations scientifiques, bonnes pratiques technologiques).
- Elle développe également un dispositif technologique unique permettant le traitement, la conservation, l'accès et l'interopérabilité des données de la recherche en SHS. Ce dispositif est composé :
    - d'une grille de services dédiés,
    - d'ISIDORE[82], une plateforme d'accès unifié,
    - et d'une procédure d'archivage à long terme (en partenariat avec le CINES et l'IN2P3).
    - Depuis juin 2014, un nouveau service d'exposition des données appelé NAKALA.
- La TGIR Huma-Num propose des guides de bonnes pratiques technologiques généralistes à destination des chercheurs[83]. Elle peut mener ponctuellement des actions d'expertise et de formation. Elle porte la participation de la France dans le projet DARIAH* en coordonnant les contributions nationales.
- La TGIR Huma-Num a établi un partenariat avec le CINES sur l'archivage à long terme des données numériques, en collaboration avec le centre de calcul de l'IN2P3.

Ces services ont été décrits avec précisions dans une brochure[84] disponible en ligne et éditée en mars 2013, qui fait un point très clair sur ce que propose la TGIR aux chercheurs et aux laboratoires de SHS. Mise à jour en mai 2014, elle intègre une présentation du nouveau service NAKALA*.

Il s'agit donc d'un acteur institutionnel relativement bien défini et de services clairement exposés. Huma-Num ne devrait donc pas poser question. Mais une autre TGIR, elle aussi dédiée aux sciences humaines, coexiste avec Huma-Num. Il s'agit de PROGEDO, dont le réseau Quetelet est une composante. Il est un peu troublant que 2 TGIR s'occupent des données en SHS.
La comparaison sous forme d'un face à face[85] réalisée en mai 2014 par l'Alliance ATHENA de ces deux entités se révèle intéressante pour comprendre les petites

[82] Accessible en ligne <http://www.rechercheisidore.fr/>
[83] Cinq titres disponibles en ligne au 27/10/2014 <http://www.huma-num.fr/ressources/guides>.
[84] HUMA-NUM. Les services de conservation de données proposés par Huma-Num. Mai 2014 <http://www.huma-num.fr/sites/default/files/ressourcesdoc/dossier-thematique-mai2014.pdf> [En ligne] Consulté le 25/06/2014.
[85] Le numéro de mai 2014 de la lettre d'ATHENA, newsletter de l'Alliance ATHENA, est entièrement consacré à ces deux TGIR du numérique pour et en SHS.

différences d'objectifs et de services. Mais malheureusement insuffisante. En effet, examinons par exemple le face à face des objets et des missions de chacune des deux TGIR :

- PROGEDO se situe du côté des données quantitatives, et Huma-Num s'occupe de corpus de données de toute nature. Mais le réseau Quetelet, qui est une partie de cette TGIR, comprend un volet « qualitatif » à travers l'équipe et le portail beQuali.
- Les deux TGIR revendiquent l'accès aux données comme une de leur mission.

Bien sûr, certaines orientations sont différentes, comme les participations à des projets européens :

- PROGEDO est la porte d'accès française aux données européennes des grandes enquêtes ESS (*European Social Survey*) et SHARE[86] ; tandis que Huma-Num coordonne la participation française à l'infrastructure numérique européenne de soutien de la recherche en SHS DARIAH.
- Les services proposés sont également distincts, entre autre parce que PROGEDO ne propose pas de service d'archivage ou d'aide à la gestion de données. Mais les deux acteurs se définissent comme des portes d'accès aux données en SHS. Seulement, ce ne sont pas pour les mêmes données.

Dès lors, où se tourner lors d'une recherche de jeux de données en SHS ? Est-ce que tout dépend de ce que vous cherchez comme type de données ? Ou existe-t-il d'autres subtilités à connaitre avant toute exploration ?

Au vue de cette brève compilation des présentations, il semble que les dispositifs qui doivent simplifier l'accès aux données SHS et la maitrise de leur gestion par la communauté scientifique sont eux même difficiles d'accès.

La norme de description DDI, une initiative internationale

Située dans le quadrant sud-ouest du schéma, il s'agit d'une initiative et plus précisément, d'un outil de normalisation de description.

---

http://www.allianceathena.fr/sites/default/files/telechargements/la_mai_14_a4_2.pdf [En ligne] Consulté le 02/10/2014. Il serait intéressant d'étudier le positionnement des Maisons des Sciences de l'Homme (MSH), de la Fondation des Maisons des Sciences de l'Homme (FMSH) et de l'Alliance ATHENA comme acteurs français quant aux données de la recherche en SHS. Ils ne sont pas apparus dans nos lectures comme y jouant un rôle, mais l'éditorial de F. Thibault (Déléguée générale de l'ATHENA et directrice scientifique à la FMSH) indique une volonté nette de participer aux –voir d'influencer les- débats.

[86] « Réalisé tous les deux ans depuis 2002, auprès de 40 000 individus, l'European Social Survey (ESS) est un programme de production d'une enquête comparative européenne destinée à mesurer les comportements et les attitudes des citoyens des pays membres et de pays non-membres de l'Union européenne sur un ensemble de thèmes socio-politiques. » Présentation disponible en ligne sur le site du CDSP <http://cdsp.sciences-po.fr/enquetes.php?lang=FR&idRubrique=enquetesINT&idTheme=10>

Data Documentation Initiative[87] (DDI) est un projet international (programme de *l'University of Michigan* (UM), géré par la *DDI Alliance*[88]) initié en 1995 par un groupe de travail constitué dans le cadre de IASSIST (*International Association for Social Science Information Service & Technology*). Il s'agissait de créer et maintenir un standard de documentation technique pour décrire et conserver les informations statistiques et, plus globalement, les informations et données d'enquêtes en sciences humaines et sociales. DDI utilise le langage XML ; les schémas XML comme l'ensemble de la documentation sont en libre accès. L'objectif étant de répondre aux besoins de métadonnées nécessaires à la conservation de l'information sur les fichiers, à l'échange de données entre institutions selon des champs communs et de permettre une "recherche intelligente" sur le web.

Mais pourquoi un standard spécifique de documentation des données d'enquête en SHS ?

La réexploitation des données d'enquête nécessite une documentation détaillée et fiable pour autoriser de nouveaux traitements statistiques. Cette documentation d'enquête est constituée :

- des instruments de recueil des données (questionnaires et formulaires) ;
- des référentiels qui ont permis de coder l'information (nomenclatures, dictionnaires de codes ou de variables).

Ce sont donc des documents et des champs de description spécifiques aux données des enquêtes et statistiques des SHS, que d'autres standards ne prennent pas en compte.

La standardisation de cette documentation et du format des fichiers qui la compose facilite à la fois la recherche (variété et richesse des modes de recherche) et la réexploitation de ces données dans de nouvelles études grâce à la précision des données de contexte. Mais cela demande un important traitement et un long travail sur les métadonnées. A titre d'illustration, le *Codebook* de DDI v.1.2.2[89], traduit par le réseau Quetelet, fait une centaine de pages. Or, il s'agit d'un outil de travail. Il représente une forme allégée du standard, dans une version ancienne elle-même plus simple à découvrir que les versions 3 et suivantes, puisqu'elle n'intègre pas la notion de « cycle de vie des données ». Cette notion signifie que, à partir de la version 3 :

> « Les spécifications DDI pour les métadonnées prennent désormais en charge l'ensemble du cycle de vie des données de recherche. Les métadonnées DDI accompagnent et permettent la conceptualisation, la collecte, le traitement, la

[87] Une présentation complète de la norme DDI en anglais est disponible en ligne sur le site de la DDI Alliance <http://www.ddialliance.org/what>

[88] L'accueil du site de la DDI Alliance permet de trouver l'ensemble des informations sur le fonctionnement et l'organisation de cette association, mais aussi sur les évolutions de la norme, les groupes de travail thématiques et les utilisateurs dans le monde <http://www.ddialliance.org/> [En ligne] Consulté le 02/06/2014.

[89] On trouve cette traduction en ligne sur le site du réseau Quetelet <http://www.reseau-quetelet.cnrs.fr/spip/IMG/pdf/DDI_versionFR.pdf>

distribution, la découverte, l'analyse, la réorientation et l'archivage des données."[90]

Cette prise en compte des différentes étapes de vie des données de recherche s'accompagne de la multiplication des champs à décrire et des spécifications pour que chacun puisse le faire de façon relativement uniforme.

Voici le schéma du cycle de vie tel que le propose le site de la DDI Alliance :

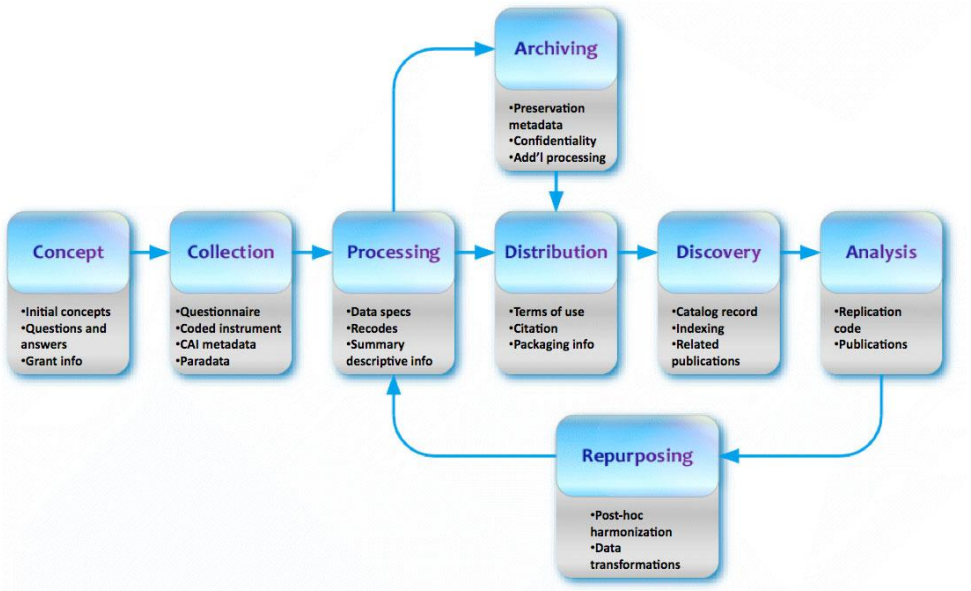


Figure 2 : Schéma du cycle de vie des données

Source : http://www.ddialliance.org/sites/default/files/what-is-ddi-diagram.jpg

Pour intégrer ce cycle de vie, il faut donc décrire et documenter chaque étape. Cela représente un travail considérable. C'est pourquoi peu de projets en 2014 utilisent les versions 3 et suivantes de DDI[91].

[90] *What is DDI?* Extrait du texte de présentation en anglais de la norme DDI sur le site DDI Alliance <http://www.ddialliance.org/what> [En ligne] Consulté le 02/06/2014.

[91] Il ne nous pas été possible de déterminer précisément le nombre de projets qui ont opté pour les versions les plus complètes de DDI. Pour des exemples d'utilisation, on peut lire la page du site DDI Alliance consacrée aux réalisations < http://www.ddialliance.org/ddi-at-work/projects> [En ligne] Consulté le 05/10/2014. En parcourant la liste des initiatives utilisant DDI, on voit néanmoins qu'elles ne sont pas majoritaires à user des versions 3 et suivantes. Nos entretiens et rencontres avec des membres de Quetelet confirment cette impression, mais des projets d'utilisation des versions récentes sont à l'étude.

Les 34 membres actuels de l'Alliance sont des institutions spécialisées de très nombreux pays, dont le Réseau Canadien des Centres de Données de Recherche (RCCDR) ou Statistique Canada, et le Réseau Quetelet en France.

Les partenaires de Quetelet travaillent tous avec la DDI, en général dans sa version 2.5. Le CDSP utilise également DDI pour la description des enquêtes, tant quantitatives que qualitatives. Mais pour ces dernières, l'équipe beQuali a créé sa propre grille de champs de description, en utilisant également d'autres standards, dont le Dublincore, afin de répondre aux besoins très particuliers du type de données traitées (ce sont par exemple des entretiens semi-directifs). Pour les données quantitatives, l'équipe projette en 2014 de passer à la version 3, qui permet la prise en compte du cycle de vie des données.

OpenAire+, initiative européenne d'accompagnement des politiques.
Quadrant sud-est

OpenAIRE* (*Open Access Infrastructure for Research in Europe*) est un projet européen dont le but est d'accompagner l'obligation partielle de dépôt en accès libre (*deposit mandates*) décidée par la Commission Européenne et le Conseil Européen de la Recherche (ERC).

La décision de la Commission Européenne de rendre obligatoire le dépôt pour 20 % des recherches financées par le 7e PCRD (c'est-à-dire le prédécesseur d'Horizon 2020) constitue une étape importante. Elle est venue renforcer l'obligation de dépôt rendue publique fin 2007 par l'*European Research Council* (ERC) pour les recherches qu'il finance. OpenAIRE doit accompagner concrètement la mise en œuvre de deux décisions :

- la publication, le 17 décembre 2007, de recommandations de l'ERC demandant la mise en accès libre des résultats des recherches financées par l'ERC au plus tard six mois après leur publication ;
- l'annonce à l'été 2008 par la CE de l'obligation de diffusion en accès libre dans un délai de 6 ou 12 mois des publications issues de 20 % des projets financés par le 7e PCRD. Baptisé Open Access Pilot, cette décision importante est un premier pas vers une obligation de dépôt plus large.

Afin d'accompagner ces deux décisions, la Commission a lancé en 2009 un appel d'offre dans le cadre du 7^{e} PCRD afin de mettre en œuvre « une infrastructure électronique et des mécanismes de soutien permettant l'identification, le dépôt, la consultation et la gestion des articles financés par l'ERC et le 7^{e} PCRD ». Le projet OpenAIRE a été retenu au mois de juillet 2009. La Commission y attachait une importance particulière car de son succès dépendait l'extension de l'obligation de dépôt à l'ensemble des recherches financées par le 8^{e} PCRD, c'est-à-dire le programme cadre Horizon 2020.

OpenAIRE regroupe 38 institutions représentant 26 des 27 pays de l'Union Européenne (manque le Luxembourg). Le projet a débuté fin 2009 pour une durée de 36 mois et a été prolongé par le projet openAIREplus* (ou OpenAire+) qui élargit son périmètre aux données de la recherche, jusque fin 2014. La Commission travaille à la pérennisation de son infrastructure[92].

OpenAIREplus utilise l'infrastructure du projet OpenAIRE qui donne la possibilité aux scientifiques de déposer en Open Access leurs publications et leurs résultats de recherches financés par la Commission Européenne en libre accès. Grâce à OpenAIREplus, les données de recherches peuvent dorénavant être archivées et liées avec les publications qui y sont associées. Les chercheurs peuvent déposer leurs données dans le répertoire *Open Access* de leur choix ou dans l'archive orpheline ZENODO, qui se trouve sur le site web d'OpenAIRE. Le *helpdesk* OpenAIRE et les NOADs (*National Open Access Desks*) sont disponibles pour assister et informer tous ceux qui s'intéressent au projet ou à *l'Open Access* en général.

L'ambition d'OpenAire comme d'OpenAire+ est de permettre l'interopérabilité des entrepôts (qu'ils soient institutionnels ou disciplinaires) et de devenir, en les moissonnant[93], le fédérateur, le « guichet unique » de l'accès aux données de la recherche.

On retrouve là une préoccupation importante de l'ensemble des actants des données de la recherche[94] que nous avons observé[95] : faciliter le dépôt de jeux de données, tout en liant ceux-ci aux publications auxquelles ils ont donné lieu.

Mais pour le moment, ces préoccupations semblent avoir abouti à une multiplicité d'entrepôts, de plateformes et autres projets. Devant ce foisonnement, une question demeure : comment choisir une méthode, un entrepôt, un intermédiaire, pour la gestion des données[96] ? OpenAire sera-t-il « le » portail des données de la recherche pour l'Europe de la recherche ? La porte qui guiderait chacun, depuis un pallier, vers l'entrepôt (ou la méthodologie, ou le jeu de données, ou l'interlocuteur) qui lui convient ? Actuellement, OpenAire ne donne en fait accès qu'aux données issues d'un projet de recherche financé par l'Europe et/ou reliées à une publication elle-même financée par l'Europe.

---

[92] Au 28/10/2014, nous n'avons pas trouvé d'information plus précise quant aux suites qui seront données à ce projet.

[93] Selon le protocole OAI-PMH*, et avec un schéma de métadonnées issu de celui de DataCite*.

[94] Préoccupation que nous trouverons de nouveaux sur le terrain du stage, comme l'évoque la troisième partie de ce mémoire.

[95] Pour obtenir une lecture plus complète du paysage des *Research data* en France, il faudrait poursuivre l'observation du côté des relations qu'entretiennent ces actants, qu'ils soient acteurs publics ou associatifs, avec les éditeurs scientifiques. Est-ce que tout passe par le consortium Couperin ? Quels sont les rapports de pouvoir en jeu ? Et comment les éditeurs vont-ils se positionner sur le long terme ? Les éditeurs de SHS ont-ils les mêmes positions que ceux des sciences dites « dures » ?

[96] A titre de comparaison, il serait intéressant de se tourner vers l'exemple anglais, plus clairement dans l'esprit de l'*Open data*, mais parfois beaucoup moins fouillé. Par exemple, l'initiative anglaise Qualidata présente des métadonnées moins nombreuses et plus frustes que celles de la base française beQuali, à qui elle a servi de modèle. Le rôle du Jisc et de son service DCC, en lien avec les bibliothèques universitaires, pourrait aussi être comparé avec les rôles des institutions françaises et européennes. A ce sujet, on peut lire le chapitre sur les cas américains, britanniques et canadiens p.29 à 31 du mémoire de R. Gaillard déjà cité.

TROISIEME PARTIE

III Une étude de cas : un projet de longue haleine à l'IFSTTAR

Pour illustrer notre propos, cette dernière partie se focalise sur l'exemple d'un projet en cours au sein d'un institut de recherche français : l'IFSTTAR. Elle s'appuie notamment :

- Sur des rencontres avec six chercheurs du site de Marne-la-Vallée de cet institut, afin de déterminer si leur participation était possible. Lors de ces premières rencontres, préparatoires, nous leur avons demandé s'ils souhaitaient et pouvaient nous confier un jeu de données, et s'ils jugeaient que ce projet pouvait être utile pour eux. Ces six entretiens se sont déroulés entre le 6 juin et le 15 juillet 2014.
- Dans un second temps, nous avons donc mis en place, avec les chercheurs ayant accepté (et présents durant l'été) quatre séances de tests pour valider deux modèles ou « templates » destinés à formaliser la description des jeux de donnes et leur documentation dans l'outil logiciel Dataverse. Ces quatre entretiens ont eu lieu entre le 29 juillet et le 2 septembre 2014.

3.1 L'IFSTTAR, un institut de recherche finalisée

L'Institut français des sciences et technologies des transports, de l'aménagement et des réseaux (IFSTTAR) est issu d'une fusion réalisée en 2011 : celle du Laboratoire Central des Ponts et Chaussées (LCPC) et de l'Institut National de Recherche sur les Transports et leur sécurité (INRETS).

C'est - en France, l'institut de recherche finalisée sur les questions liées aux transports (aménagement, réseaux, technologies des transports) et aux matériaux des infrastructures de transport. Il s'organise en plusieurs sites en France, dont le plus important en nombre d'employés se trouve depuis un an à Marne-la-Vallée.

Il n'existe pas d'autre établissement de recherche en France avec ce périmètre, transverse quant aux disciplines[97], et unifié autour des questions relatives au transport.

[97] Organisées en cinq départements, les recherches menées au sein de l'institut font appel, notamment, aux disciplines suivantes : sciences des matériaux (physico-chimie), mécanique des solides, mécanique des structures, sciences de l'ingénieur, pour le département Matériaux et Structures ; géotechnique et sciences de la terre, pour le département Géotechnique, Environnement, Risques naturels et Sciences de la terre ; modélisation, calcul scientifique, physique numérique, informatique et génie logiciel, automatique, traitement de l'information, électronique, pour le département Composants et Systèmes ; biomécanique, épidémiologie, ergonomie, sciences cognitives, mécanique des chocs, mécanismes d'accidents, simulation, pour le département Transport, Santé, Sécurité ; sociologie, économie, sciences de l'ingénieur, informatique scientifique, anthropologie, géographie, aménagement, informatique et

Des données hétérogènes :

L'Institut a pour mission de réaliser ou faire réaliser et d'évaluer des recherches, des développements et innovations dans les domaines du génie urbain, du génie civil et des matériaux de construction, des risques naturels, de la mobilité des personnes et des biens, des systèmes et des moyens de transport et de leur sécurité, des infrastructures, de leurs usages et de leurs impacts, considérés des points de vue technique, économique, social, sanitaire, énergétique, environnemental et humain. Les données travaillées à l'IFFSTTAR sont donc issues de nombreuses disciplines et de nature extrêmement variée.

3.2 Le projet réseau GEBD, Belgrand…et ses données : un questionnement de départ et des évolutions

Nommé parfois GEBD (pour Grand équipement base de données), réseau-GEBD, Equipement Belgrand, Belgrand ou réseau Belgrand, GEBD-Belgrand et enfin Grand équipement Belgrand…A l'image de l'organisation de la recherche en France, celle du projet Belgrand est parfois difficile à appréhender. Les contenus et les objectifs ont et continuent d'évoluer au gré des questionnements et des avancées technologiques. Né en 2008, le projet Réseau-GEBD (http://belgrand-gebd.ifsttar.fr) vise à faciliter l'accès, l'usage et la valorisation des bases de données pour la recherche et l'expertise sur la ville. Ce projet est porté par les laboratoires Ville Mobilité Transport (LVMT) et Dynamiques économiques et sociales des transports (DEST) et fait appel à de nombreux partenaires, selon les lots[98].
Le premier lot ou lot A correspond à l'objectif de mise en place et d'animation d'un réseau scientifique disposant d'une infrastructure informatique d'échange, menant des actions d'audit de l'accès aux données, de dissémination d'information scientifique et de réflexion sur la formation aux systèmes d'information sur la ville. Le second lot, ou lot B, doit permettre le développement et la diffusion de méthodes partagées d'analyse de référentiels évolutifs et la reconstitution chronologique de référentiels, à partir de l'apport méthodologique des projets de recherche des partenaires et de l'expertise en outils et méthodes SIG (pour Système d'Information Géographique).

En effet, une grande partie de la valeur des données réside dans la compétence acquise par ceux qui les utilisent et éventuellement les transforment pour résoudre des problèmes de recherche. Il est important à la fois pour les chercheurs, pour l'IFSTTAR et pour ses partenaires de rendre visible, de diffuser cette expertise, tout en mutualisant les moyens pour la développer.

---

mathématiques appliquées, acoustique environnementale et psychologie des comportements pour le département Aménagement, Mobilité, Environnement.

[98] Au démarrage du projet, le premier lot impliquait l'INRETS et le LCPC, l'IGN, le Lab-Urba, et des laboratoires de l'ENPC ayant manifesté leur intention de participer à l'animation du réseau. Le second lot impliquait l'IGN, l'INRETS (DEST et LVMT), le LCPC et le CIRED (laboratoire ENPC-EHESS). Depuis 2011 et la fusion INRETS-LCPC, c'est donc l'IFSTTAR qui est impliqué.

Le projet Belgrand dans son ensemble concerne les informations relatives à la ville, notamment topographiques et géolocalisées (données de l'IGN, du Vélib...). L'un des lots du projet concerne plus particulièrement les données de la recherche utilisées par l'IFSTTAR[99] ou ses partenaires[100] : le lot A est décrit comme devant permettre la création et l'animation d'un réseau scientifique sur les données, leur accès et leur exploitation. C'est au sein de ce lot que le stage réalisé dans le cadre de la formation INTD devait apporter des connaissances en ingénierie documentaire.

Les chercheurs des laboratoires participants aux premiers volets du projet Réseau-Belgrand travaillent essentiellement sur des données statistiques, sur les trajets et déplacements, la mobilité et la ville ; ces données peuvent être issues de l'INSEE, comme les « grandes enquêtes transports ». Il s'agit également de données agrégées par les chercheurs de l'IFSTTAR, qui réalisent des études sociologiques et économétriques, des programmes de traitement de données, et des data-visualisations. Nous avons donc affaire à des données majoritairement statistiques et issues d'enquêtes ou d'études sur les transports, et à quelques programmes informatiques. Parmi les chercheurs, on retrouve donc des sociologues, des géographes, des statisticiens, des économistes et des informaticiens.

Mais d'autre part, certains laboratoires de l'IFSTTAR produisent des données sur la résistance des matériaux. Elles sont issues d'essais réalisés par des scientifiques des matériaux. L'exploitation de ces données à été intégrée au projet au cours du stage, les questions soulevées par le stage ont réactivé le réseau des chercheurs qui portent Belgrand. En effet, Jean-Paul Hubert, géographe, membre du laboratoire DEST et l'un des porteurs de GEBD, a diffusé oralement l'information auprès d'autres laboratoires et départements de l'IFSTTAR afin de trouver des chercheurs qui acceptent de coopérer et de nous confier quelques jeux de données. C'est donc de manière très informelle que ce pan du projet avance (et pose de nouvelles questions).

A l'été 2014, la phase préparatoire du projet (2008-2010) mais aussi les premières réalisations (un site Internet[101], le cycle de séminaires…) sont abouties. Notre intervention concerne en quelque sorte une phase de test pour un « projet dans le projet », une idée qui est née au long des phases précédentes : et si les laboratoires impliqués automatisaient et élargissaient le partage de leurs données, celles sur lesquelles ils ont des compétences et qu'ils travaillent pour leurs études ? En effet, les nouvelles directives du MESR comme le nouvel agencement géographique du site de Marne-la-Vallée tendent à rapprocher l'institut de ses partenaires : l'Ecole des Ponts ParisTech, l'Ecole Nationale des Sciences Géographiques (ENSG, membre de l'Université Paris-Est). Réunies en PRES puis au sein de COMUE, ces institutions sont incitées à travailler davantage en réseau, afin de dynamiser la recherche malgré des budgets orientés à la baisse. Il s'agit également de montrer aux évaluateurs de la recherche que les différents laboratoires impliqués ont des compétences d'ingénierie des données, que ce travail est important et doit être valorisé et reconnu. Il est

[99] L'IFSTTAR produit peu de données en SHS, mais en utilise beaucoup. Voir le rapport 2014 sur le projet Belgrand.
[100] Entre autres, l'IGN, impliqué dans l'ensemble des lots du projet.
[101] Disponible en ligne <http://belgrand-gebd.ifsttar.fr> Consulté le 8/06/2014.

chronophage mais indispensable pour nourrir les publications, même s'il peut en ralentir le rythme.

Plutôt que de partager des outils de type BDD, ou une BDD de BDD (ce qui était envisagé au commencement du projet), pourquoi ne pas se tourner vers l'*Open access* et partager les données dans un outil libre ?

Cette préoccupation de valorisation rejoint également un constat quasi quotidien des chercheurs de l'IFSTTAR : il est parfois long et compliqué de se procurer des données, y compris lorsqu'elles « appartiennent » à l'institution au sein de laquelle ils travaillent, ou lorsqu'ils ont déjà eu accès à ces données pour de précédentes recherches.

Les injonctions politiques, les restrictions budgétaires, l'existence d'un projet financé et concernant les données, le besoin d'efficacité et d'une visibilité nécessaire à la valorisation d'un travail souvent invisible : voilà donc les moteurs qui ont poussés à la naissance de cette partie du projet Réseau-GEBD.

Enfin, Jean-Paul Hubert et Olivier Bonin, les co-pilotes du projet, ont rapidement identifié le besoin de s'adjoindre des compétences sur les questions liées aux métadonnées et aux formats de description des jeux de données. Ils se sont très naturellement tournés vers le service Documentation Multimédia scientifique et technique (DMST), et c'est Alain Drouet, responsable délégué du service pour le site de Marne la Vallée, qui a proposé le recrutement d'un stagiaire de l'INTD. On voit ici l'importance d'une culture de l'IST au sein des établissements de recherche. C'est cette connaissance des compétences du service de documentation qui a permis aux chercheurs d'identifier rapidement et précisément leur besoin.

3.3 Politiques de gestion de données, DMP et logiques à l'œuvre :

En l'état actuel, il n'y a pas de prise en compte spécifique des données de recherche à l'IFSTTAR. Il n'existe pas –encore- de politique de gestion définie, ni à l'étude au niveau de la direction scientifique de l'établissement. En fait, le projet GEBD semble bien être le premier[102], à l'IFSTTAR, à se préoccuper de définir une gestion de données.

Or ce projet émane de chercheurs, et les évolutions des objectifs et étapes de GEBD Belgrand sont la traduction d'un cheminement qui s'est effectué au sein d'un réseau de chercheurs et de professionnels de l'IST de différentes institutions[103]. Cette logique « bottom up » semble plutôt atypique pour la mise en place d'une gestion de données

[102] Nous n'avons pas réalisé de véritable enquête à ce sujet, mais avons pu parler avec de nombreux chercheurs, notamment au restaurant d'entreprise ou à la cafétéria. Et bien sûr, nous avons interrogé le responsable délégué du service DMST : aucun projet n'était à l'étude concernant un management des données avant GEBD.

[103] Au cours d'échanges avec des chercheurs de l'IGN et d'autres partenaires, et avec les professionnels de l'ADISP, membre du réseau Quetelet.

de la recherche ; dans la plupart des études de cas que nous avons lues[104], c'est un fonctionnement « top down » que nous avons noté.

A l'IFSTTAR, et malgré des politiques volontaristes exprimées par le MESR en 2013[105] et l'Europe ces dernières années[106], c'est de chercheurs et non de la direction que vient ce premier pas vers un management des *Research data*.

Dans la littérature que nous avons consulté au sujet des données de la recherche, l'importance des Data Management Plan (DMP), c'est-à-dire de plan de gestion des données, est mainte fois soulignée. Rémi Gaillard tire même comme bilan de son étude qu'il s'agit de l'outil incontournable pour impulser une prise en compte des *Research data* au sein d'un établissement. Ces DMP sont des documents établis en amont d'un projet de recherche par les chercheurs, et qui décrivent la planification de la gestion envisagée des données qui seront produites pour ce projet. Ainsi R. Gaillard[107] note :

> *« Le plan de gestion constitue le pendant, concret, d'une politique de données, puisqu'il est censé expliciter la manière dont le chercheur entend se conformer à la Data policy de son agence de financement ou de son université et garantir, à terme, la diffusion des données produites. Il est donc un outil indispensable à l'ouverture. »*

C'est ce document[108] qui, prenant en compte les types de données qui seront produites, leur cycle de vie, les questions de droit, les problèmes de formats informatiques, leur valeur et leur degré de « réutilisabilité », permettra de fixer d'éventuelles durées d'embargo, durant lesquelles les données ne seront utilisables que par ceux qui les ont recueillies et fabriquées, ou par un partenaire privé qui aurait financé le projet, afin de permettre la publication des résultats de la recherche ; puis fixera un cadre pour une diffusion qui permettra ou non un accès ouvert et les durées de conservation et d'archivage ad hoc. Il est donc aisé de comprendre l'intérêt, pour une bonne gestion des données, de ce document qui permet d'anticiper et de planifier, en se posant les bonnes questions, ce que l'on fera de chaque type de données récoltées au cours d'une recherche.

---

[104] Comme par exemple les nombreux cas, européens et nord-américains, cités par R. Gaillard dans son mémoire, mais aussi ceux exposées dans le numéro 73 d'Arabesque, ou les études anglaises évaluant l'impact de projets de *Data Centers*…

[105] Voir le discours de la secrétaire d'Etat à la recherche, G. Fioraso, cité plus haut.

[106] Avec le programme cadre H2020, par exemple, ou le dispositif OpenAire+.

[107] GAILLARD Rémi, op. cit. p.38.

[108] Il existe de nombreux modèles de DMP, la plupart en anglais. Le DCC anglais fournit une très pratique « checklist » qui liste les questions à se poser pour établir un plan de gestion des données : DCC, 2013. Checklist for a Data Management Plan.v.4.0. Edinburgh: *Digital Curation Centre*, et un outil en ligne, de type formulaire à remplir. Ces ressources sont accessibles, avec d'autres, à cette adresse : <http://www.dcc.ac.uk/resources/data-management-plans> [En ligne] Consulté le 05/06/2014. On trouve une trame de DMP en français, adaptée de la checklist du DCC , sur le site CoopIST, dans le dossier d'initiation à la gestion des données de recherche déjà cité <http://coop-ist.cirad.fr/gestion-de-l-information/gestion-des-donnees-de-la-recherche/decouvrir-des-plans-de-gestion-de-donnees-de-la-recherche/3-exemple-de-trame-d-un-plan-de-gestion-de-donnees-pgd> [En ligne] Consulté le 27/11/2014. Ce document est reproduit en annexe 2.

Mais dans notre cas d'étude, que faire de cette recommandation ? Comme le note bien R. Gaillard, il s'agit de créer un document en accord avec une politique d'établissement, qui, à l'IFSTTAR, n'est pas encore définie.

D'autre part, les DMP concernent plutôt des données non-encore existantes. On y prévoit une gestion, en fonction de ce que l'on peut anticiper, des données qui seront à réunir, recueillir ou à construire[109] lors d'une recherche. Mais que faire des données déjà utilisées/recueillies/agrégées, dormant dans les disques durs d'ordinateurs individuels, voir sur des supports de stockage aux formats obsolètes ? Sans aller jusqu'à envisager une complète reprise de l'existant, doit on simplement abandonner ces données ?

Ces difficultés signifient elles que le projet GEBD prend la question des données à l'envers ? Faut-il nécessairement qu'une logique descendante soit à l'œuvre pour qu'une politique de gestion des données de recherche démarre ?

Notons également une difficulté qui vient du fait même que les DMP sont des réponses que doivent faire les laboratoires à une injonction émanant d'une direction : les chercheurs sont-ils bien disposés face à ce qu'ils peuvent vivre comme une intrusion dans leur mode de fonctionnement ? Et ont-ils le temps de réaliser avec soins ces documents ? N'y a-t-il pas un risque qu'ils perçoivent la question des données comme une lourdeur administrative supplémentaire ?

Nous n'avons pas de réponse définitive à ce sujet, mais nous avons décidé de prendre ce risque en compte en acceptant de travailler le projet GEBD là où il en était, avec sa logique « Bottom up ». Les chercheurs portant ce projet ont déjà beaucoup avancé, entre autre en optant pour un outil de stockage et de diffusion en ligne et en réseau, créé par des chercheurs pour des chercheurs. Nous faisons le pari que ce mode de fonctionnement, peu orthodoxe (pas de cahier des charges avant de choisir un outil, pas de politique définie au niveau de la direction, pas de plan de management des données…), peut se révéler une force, au moins dans un premier temps.

- Cela permet de s'appuyer sur le réseau des chercheurs qui mènent ce projet, et donc de convaincre d'autres chercheurs de l'établissement de l'utilité du projet, plus aisément que si il s'agissait d'une directive hiérarchique.
- C'est également un bon moyen de tester l'intérêt et la viabilité d'un projet de gestion des données à l'IFSTTAR.
- Même sans être exhaustif, c'est une façon de faire un premier point sur le type de données de recherche présentes au sein de l'établissement, et des besoins des chercheurs.
- Si les conclusions devaient être qu'il serait plus intéressant d'entreposer ces données dans un entrepôt disciplinaire (ou dédié à la longue traine des données), cela aura permis de se poser l'ensemble des questions préalables et d'étudier les différentes possibilités qui se présentent (par exemple, les

[109] Voir l'animation de la *NYU Health Sciences Library*, qui figure en fin de présentation canadienne de Nathalie Clairoux.

nombreux entrepôts existants, mais aussi les services développés pour aider les laboratoires à gérer leurs données).

- Enfin, une modeste réalisation peut aider à argumenter en faveur de la définition d'une véritable politique de gestion des données, à l'échelle de l'établissement.

3.4 Open data ou accès restreint ?

Quel type d'accès choisir, lorsqu'une institution souhaite permettre une diffusion de ses données de recherche ? Comme nous l'avons vu en première partie, cela dépend essentiellement du public auquel on souhaite s'adresser. Partager ce que certains voient comme un « trésor de guerre » ou encore le « nouveau pétrole », certes, mais avec qui ? Que veut-on favoriser avec ce partage ?

Cette question revêt à l'IFSTTAR une importance pratique. En effet, tandis que Jean-Paul Hubert et Olivier Bonin, porteurs du projet GEBD-Belgrand, penchent pour un accès restreint aux jeux de données mais avec une description visible par tous, s'inspirant de ce qu'ils utilisent avec le Réseau Quetelet, un autre « projet Open Data » a vu le jour dans l'établissement. Plus classiquement « Open », il s'agit d'une initiative individuelle d'un chercheur très intéressé par les possibilités offertes par les TIC et le mouvement d'ouverture des données publiques qu'opère le gouvernement français depuis quelques années.

L'idée est de conclure un accord avec Etalab[110] pour déposer des jeux de données des chercheurs de l'institut sur la plateforme de diffusion de données publiques françaises data.gouv.fr, et d'assurer ainsi une certaine visibilité de l'IFSTTAR. En effet, l'institut souffre d'un déficit de reconnaissance, y compris au sein de la communauté de la recherche française, depuis la fusion en 2011 du LCPC et de l'INRETS et l'attribution d'un nouveau nom pour cette jeune entité. Une diffusion sur une plateforme gouvernementale et dont l'objectif vise un large public pourrait donc contribuer à assurer à l'établissement une meilleure exposition.

Nous sommes donc devant deux modèles de diffusion des données très différents, y compris en terme d'objectifs. Le réseau Quetelet travaille sous l'égide du Comité de Concertation pour les Données en SHS (CCDSHS), créé en 2001, et diffuse depuis plus de dix ans des données de recherche pour les chercheurs ; la plateforme data.gouv.fr

[110] La politique d'ouverture et de partage des données publiques (« Open data ») est pilotée, sous l'autorité du Premier ministre, par la mission Etalab, dirigée par M. Henri Verdier. La mission Etalab fait partie du Secrétariat général pour la modernisation de l'action publique, dont la direction est assurée par Mme Laure de La Bretèche. Etalab administre le portail unique interministériel data.gouv.fr destiné à rassembler et à mettre à disposition librement l'ensemble des informations publiques de l'Etat, de ses établissements publics et, si elles le souhaitent, des collectivités territoriales et des personnes de droit public ou de droit privé chargées d'une mission de service public. Treize personnes, directeur compris, sont recensées dans la présentation de l'équipe des employés de la mission Etalab. <https://www.etalab.gouv.fr/lequipe> [En ligne] Consulté le 02/10/2014.

réalisée par Etalab sous l'autorité du Premier Ministre est en ligne depuis le 5 janvier 2011 pour assurer la diffusion en libre accès des données publiques.

Afin de comprendre les fonctionnalités de ces deux modèles que représentent le catalogue du réseau Quetelet et celui du site open.data.gouv, voici une brève comparaison de ces deux « modèles » sous forme de tableau.

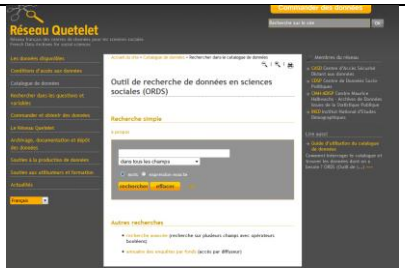
Responsabilité Site Web	**Réseau Quetelet** http://www.reseau-quetelet.cnrs.fr	**Etalab** https://www.data.gouv.fr/fr/
Page de recherche		
Périmètre couvert	Données en sciences humaines et sociales (voir la cartographie en annexe pour plus de précision)	Tout type de données publiques, que les institutions productrices ont bien voulu déposer.
Accès aux jeux de données	Open access, avec accès restreint à la communauté des chercheurs, et sur justification d'un projet de recherche.	Open data : accès libre.
Accès aux métadonnées	Libre	Libre
Documentation, description, modèles de métadonnées	DDI et adaptation de celle-ci, avec éventuellement utilisation de modèles complémentaires. Pour certaines enquêtes, la description va jusqu'à la variable. Ce travail est effectué par des chercheurs et professionnels de l'IST et des statistiques.	Pas de modèle de référence. Titre et nom du déposant obligatoires, puis description libre, effectuée par le déposant du jeu de données. Un guide de publication recommande d'aller plus loin dans la description.
Recherche	2 modes de recherche : simple (une seule cartouche) ou avancée (dans quatre champs et avec opération booléenne).	2 modes de recherche : par catégorie parmi les 9 définies, ou avec le moteur de recherche.
Objectifs, publics et utilisations ciblées	Stimuler la recherche. Réutilisation des données par des chercheurs pour de nouvelles recherches.	Transparence démocratique et stimulation économique. Réutilisation des données par des développeurs individuels et des entreprises pour créer des applications grand public.

Figure 3 : Tableau comparatif des catalogues Quetelet/data.gouv.fr

Lors d'une recherche simple avec le terme « transports » dans ces deux outils, voici un bref comparatif des résultats.

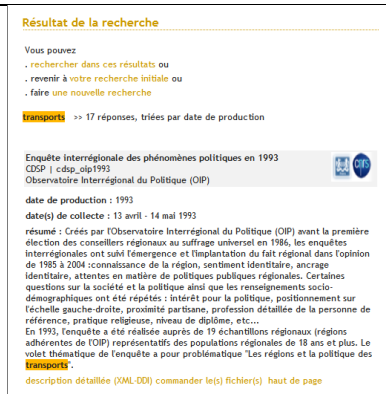
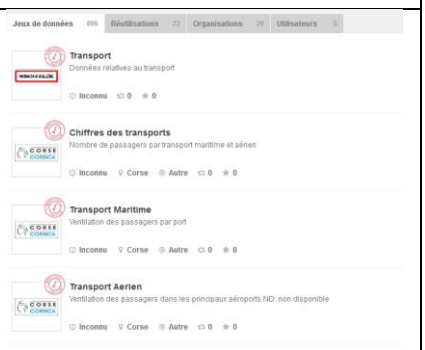
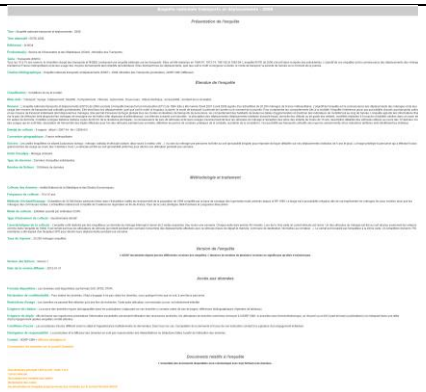
Outil	**Quetelet**	**Data.gouv.fr**
Nombre de résultats retournés	17	895
Affichage des résultats		
Exemple de page de description détaillée d'un item (issu de la première page de résultats)		
Contenu de la description détaillée	32 champs de description de l'enqête, en 5 grands aspects : Présentation de l'enquête, 6 champs Etendue de l'enquête, 9 champs Méthodologie et traitement, 7 champs Version de l'enquête, 2 champs Accès aux données, 8 champs Plus une liste des documents relatifs à l'enquête, directement accessibles.	5 champs de decription : Titre Date de publication Fournisseur Résumé Liste des tableau disponibles Plus les liens directs vers l'accès aux jeux de données.

Figure 4 : Tableau comparatif des résultats de requête simple dans Quetelet/data.gouv.fr

L'intérêt principal de la plateforme Data.gouv.fr réside dans l'accès direct et immédiat aux jeux de données. Alors que celui de Quetelet se trouve dans la minutieuse description et la documentation très complète des jeux de données. Cela correspond au fait que les objectifs poursuivis par les deux outils sont très différents.

L'aspect recherche de chacun des deux catalogues est à l'avenant : celle de Quetelet est assez fine, et renvoie à une liste de résultats ciblés, parfois peu nombreux et frôlant le silence. Celle de data.gouv.fr retourne par contre une masse de résultats (allant jusqu'au bruit) sans aucune possibilité de tri parmi ceux-ci. Comme peu de champs sont renseignés, une recherche peu tourner court. D'autre part, lorsque l'on met le nom d'un déposant de *datasets* ou de réutilisation de données dans le cartouche de recherche, le moteur ne renvoie pas les résultats escomptés : manifestement, le champ « nom du déposant » n'est pas pris en compte, ou alors à la marge. Les données recherchées figurent peut être quelque part, parmi les centaines de résultats...

Chacun des deux outils a une orientation très nette. Celui d'Etalab favorise une sorte de « flânerie » parmi les *datasets* en ligne. Comme il n'existe pas de description normalisée, il faut fouiller, ouvrir les fichiers (et pour cela, posséder les logiciels qui le permettent, car de très nombreux formats spécialisés cohabitent dans le catalogue) pour comprendre ce qui s'y trouve. Les résumés –quand il y en a- sont parfois elliptiques. Le public visé n'est pas tant le grand public que la communauté des « geek »[111], à l'aise avec les TIC au point de savoir coder dans plusieurs langages informatiques et de différencier des formats très spécifiques (par exemple, les format gtfs[112], json[113], gzip, xml, xls, tsv, ods, kml, shp…tous rencontrés avec cette simple requête sur le terme « transports » et en ouvrant les cinq premiers résultats).

L'outil de Quetelet se tourne, explicitement, vers le monde de la recherche, au point que quelqu'un qui ne peut justifier d'un projet de recherche ne peut avoir accès aux données. Mais la recherche est facilitée par l'énorme travail de description et de documentation des jeux de données, qui sont regroupés par enquête. L'utilisation d'une norme internationale, la DDI, connue et reconnue dans la communauté des sociologues, statisticiens et économistes du monde entier, facilite une recherche pointue et orientée vers l'obtention de données à réutiliser pour des études et

[111] D'après le Larousse en ligne : Fan d'informatique, de science-fiction, de jeux vidéo, etc., toujours à l'affût des nouveautés et des améliorations à apporter aux technologies numériques.
[112] General Transit Feed Specification (GTFS, traduction littérale : spécification générale pour les flux relatifs aux transports en commun) est un format informatique standardisé pour communiquer des horaires de transports en commun et les informations géographiques associées (topographie d'un réseau : emplacement des arrêts, tracé des lignes). Source : Wikipedia
[113] JSON (*JavaScript Object Notation*) est un format de données textuelles dérivé de la notation des objets du langage JavaScript. Il permet de représenter de l'information structurée comme le permet XML par exemple. Source : Wikipedia. On peut trouver sur cette encyclopédie collaborative des articles courts sur l'ensemble de ces formats spécialisés.

recherches dans ces domaines. D'ailleurs, l'outil utilisé pour diffuser ces données est moissonné par ISIDORE[114], la plateforme de recherche et d'accès aux données de SHS.

Certains chercheurs de l'IFSTTAR penchent plutôt vers le projet *Open data* et la plateforme d'Etalab, et nous formulons l'hypothèse que cela peut s'expliquer par leur profil de type « geek » (jeune et œuvrant dans une discipline liée à l'informatique, avec un goût prononcé pour une forme de « flânerie » sur le web et la sérendipité). Par exemple, parmi les six chercheurs que nous avons rencontrés pour des entretiens puis pour trois d'entre eux, pour des tests de description d'un *dataset (voir 3.5)*, l'un d'entre eux a manifesté cette préférence. C'est un jeune chercheur[115] en technologie de l'information et des systèmes. Il est le seul des six à avoir ce profil et à opter clairement pour *l'Open data*.

La demande d'Olivier Bonin et Jean-Paul Hubert, les chercheurs porteurs de GEBD-Belgrand, se situe clairement du côté d'un accès contrôlé, et ce pour plusieurs raisons :

- L'un des moteurs du projet d'ouverture est l'idée de permettre la citation des jeux de données (re)travaillées par des chercheurs de l'institut, afin que ce travail sur les données gagne en visibilité et éventuellement, soit pris en compte lors des évaluations des financeurs. Dans ce cadre, la plateforme data.gouv.fr n'offre pas d'intérêt, puisqu'elle se situe hors du monde de la recherche, et ne propose pas de référence pré-écrite pour la citation des jeux de données.
- Les chercheurs de l'IFSTTAR utilisent beaucoup de *datasets* issus d'autres institutions. La question des droits de diffusion de données retravaillées les met mal à l'aise s'il s'agit d'une diffusion sans restriction ni contrôle. Alors que si cela reste au sein de la communauté de la recherche, et sur demande, un contact peut être établi avec chaque demandeur afin de savoir quel type d'usage il compte faire des données. C'est également l'occasion d'insister sur l'importance de la citation complète des « auteurs », des premiers créateurs des données à l'ensemble de ceux qui les ont agrégées, compulsées, nettoyées...
- D'autre part, ce contrôle et ce lien entre chercheurs comme « passage obligé » pour accéder aux données renforce le fait que la recherche est une activité qui ne saurait se passer de réseau. C'est un aspect auquel les initiateurs de GEBD-Belgrand tiennent particulièrement.
- Au vu du type de réutilisations qu'ils souhaitent favoriser, les chercheurs ont besoin que l'aspect descriptif de leurs jeux de données soit relativement complet. Sans aller jusqu'au travail très poussé effectué par les partenaires de Quetelet, ils ont besoin d'une documentation qui réponde aux normes internationales en vigueur au sein de leurs disciplines.

[114] Cela concerne pour le moment le corpus des enquêtes de l'ADISP. <http://www.reseau-quetelet.cnrs.fr/spip/breve.php3?id_breve=122&var_recherche=isidore> [En ligne] Consulté le 02/10/2014.

[115] Ce chercheur met d'ailleurs déjà des réutilisations de données en libre accès sur data.gouv.fr <https://www.data.gouv.fr/fr/users/etienne-come/reuses/> [En ligne] Consulté le 15/11/2014.

Mais parmi les différences entre les deux outils, il en est une qui pèse en faveur de la plateforme mise en ligne par Etalab : celle-ci demande moins de travail préparatoire avant un dépôt, moins de (voir aucun) personnel qualifié pour décrire les données, permet donc une mise en route du projet très rapide et, en un mot, ne nécessite pas de dégager un budget spécifique –ou du temps de travail.

3.5 Des métadonnées pour décrire les données : comment choisir une norme ?

Toutes les recommandations pour les projets d'ouverture de *Research data*, qu'elles émanent des autorités européennes, du JISC, du gouvernement français, ou des institutions de l'IST en France, préconisent de rechercher le ou les formats de description de données qui, dans une discipline ou un champ de recherche, font autorité au niveau international. Cela doit permettre les échanges et les réutilisations de données. Car si chacun partage la même grille de description et s'entend sur ce que l'on met comme information dans un champ, alors - et alors seulement - peut-on trouver et réutiliser des données en sachant bien ce qui s'y trouve, ce qu'elles contiennent, y compris en terme de mode de recueil, de méthode de traitement, de format informatique et de droits. La description et la documentation d'un *dataset* constituent en quelque sorte son mode d'emploi, sans lequel les données sont quasi introuvables et inutilisables.

Mais cela permet aussi d'envisager qu'un entrepôt de données institutionnel ou disciplinaire très spécifique soit moissonné – selon la terminologie en vigueur - par d'autres, par des plateformes internationales ou nationales, regroupant de nombreuses sources de données, éventuellement multi-disciplinaires, et qui sont des portes d'accès mutualisées, à large visibilité.

Pour un professionnel de l'information, l'une des premières tâches lors de la construction d'un projet d'ouverture de données de recherche est donc d'identifier les normes de description et de documentation ad hoc.

DDI, une norme pour les données d'enquêtes en SHS

Sur notre terrain, l'identification d'une norme se trouve, au moins au départ, grandement facilitée par le fait que le projet GEBD-Belgrand concerne des données de recherches quantitatives de SHS. Les chercheurs qui coordonnent ce projet connaissent bien les interlocuteurs incontournables pour ces données : il s'agit, en France, du Réseau Quetelet, dont nous avons décrit le fonctionnement en seconde partie, et dont nous avons rencontrés deux des partenaires (l'ADISP et le CDSP).

Sur le site du réseau, se trouvent de nombreuses informations et bases méthodologiques pour faciliter et favoriser le partage des données en SHS. Et notamment, sur la question des métadonnées, une brève présentation du standard DDI :

> « *Le développement d'Internet a conduit l'ensemble des archives de données pour les sciences sociales à élaborer des techniques adaptées pour donner accès à l'information sur les données disponibles afin de faciliter l'accès aux données elles-mêmes. En 1995, un groupe de travail constitué dans le cadre de IASSIST (International Association for Social Science Information Service & Technology) impulsé par l'ICPSR et connu sous l'étiquette "Data Documentation Initiative" ou DDI a entrepris d'utiliser le langage XML pour répondre aux besoins de métadonnées nécessaires à la conservation de l'information sur les fichiers, à l'échange de données entre institutions selon des champs communs et permettre une "recherche intelligente" sur le web.* »[116]

Nous avons donc observé le travail réalisé par ce réseau et chacun de ses partenaires, et vérifié que la norme DDI est celle qui convient aux données que GEBD-Belgrand doit permettre de partager. Car il ne suffit pas qu'elle soit reconnue et utilisée par les acteurs de la statistique publique[117], ou ceux des enquêtes en sciences sociales au niveau international[118]. Encore faut-il qu'elle convienne aux données dont le projet s'occupe (et dont il n'existe pas d'inventaire[119]). Chaque organisme adapte cette norme, en fonction des données à traiter, en sélectionnant parmi les champs qu'elle propose, ceux qui ont du sens dans leur cadre. Cela explique le volume important du *Codebook* : on y trouve une liste très impressionnante de champs, mais il n'est pas question que tout utilisateur utilise obligatoirement l'ensemble de ces champs.

Et c'est pourquoi nous avons conçu des tests : afin de vérifier l'adéquation entre les champs proposés par DDI et les jeux de données présents (et susceptibles d'être partagés sans souci juridique) chez les quelques chercheurs ayant accepté de participer à ce pré-projet.

En nous inspirant de ce qui existe au sein du réseau Quetelet[120], des propos et descriptions des chercheurs interrogés sur leurs données au cours d'une première série d'entretiens, et de ce qui se fait déjà à l'IFSTTAR pour les publications et leur dépôt dans l'archive ouverte HAL[121], nous avons élaborer une première version d'un questionnaire destiné aux chercheurs déposant des données. Celui-ci était volontairement très large, comprenant donc un maximum de champs issus de DDI-

---

[116] Page « Documenter les données » du site du réseau Quetelet <http://www.reseau-quetelet.cnrs.fr/spip/article.php3?id_article=140> [En ligne] Consulté le 02/06/2014.

[117] Par exemple, son utilisation est valorisée dès 2003 par l'INSEE dans sa publication (désormais arrêtée) *Le courrier des statistiques*. RIANDEY Benoît. Centre Quetelet, deuxième anniversaire in *Courrier des statistiques* n° 107, septembre 2003.

[118] La liste complète est disponible sur le site officiel de la *DDI Alliance* <http://www.ddialliance.org/> [En ligne] Consulté le 10/06/2014.

[119] Et pour cause : il faut encore convaincre les chercheurs de confier des données pour alimenter le projet, puisqu'il ne s'agit pas d'une obligation. La liste des jeux de données n'est donc pas arrêtée ; elle démarre tout juste.

[120] Notamment le guide du déposant du CDSP, disponible en ligne <http://cdsp.sciences-po.fr/fichiers_guide/guideFR.pdf> Ce document traduit des champs DDI en langage compréhensible par les chercheurs, et permet donc de demander à ces derniers, qui seuls connaissent bien le contenu et les méthodes concernées dans leurs *datasets*, de décrire ceux-ci afin que les professionnels de l'IST puissent ensuite « cataloguer » ces données dans un outil informatique.

[121] Voir en annexe 3 la note de procédure rédigée par le service DMST pour le dépôt des publications dans le catalogue de l'IFSTTAR et dans HAL.

codebook[122], puisqu'il s'agissait plutôt de le restreindre ensuite, en fonction des commentaires et besoins exprimés par les chercheurs auprès desquels nous avons fait le test.

C'est cette version[123] d'un guide du déposant de jeu de données que trois chercheurs ont testée. Nous leur avons demandé de choisir un jeu de données qu'ils souhaitaient partager, et les avons observés pendant qu'ils remplissaient le document comme s'il s'agissait réellement de le déposer auprès d'un professionnel de l'IST. Nous avons noté aussi bien leurs réflexions, leurs questions, que les réponses qu'ils notaient dans le guide, en intervenant le moins possible afin de ne pas fausser les résultats.

Un tableau[124] comparant leur réponses et doutes nous a permis de forger la version finale de ce guide[125], en analysant la pertinence de leurs propositions (de suppression ou d'ajout de champs) au vu du type de données à décrire, en faisant une synthèse des trois réponses. Leurs incompréhensions de certaines questions nous ont permis de les reformuler, de clarifier certains intitulés.

Le format proposé par DataCite, pour tout type de données

Jean-Michel Torrenti, Directeur adjoint pour la recherche et le développement du département Matériaux et structures (MAST), a manifesté son intérêt pour le projet au cours de conversations avec l'un des porteurs de GEBD-Belgrand, et a accepté de nous rencontrer. En effet, ce chercheur mettait auparavant des données de ses recherches en lignes, en libre accès, sur sa page personnelle du site de son laboratoire. Avec la fusion du LCPC et de l'INRETS, la nouvelle entité a conçu un nouveau site. Celui-ci ne permet pas (encore ?) de déposer des jeux de données accessibles par une page personnelle. M. Torrenti regrette de ne pouvoir partager certaines données. Il travaille entre autre sur la résistance de matériaux (ciments et bétons, notamment). Les expériences qu'il mène sont couteuses à réaliser, et leurs résultats facilement réutilisables. Il est donc regrettable de ne plus pouvoir les partager simplement avec la communauté des scientifiques des matériaux et des ingénieurs concevant des structures avec ces mêmes matériaux.

Bien que ces données soient fort différentes de celles des laboratoires DEST et LVMT, avec qui le projet GEBD-Belgrand se construit essentiellement, il nous a semblé intéressant de répondre à ce besoin, et d'envisager de faire de l'entrepôt de données du projet un entrepôt de données de l'institution, ou du moins son embryon.

Mais comment décrire des données issues de capteurs et mesurant la résistance d'un béton immergé dans l'eau ? Au vu d'un premier jeu de données que nous a confié M. Torrenti, il est apparu évident que la DDI ne pouvait être utilisée. D'une part, les champs concernant l'aspect enquête (la population interrogée ou la méthode d'enquête

[122] Le choix de cette version de DDI s'est effectué pour répondre à la demande de simplicité émanant des porteurs du projet, et sur les conseils, après discussions, de professionnels du réseau Quetelet.
[123] Disponible en annexe 4.
[124] Disponible en annexe 5.
[125] Disponible en annexe 6.

ou de questionnaire, par exemple) n'était pas pertinents. D'autre part, M. Torrenti insistait sur l'importance d'un questionnaire court et simple à remplir, si l'on souhaitait que des chercheurs acceptent de participer. Et enfin, il ne connaissait pas lui-même de format de description de données dans sa discipline qui fasse autorité. Et, n'ayant aucune notion de physique des matériaux, nous n'avons pas réussi, malgré quelques brèves recherches[126], à déterminer le format le plus pertinent.

Dans l'idée que, peut-être, d'autres département ou laboratoires de l'IFSTTAR pourraient se joindre au projet si celui-ci perdure, nous avons donc recherché un format qui pourrait être valable pour des types très différents de données, tout en étant suffisamment reconnu internationalement pour être, potentiellement, moissonné par différents entrepôts ou plateformes de données.

C'est ainsi que nous avons utilisé le format recommandé par DataCite, acteur international des *Research data*. Il s'agit d'un standard généraliste, simple et adaptable. Les usages qui en sont d'ores et déjà faits nous ont convaincu qu'il était suffisamment reconnu au sein de la communauté de la recherche.

Nous avons donc créé un guide du déposant à partir de ce format, que M. Torrenti a accepté de tester. La version finale découle de ce test, unique, puisque ce chercheur était à ce moment-là le seul à participer au projet[127].

Si l'on compare avec les tests du guide issu de la DDI, ce test s'est déroulé de façon extrêmement rapide et fluide, M. Torrenti comprenant d'emblée l'ensemble des questions posées, et formulant des réponses très claires. Nous ne pouvons déterminer si cette facilité découle de l'habitude de ce chercheur de partager ses données, ou du modèle de DataCite, beaucoup plus épuré que la DDI, puisqu'il doit s'adapter à tout type de données. Une autre hypothèse serait que cela vient du type de données elles-mêmes : les données issues de capteurs sont relativement simples, dans le cas qui nous occupe tout du moins[128].

Mais la question de savoir si la description universelle est possible reste entière. Pour savoir si ce format DataCite pourrait être étendu à d'autres types de données de l'institut, il faudrait poursuivre les tests, avec d'autres laboratoires et d'autres disciplines. En fait, nous émettons l'hypothèse qu'il s'agit de savoir quel type de réutilisation est envisagé : seule l'étude au cas par cas des situations et des besoins – et pas seulement du type de données- permet de se décider pour un format de description généraliste ou spécifique.

Nous aurions également aimé aborder la question des données qualitatives : après avoir rencontré l'équipe beQuali du CDSP et découvert leur travail sur ce type de données en SHS, nous avons compris qu'il ne peut s'agir que d'un projet à part entière, tant les questions que ce type de données soulève sont nombreuses. Néanmoins, cette

[126] Voir les 33 normes et standards listés par le DCC.
[127] Disponible en annexe 7.

[128] Il s'agit d'un tableau avec peu de variables, qui mesure le degré d'usure dans le temps d'un béton immergé dans une solution aqueuse. Malgré le petit nombre de champs de métadonnées, la méthode utilisée est décrite dans le champ « résumé », ce qui permet d'envisager de refaire l'expérience.

visite au CDSP nous a conforté dans l'idée que chaque projet d'ouverture des données de recherche est unique, et se construit dans le temps, guidé par une histoire – voir une tradition - et un contexte. L'équipe beQuali s'est ainsi inspirée, à son commencement, de la base anglaise nommée ESDS Qualidata[129]. C'est cet exemple qui a donné à penser que le partage de sonnées qualitatives était possible, et même souhaitable, même si l'on ne peut imaginer des retraitements automatiques à grande échelle, à l'instar de ce qui de fait pour des données quantitatives. Pourtant, au fur et à mesure de la construction de ce projet, l'équipe s'est éloignée de ce modèle en décrivant bien plus précisément les enquêtes et leurs données que ne le font les anglais. Ainsi, beQuali réalise une « enquête sur l'enquête » auprès de tout chercheur acceptant de déposer ses données, sous la forme de plusieurs entretiens. Bien sûr, cela implique un temps de travail très long et très spécialisé avant la mise à disposition de ces données…et donc des coûts. Mais, en plus de proposer des descriptions et documentations très fines, cela permet de rassurer les chercheurs en SHS, qui peuvent avoir de grandes difficultés à « lâcher » leurs données de terrain, avec lesquelles ils vivent parfois une forme d'intimité.

3.6 Des questions juridiques, techniques et de gestion en suspend

De nombreux aspects restent à aborder avant que ce projet puisse aboutir sous la forme d'un livrable.

Les questions juridiques, notamment, devront être tranchées. Quelles sont les données qui pourront être ouvertes ? Un nombre important d'études est réalisé en partenariat avec des entreprises ; il y a des contrats à respectées. L'IFSTTAR travaille également sur des données d'accidentologie : seront-elles un jour partagées, malgré les soucis qu'elles posent en termes de données personnelles ? Il faudra probablement prendre des décisions avec le concours de juristes, et opérer au fur à mesure, par type de données rencontrées.

Mais pour entamer le processus, et afin d'effectuer des choix pour les données les plus facilement partageables, nous avons rédigé une synthèse sur les licences ouvertes[130] actuellement envisageables dans le cadre du projet GEBD-Belgrand. Il s'agit de permettre aux porteurs du projet de choisir les options qui correspondront le mieux à leurs préoccupations éthiques et juridiques.

L'outil informatique est choisi depuis un certain temps, et en partie opérationnel : les porteurs de GEBD-Belgrand ont opté, avec l'aide et les conseils d'un chercheur en informatique de l'IFSTTAR, pour un logiciel libre en réseau, Dataverse[131], créé par des chercheurs de Harvard pour l'ouverture et le partage entre chercheurs de données

[129] Accessible depuis 2012 sur le site du *UK Data Service* <http://ukdataservice.ac.uk/get-data/key-data/qualitative-and-mixed-methods-data.aspx> [En ligne] Consulté le 09/09/2014.
[130] Disponible en annexe 8.
[131] Voir la présentation du *Dataverse Network Project* sur le site du projet <http://thedata.org/> [En ligne] Consulté le 16/06/2014. Le sous-titre du projet est explicite : *« A Web Application for Publishing, Citing, Analyzing, and Preserving Research Data »*.

scientifiques convenablement documentées. Cette solution, souple et relativement facile à utiliser, a été installée puis testée bien avant que notre stage n'ait lieu.

C'est donc dans cette application que nous avons essayé de transposer les résultats de nos tests sous la forme de deux *templates*[132], l'un correspondant au modèle issu de la DDI, et l'autre du standard de DataCite. Ce qui a fonctionné, avec quelques difficultés pour faire correspondre certains champs issus de la DDI.

Mais une nouvelle version de Dataverse, qui comprend un *template* correspondant très précisément à la DDI, a été mise en ligne (puis installée sur un serveur d l'IFSTAR) peu de temps avant notre départ. Cela devrait faciliter grandement la transposition du modèle que nous avons mis au point. Actuellement, le Dataverse de l'IFSTTAR, en phase de test, n'est accessible que depuis l'institut lui-même. Il reste en effet des étapes à réaliser avant de pouvoir diffuser les métadonnées en libre accès sur Internet, dans le réseau des Dataverses.

Outre les questions techniques, concernant l'interopérabilité des systèmes informatiques, d'éventuels moissonnages et l'archivage intermédiaire et/ou pérenne (en fonction du cycle de vie des donnée, encore à déterminer), et qui dépassent nos compétences, il faudra notamment se préoccuper de la problématique des identifiants pérennes. Au cours de discussions avec l'informaticien travaillant avec le réseau Quetelet comme au cours de la conférence DataCite 2014, à l'Inist, nous avons pris conscience de l'importance de cette identification, entre autre pour que les jeux de données ouverts puissent être cités[133] lors de toute réutilisation, sans soucis liés au *versionning*.[134] Cela impliquera un choix : plusieurs types d'identifiant pérennes coexistent, comme le DOI (Digital Object Identifier)[135], géré par DataCite, le Handle[136], et d'autres encore.

En termes d'organisation et de gestion, si le Dataverse de l'IFSTTAR se pérennise, il sera nécessaire d'envisager une organisation du travail de description et de dépôt des *datasets*. Les chercheurs ne pourront probablement pas effectuer ces tâches eux-mêmes, faute de formation et, surtout, de temps. Si le partage des données est chronophage pour eux, ils perdront toute motivation à participer au mouvement d'ouverture. Malheureusement, le service Documentation Scientifique, Multimédia et Technique n'a pas, à court ou moyen terme, la latitude pour embaucher une personne sur cette activité spécifique, ni la possibilité de détacher quelqu'un : il y a déjà beaucoup à faire quant aux publications.

---

[132] C'est ainsi que se nomment les masques de saisie dans Dataverse.
[133] DataCite propose depuis peu sur son site un outil très pratique pour créer des citations de jeu de donnes sous une forme complète et compréhensible sur la page « *Format your citation* » <https://www.datacite.org/services/format-your-citation.html> [En ligne] Consulté le 02/12/2014.
[134] Voir à ce sujet les pages du site de DataCite « *What do we do?* » et « *Cite your data* » https://www.datacite.org/ [En ligne] Consulté le 01/11/2014.
[135] En termes techniques, il semble que le DOI est une sorte particulière de Handle, mais avec des services fournis par DataCite, en échange d'une participation financière raisonnable qui se présente comme une souscription à l'association.
[136] Notons d'ailleurs que l'application Dataverse est prévue pour gérer DOI ou Handle, au choix.

Avant de proposer quelques pistes des différents scénarios qui permettraient de pérenniser le projet, il reste une question centrale : pourquoi avoir opté pour un entrepôt institutionnel ? Pourquoi ne pas choisir de déposer sur des plateformes disciplinaires et internationales ? Ou encore multi-disciplinaires et nationales ?

Bien sûr, une première réponse réside dans le fait que il n'existe pas d'entrepôt de données scientifiques couvrant le même périmètre, à savoir le transport et les réseaux, avec le regard de disciplines aussi variées que la sociologie, l'économie d'un côté, et les sciences des matériaux de l'autre.

L'entrepôt institutionnel permet également, comme son nom l'indique, une visibilité de l'institution tout entière, ce qui n'est pas négligeable dans le cas qui nous occupe, puisqu'il s'agit d'un établissement de création très récente qui doit se faire connaitre.

Mais aussi, un entrepôt institutionnel permet de conserver un certain contrôle sur les données : « Open… mais pas trop », pourrait-on dire familièrement.

Pourtant, n'eut il pas été plus simple d'opter pour un dépôt dans DRYAD*, qui accueille tout type de données scientifiques (les anglo-saxons disent « *domain-agnostic* » pour désigner ces bases ouvertes à toutes disciplines) pourvu qu'elles soient en lien avec une publication ? D'autant que cette préoccupation du lien entre *datasets* et publication à laquelle il a donné lieu était une demande très forte des porteurs de Belgrand-GEBD. Il existe également un entrepôt destiné à accueillir la « longue traine »[137] des données de la recherche : c'est ZENODO*. Plusieurs solutions probablement plus simples à mettre en place et peu coûteuses existaient donc.

Nous pensons – mais ce n'est qu'une hypothèse, assise sur des observations et conversations informelles - que la réponse réside dans la construction et les objectifs du projet Réseau-GEBD devenu Belgrand-GEBD, dans sa globalité et sa temporalité. Personne n'a, à proprement parler, opté pour un type d'entrepôt. Au fil du temps, les chercheurs qui portent Belgrand ont réorienté les contours du projet et l'ont renommé, en affinant la définition de leur besoin et en s'adaptant au contexte technico-informationnel et politique mouvant (celui de la recherche, par exemple avec la LRU, et celui de l'établissement, notamment avec la fusion en 2011). Tout en conservant comme objectif central le travail en réseau de chercheurs de différentes disciplines, le projet de base de données plus classique, en silo, s'est transformé peu à peu pour aboutir à l'idée du partage des données en *Open access* sur Internet. C'est aussi en prenant exemple sur le travail du réseau Quetelet, dont les pionniers de GEBD-Belgrand sont des utilisateurs, que cette histoire s'est construite. En échangeant avec Quetelet, ils ont pu se poser les questions juridiques et techniques de la diffusion de données de recherche, et imaginer des pistes de réponses. Rester dans le cadre d'une diffusion pour la recherche, voir interne, limites les risques juridiques comme les réutilisations intempestives et permet de prévoir à plus long terme un moissonnage par les « grands » entrepôts, pour assurer une visibilité plus large, tout en renforçant les

[137] Voir les cours de G. CHARTRON pour le Titre I de l'INTD, 2013/2014.

contacts entre chercheurs. Cette solution présente de réelles réponses aux besoins et objectifs en jeu.

3.7 Perspectives et scénarios

A la fin de notre intervention, à l'automne 2014, et dans l'objectif d'élargir et de pérenniser ce projet de partage des données de l'IFSTTAR, deux scénarios sont envisageables.

Comme les porteurs de Belgrand souhaitent embaucher temporairement[138] un ingénieur d'étude (IE) afin de poursuivre ce travail, cela pourrait permettre de finaliser l'outil puis d'imaginer une formation de quelques chercheurs à l'utilisation du Dataverse et de la DDI. Mais est-ce une solution de long terme, si la direction n'impulse pas rapidement une véritable politique de gestion des données[139] ? Et les chercheurs formés ne seront-ils pas rapidement repris dans le tourbillon du quotidien, déjà chargé, du travail de recherche cumulé à la gestion des laboratoires ?

L'autre scénario ferait de ce même poste d'IE un temps consacré à la simplification du *template* issu de la DDI, pour se diriger vers un modèle plus proche de celui de DataCite ou du DublinCore. Complété par une phase de formation, ceci rendrait plus réaliste l'option d'un auto-dépôt et d'une auto-description, par les chercheurs déposants, des jeux de données. Mais il serait sans doute un peu frustrant de se passer des possibilités offertes par la riche description que propose le *template* issu de la DDI pour les données d'enquêtes.

Un troisième scénario est apparu alors que le stage était terminé depuis peu, qui consisterait à utiliser le nouveau service proposé depuis juin 2014 par la TGIR Huma-Num : NAKALA. Nous n'avions pas identifié ce service comme pertinent dans un premier temps, car nous n'avons découvert son existence qu'au moment où des données de sciences des matériaux s'incluaient dans le projet. Or, Huma-Num est dédié aux SHS. Mais l'un des porteurs de GEBD-Belgrand l'ayant mentionné par mail, nous nous sommes penchés plus avant sur la présentation de NAKALA[140].

> *« Nakala permet […] à des équipes de recherche, qui en font la demande, de déposer leurs données numériques (fichiers texte, son, images, vidéo) dans un entrepôt sécurisé qui assure à la fois l'accessibilité permanente aux données et leur citabilité dans le temps. Les technologies mises en œuvre permettent de rendre interopérables les métadonnées, c'est-à-dire de pouvoir les connecter à d'autres entrepôts existants, et de les rendre moissonnables par des services spécialisés, ce qui permet d'en accroître la visibilité. »*

[138] Sur le budget du projet Belgrand.
[139] Et le peut-elle, alors que l'IFSTTAR subit, comme tout organisme de recherche publique, des baisses de budget ? Nous n'avons pas connaissance de leurs priorités, ni des directives du MEDDE pour 2015.
[140] HUMA-NUM, op.cit.

En somme, il s'agit d'un service très complet qui permet aux équipes de recherche de « se concentrer sur la valorisation scientifique de leurs données »[141], et qui semble bien répondre à l'ensemble des attentes du projet en cours.

Toutefois, il faudrait prendre contact avec Huma-Num avant d'opter pour cette solution afin d'éclaircir quelques points, entre autre :

- Huma-Num étant dédié aux SHS, le service NAKALA peut-il prendre en charges des données de sciences dites « dures » ?
- La brochure mentionne une description et des métadonnées aux formats DublinCore. Est-il possible d'utiliser d'autres standards, comme la DDI ?
- Peut-on moduler les accès et gérer finement les droits des utilisateurs, et ainsi permettre un *Open access* total pour les membres de l'établissement, mais un accès sur demande et validation pour tout autre usager ?

Nous tirons de cette anecdote de la découverte de ce service deux enseignements.

Le premier rejoint notre propos exposé en seconde partie : deux chercheurs, particulièrement au fait de l'actualité des données de la recherche car impliqué dans un projet de gestion de données, ont pris connaissance de l'existence de services qui pouvaient répondre à leur besoins quelque peu tardivement. Car, si NAKALA est récent, Huma-Num (et son « ancêtre » le TGE ADONIS avant lui) propose depuis plus longtemps des aides au partage et à l'exposition de données de recherche en SHS. Mais les porteurs de Belgrand travaillent depuis des années avec l'autre TGIR des SHS, PROGEDO (et avant l'existence de cette TGIR, avec le réseau Quetelet qui en fait partie). On peut supposer que cette co-existence de deux structures si proches a comme troublé la communication, et n'a pas permis à l'information de parvenir rapidement à ses destinataires, du moins dans le cas qui nous occupe.

Un second enseignement, de notre point de vue, est plus optimiste : ce projet d'ouverture, conçu et porté par des chercheurs, et le fait qu'ils envisagent de confier leurs données à une organisation extérieure – ici, la TGIR Huma-Num - montrent une réelle sensibilisation aux problématiques de l'*Open access* concernant les *Research data*. L'inquiétude existe, au sein du mouvement de l'*Open data*, que les chercheurs refusent de se dessaisir de leurs données, et qu'ils restent comme crispés autour d'un trésor. La généralisation serait sans doute abusive, mais nous ne pouvons que constater qu'à l'IFSTTAR, au moins, si c'est une co-construction de la gestion qui est à l'œuvre, les chercheurs sont prêts au partage et à l'ouverture bien pensée de leurs données.

[141] Ibid.

Conclusion

Conclusion

Le projet Belgrand et ses multiples noms et composantes s'est développé en parallèle de réflexions plus globales. Comme dans le secteur de l'information et du numérique en général, on est passé d'un objectif de tout mettre en bases de données pour créer une « BDD totale », au web de données et à ses outils de structuration et de description. On ne penses plus tant à créer ex nihilo des BDD qui répertorieraient le monde, qu'à faire du Web une toile suffisamment « intelligente », où toute information serait liée à d'autres, pour répondre à toute demande. Pour le projet, cela signifie qu'il s'agit de créer du lien entre des entrepôts de données accessibles sur le Web, plutôt qu'une base « en silo », accessible seulement à certains, ou depuis quelques serveurs.

Orienté au départ vers un échange de données entre chercheur de plusieurs établissements et disciplines de SHS, mais autour de la notion de ville, le Dataverse en cours de fabrication est, pour l'instant, en train de devenir un entrepôt de l'IFSTTAR, offrant une visibilité sur les laboratoires et les projets d'études, en prenant en compte le contexte pluridisciplinaire de l'IFSTTAR. Cela peut encore évoluer, s'ouvrir à d'autres établissements, ou encore devenir une partie d'un outil national, si le choix est fait de souscrire aux services proposés par la TGIR Huma-Num, voir si l'on découvre une autre offre de services. Mais quoi qu'il en soit, cela aura lieu sur le Web.

De nombreux choix restent à trancher. Les suites du projet Belgrand, qui devraient concerner les données géographiques –et donc la directive INSPIRE- pourraient être éclairantes quant au choix de méthode. Les questions juridiques, enfin, et notamment les recommandations sur les licences utilisables pour protéger et partager les données dans le cadre du Dataverse de l'IFSTTAR, constitueront une indication sur les logiques en œuvre.

Plus largement, du coté du mouvement de l'ouverture en France des données de la recherche, de nombreuses pistes seraient à suivre et creuser. Quelles seront les positions, sur un plus long terme, des actants de la recherche et de la donnée ? Les éditeurs scientifiques investiront-ils massivement ce champ en proposant de nouveaux services, comme cela arrive déjà pour certains ? L'Europe et le gouvernement français opteront-ils pour des politiques plus contraignantes ou pour des concertations et une plus grande participation des actants ? Il serait également intéressant d'écouter ce que les champs disciplinaires et les chercheurs ont à dire. Existe-t-il dans la communauté de la recherche des positionnements selon les disciplines, ou selon d'autres critères ? Concernant spécifiquement le standard DDI, il faudrait observer les utilisations des nouvelles versions : la notion de cycle de vie des données sera-t-elle opérante, et qui s'en saisira ? Et pourrons nous gérer des données qualitatives de SHS de façon satisfaisante à l'aide de cette norme de description très riche ?

Pour conclure, nous souhaiterions attirer l'attention sur ce qu'on pourrait appeler la grandeur et les limites de l'*Open data* pour les données de la recherche. L'ouverture

des *Research data* nous aidera très probablement à construire de nouvelles connaissances scientifiques, à découvrir mieux et plus vite, à innover, et à améliorer notre compréhension du monde en favorisant le partage et l'échange de savoirs.

Mais ces promesses ne doivent pas nous aveugler. Restons vigilants face aux mirages du « tout open » et des discours qui en font une réponse presque magique. Ne laissons pas le champ lexical de l'Open data et de la donnée brute, disponible et que nous n'aurions qu'à ouvrir, comme un coffre contenant un trésor, pour que la science et la connaissance fassent brusquement des bonds de géant, enfin accessibles à tous.

« *Le discours médiatique et politique sur l'open data présente l'universalité des données comme une évidence et efface leurs conditions de production*. »[142] C'est à dire un travail, important, qui permet à ces données d'exister, puis d'être diffusées et accessibles. L'activité de mise en valeur de cette production ne devrait pas être négligée.

Nous aimerions terminer avec une anecdote ayant eu lieu à l'IFSTTAR, et qui concerne l'*Open data* d'une façon plus large.

Le 3 juin 2014, Jean-Paul Hubert nous présente une juriste qui travaille sur les donnes d'accidentologie. Lors de cet entretien à trois, nous tachons, avec M. Hubert, d'expliquer le projet d'entrepôt de données, et les avantages de l'*Open data*, à cette juriste, et peut être d'obtenir d'elle quelques pistes pour des jeux de données qui nous serviraient de test.

Et nous lui racontons donc la belle aventure de Rennes Métropole qui, ayant ouvert de nombreuses données de l'agglomération, a vu certaines d'entre elles réutilisées pour une application gratuite très utile[143]. Celle-ci permet de prévoir les trajets des personnes à mobilité réduite, en leur signalant les points difficiles : hauteur des trottoirs, portes et transports ne répondant pas aux normes, etc. La juriste commente notre enthousiasme, un peu enfantin, et nous répond qu'il « serait plus intéressant de mettre l'argent public dans les travaux pour rendre réellement accessibles les trottoirs et les transports ! ».

Et en effet, lorsque l'on observe les efforts déployés pour rendre accessible des données, les compétences mobilisées, le temps et les financements nécessaires…pour permettre de se déplacer dans un espace hostile, au lieu de rendre ce territoire accueillant, on peut se demander si « le jeu en vaut la chandelle ». Quoi qu'il en soit, l'argument nous a marqué. Malheureusement, les travaux coûteraient probablement bien davantage que la mise en ligne de données, effectuées par des individus ou des institutions mais sans création de poste.

---

[142] GOETA Samuel. Ouvrir la boite noire d l'Open data : quelques premières pistes issues des coulisses. 27/02/2013. Billet du blog Les coulisses de l'Open data, carnet de recherche de S. Goëta <http://www.coulisses-opendata.com/2013/02/27/ouvrir-la-boite-noire-de-lopen-data-quelques-premieres-pistes-issues-des-coulisses/> [En ligne] Consulté le 05/12/2014.

[143] Cette réalisation est présentée dans l'ouvrage CHIGNARD Simon. Open data: comprendre l'ouverture des données publiques. Editions FYP. Paris, 2012. 191 p.

L'*open research data* est riche de promesses en terme de partage d'information et de transparence. Les tiendra-t-il ? Il ne pourra résoudre à lui seul l'ensemble des soucis. Car si être informé d'un manque est utile, y apporter une solution est tout de même plus satisfaisant. Il s'agit d'un chantier ouvert, d'un nouveau territoire d'information à construire : allons y donc avec enthousiasme, mais aussi quelques précautions.

Bibliographie

Bibliographie

La bibliographie analytique suivante, arrêtée au 20 novembre 2014, est classée par ordre alphabétique d'auteur et propose une liste des références les plus utilisées pour rédiger ce mémoire.

[24] AAF (Association des archivistes français). Les archives des établissements d'enseignement supérieur et de recherche. In *La Gazette des archives*, n°231 (2013-3).
Ce numéro de la Gazette des Archives est entièrement consacré aux archives et données de la recherche en France. Outre des exemples et une histoire de ces archives, certains articles permettent d'aborder des notions d'archivistique, comme celle d'archivage intermédiaire.

[19] ABES. « Semer, essaimer : La valorisation des données de la recherche ». *Arabesque*, n° 73, janvier 2014, 27 p. <http://www.abes.fr/Arabesques/Arabesques-n-73> [En ligne] Consulté le 03/06/2014.
Ce numéro d'Arabesques fait le point sur le rôle des infrastructures de l'IST dans la valorisation des données de la recherche.

[20] ADBS. Libre accès aux résultats de la recherche : l'ADBS dit oui à la concertation. Mis en ligne le 25 mars 2013. <http://www.adbs.fr/libre-acces-aux-resultats-de-la-recherche-l-adbs-dit-oui-a-la-concertation-127097.htm?RH=1245421882337> [En ligne] Consulté le 03/11/2014.
A propos du « Libre accès aux résultats de la recherche », l'association des documentalistes met en ligne un texte qui explique sa position qui tâche de concilier deux textes qui s'opposent : d'une part, la motion des éditeurs de revues SHS, qui craignent que l'Open access fragilise leur travail, et d'autre part, la pétition d'universitaires en faveur de l'Open access parue dans le journal Le Monde en mars 2013.

[38] ALAMI Sophie, DESJEUX Dominique, GARABUAU-MOUSSAOUI Isabelle. Les méthodes qualitatives. Coll. Que sais-je. Paris, PUF, 2013. 128 p.
Cet ouvrage explique, de manière concrète, comment se construisent des enquêtes qualitatives. Il présente les principales méthodes qualitatives, les techniques de recueil de données et souligne l'importance de la flexibilité et du sens de l'observation dont doit faire preuve l'enquêteur sur le terrain.

[10] ANDS (Australian National Data Service). « What is research data ». Site web de l'Australian National Data Service <http://ands.org.au/guides/what-is-research-data.html> [En ligne] Consulté le 03/08/2014.
L'Australian National Data Service, l'initiative australienne dédiée à la construction d'une infrastructure de données, propose sur cette page web une définition des données de la recherche qui inclut l'ensemble des éléments produits par les chercheurs dans le cours de leur activité scientifique.

[4] BEAGRIE Neil, HOUGHTON John. The Value and Impact of Data Sharing and Curation. Jisc. 2014, 26 p. <http://webarchive.nationalarchives.gov.uk/20140702233839/http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf> [En ligne] Consulté le 4/07/2014.
Ce rapport synthétise les résultats de trois enquêtes évaluant trois *Research data centers* britanniques après plusieurs années de fonctionnement. Il propose un résumé des principales conclusions et une réflexion sur les méthodes d'analyse qui peuvent être utilisées pour évaluer les impacts, avantages et coûts du partage et de la curation de données pour la recherche. Les trois études montrent que le partage de données par ces centres de données ont un impact mesurable positif sur l'efficacité de la recherche, et que d'un point de vue économique, on peut parler d'un bon retour sur investissement.

[30] BENECH Christophe. Protection et propriété des données sur Academia.edu et ResearchGate. Mis en ligne sur le blog ArchéOrient le 15 mars 2014 : <http://archeorient.hypotheses.org/2554> [En ligne] Consultés le 2/10/2014.
Ce billet de blog compare les deux principaux réseaux sociaux de chercheurs du point de vue de la protection et des usages des données et articles mis en ligne.

[13] BERHUET Solen. *Les chômeurs et les intermédiaires de l'emploi : Une sociologie dynamique de leurs trajectoires au sein d'une Maison de l'Emploi*. Thèse de sociologie, dir. M. Lallement et P. Nivolle. Cnam, 2013, 540 p.
Cette thèse de sociologie n'a pas de rapport direct avec les données de la recherche, mais permet de se familiariser avec la recherche en sciences sociales et présente de nombreuses références théoriques très utiles, expliquées et contextualisées de façon abordable pour les non sociologues. Elle est également source de bons exemples concrets pour les questions relatives aux données qualitatives.

[11] BERNERS-LEE Tim. *The next web*. 16mn23s. TED2009 conference. <http://www.ted.com/talks/tim_berners_lee_on_the_next_web> [En ligne] Consulté le 17/10/2014.
Il ya 20 ans, Tim Berners-Lee inventait le Web. Son projet suivant –consiste en la construction d'un web de données ouvertes et liées, qui pourraient faire pour les chiffres ce que le Web a fait pour les mots, les images, et la vidéo : débloquer nos données et changer la façon dont nous les utilisons, ensemble. Lors de cette désormais célèbre conférence TED, en 2009, il expose ce projet et demande au public de scander : « nous voulons des données brutes, maintenant ! ».

[16] BOAI *(*Budapest Open Access Initiative*)*. Dix ans après l'Initiative de Budapest : ce sera le libre accès par défaut. Budapest, 2012. <http://www.budapestopenaccessinitiative.org/boai-10-translations/french> [En ligne] Consulté le 17/09/2014.
Cette déclaration réaffirme, dix ans plus tard, les buts et les moyens posés par

l'Initiative de Budapest de 2002. Mais, de surcroît, un nouvel objectif est établi : dans dix ans, le libre accès doit être la solution par défaut pour la dissémination de nouvelles recherches dans tous les domaines et dans tous les pays.

[15] CHARTRON Ghislaine, « Open access et SHS : Controverses », *Revue européenne des sciences sociales* 1/ 2014 (52-1), p. 37-63 <www.cairn.info/revue-europeenne-des-sciences-sociales-2014-1-page-37.htm> [En ligne] Consulté le 30/07/2014.
Cet article s'attache à discuter le bien-fondé de l'injonction politique en faveur du libre accès aux publications scientifiques au regard des spécificités de la recherche en sciences humaines et sociales. Partant du constat que les politiques publiques s'élaborent le plus souvent selon des caractéristiques empruntées aux sciences dites « dures », l'auteur questionne la pertinence des postulats avancés, en particulier dans la perspective des données ouvertes et du datamining.

[37] CHARTRON Ghislaine. La valeur des services documentaires en prise avec le numérique. In BBF, vol. 57, n°5, 2012, p. 14-18, <http://bbf.enssib.fr/consulter/bbf-2012-05-0014-003> [En ligne] Consulté le 12/1/2014.

[36] CHIGNARD Simon. Open data: comprendre l'ouverture des données publiques. Editions FYP. Paris, 2012. 191 p.
Cet ouvrage constitue un guide pour comprendre l'Open data et son arrivée en France. De nombreux exemples de réalisations illustrent les propos de l'auteur, partisan de l'ouverture des données.

[23] COLLECTIF. Les métiers de l'information et la « donnée » : analyse d'un monde en mutation. In *Documentaliste-Sciences de l'Information* 50, n° 3, 2013, p.26-41. <http://www.cairn.info/revue-documentaliste-sciences-de-l-information-2013-3-page-26.htm> [En ligne] Consulté le 16/06/2014.
Ce numéro, entièrement consacré aux métiers de l'information du point de vue de la gestion des données, illustre celui-ci avec de nombreux exemples, et propose des pistes méthodologiques.

[21] COLLECTIF. Qui a peur de l'open access ? Pétition parue dans *Le Monde* du 15/03/2013.
<http://www.lemonde.fr/sciences/article/2013/03/15/qui-a-peur-de-l-open-acces_1848930_1650684.html> [En ligne] Consulté le 05/11/2014.
En mars 2013, une communauté de responsables d'universités (d'enseignants-chercheurs, d'éditeurs, de responsables de bibliothèques) répond à la motion d'éditeurs de revues de SHS, qui craignent un déséquilibre économique pour leurs petites publications, et appuie le mouvement de 'Open access des publications scientifiques.

[35] DDI ALLIANCE ; BARNIER Julien, BOUVIER Jean-Noël et al. (trad.). Traduction française du dictionnaire de balises de la norme DDI. Réseau Quetelet, 2006. <http://www.reseau-quetelet.cnrs.fr/spip/IMG/pdf/DDI_versionFR.pdf> [En ligne] Consulté le 02/06/2014.
Il s'agit de la traduction en français du *Codebook* de DDI v.1.2.2. réalisée par des

membres du réseau Quetelet. Cette version de DDI représente une forme allégée du standard, dans une version ancienne, plus simple à découvrir que les versions 3 et suivantes, puisqu'elle n'intègre pas la notion de « cycle de vie des données ».

[8] DENIS Jérôme et GOËTA Samuel, « La fabrique des données brutes. Le travail en coulisses de l'open data ». Paris, 2013, 19 p. <http://halshs.archives-ouvertes.fr/halshs-00990771> [En ligne] Consulté le 21/07/2014.
Le vocabulaire de la « libération », de la « transparence » et plus encore celui de la « donnée brute » effacent toute trace des conditions de production des données, des contextes de leurs usages initiaux et pose leur universalité comme une évidence. Ce texte explore les coulisses de l'open data afin de retrouver les traces de cette production et d'en comprendre les spécificités. À partir d'une série d'entretiens ethnographiques, il décrit la fabrique des données brutes, dont l'ouverture ne se résume jamais à une mise à disponibilité immédiate, évidente et universelle.

[2] FAYET Sylvie. « "Données" de la recherche, les mal-nommées ». *URFIST Info*, 15 novembre 2013. <http://urfistinfo.hypotheses.org/2581> [En ligne] Consulté le 4/06/2014.
Ce billet de blog fait, concrètement et avec humour, le tour des questions soulevées, du point de vue des professionnels de l'IST, par la gestion de l'ouverture des données de la recherche en France.

[17] FIORASO Geneviève. Discours du 25/01/2013 de la secrétaire d'Etat à la Recherche lors des cinquièmes journées Open Access. Disponible en ligne sur le site du MESR <http://www.enseignementsup-recherche.gouv.fr/cid66992/discours-de-genevieve-fioraso-lors-des-5e-journees-open-access.html> [En ligne] Consulté le 02/11/2014.
La secrétaire d'Etat à la Recherche expose dans ce discours la position du MESR : au regard de l'importance des enjeux scientifiques, économiques et sociétaux, le gouvernement français réaffirme son soutien au principe du libre accès à l'information scientifique et renouvelle son soutien aux archives ouvertes. Pour les publications et l'édition scientifique, le gouvernement souhaite développer l'Open Access Green, accompagner l'évolution de l'Open Access Gold, et promouvoir le développement d'une troisième voie innovante et durable, dites Platinium Road.

[9] FOURNIER Thierry, « Les données de la recherche : définition et enjeux. » In *Arabesques* n°73, ABES, p. 4-6.
Thierry Fournier, responsable scientifique de l'ADBU, retrace ici les contours de l'ouverture des données de la recherche, et de ce que cela implique pour la documentation.

[26] FRANCE Ministère de l'enseignement supérieur et de la Recherche. L'état de l'Enseignement supérieur et de la Recherche, n°7, Mars 2014. <http://cache.media.enseignementsup-recherche.gouv.fr/file/EESR_2014/60/7/EESR7_316607.pdf> [En ligne] Consulté le 8/08/2014.
Comme les éditions précédentes, cette 7eédition de L'état de l'enseignement

supérieur et de la recherche présente un état des lieux annuel et chiffré du système français, de ses évolutions, des moyens qu'il met en œuvre et de ses résultats, en le situant, chaque fois que les données le permettent, au niveau international.

[32] FRANCE. Décret no 2001-139 du 12 février 2001 portant création du comité de concertation pour les données en sciences humaines et sociales. <http://www.reseau-quetelet.cnrs.fr/spip/IMG/pdf/Decret_01-139-3.pdf> >[En ligne] Consultés le 5/06/2014.
Comme son titre l'indique, ce décret définit la création, les missions et le fonctionnement du CDSHS.

[27] FRANCE. Loi n°2013-660 du 22 juillet 2013 relative à l'enseignement supérieur et à la recherche. <http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000027735009> [En ligne] Consulté le 10/07/2014.
Dernière loi en date concernant la stratégie de l'Enseignement supérieur et de la recherche en France, ce texte organise notamment la coordination avec la stratégie européenne et la cohérence avec les priorités du plan-cadre Horizon 2020.

[1] GAILLARD Rémi. « De l'Open data à l'Open research data : quelle(s) politique(s) pour les données de recherche ? ». Mémoire, Enssib, 2014, 104 p. <http://www.enssib.fr/bibliotheque-numerique/notices/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche> [En ligne] Consulté le 20/04/2014.
Ce mémoire fait le point sur la situation du mouvement du libre-accès aux publications scientifiques, qui s'élargit de plus en plus aux données de la recherche. Des initiatives pour garantir l'accessibilité et la complète réutilisation de ces données sont prises par une grande diversité d'acteurs – États, agences de financement de la recherche, éditeurs, communautés scientifiques. L'ouverture des données de la recherche est rendue possible par la définition de politiques incitatives ou contraignantes, l'adoption de solutions juridiques et techniques, mais repose avant tout sur de bonnes pratiques de gestion des données.

[22] GFII. 1ères observations du GFII sur les recommandations de la CE en matière d'Open Access. 11/01/2013. <http://www.gfii.fr/fr/document/recommandations-de-la-commission-europeenne-en-matiere-d-open-access-premieres-observations-du-gfii> [En ligne] Consulté le 05/11/2014.
Le GFII, groupement interprofessionnel réunissant les acteurs, publics et privés impliqués dans l'industrie de l'information et de la connaissance, informe dans ce texte les pouvoirs publics des premières conclusions auxquelles est arrivé son groupe de travail Open Access.

[12] GITELMAN Lisa (dir.). *Raw Data is An Oxymoron.* Cambridge, MIT Press, 2013, 182p.
Cet ouvrage collectif rappelle que nous vivons dans l'ère du Big Data, où la collecte de données est constante, voir insidieuse. La thèse qui y est développée pose que les

données ne sauraient jamais être «brutes» ; il ne faudrait pas les considérer comme une ressource naturelle mais comme une ressource culturelle, devant être créée, protégée, et interprétée.

[39] HANSON Karen, SURKIS Alisa, YACOBUCCI Karen. Data Sharing and Management Snafu in 3 Short Acts. New York, 2012. NYU Health Sciences Library, 4.40 mn. <http://www.youtube.com/watch?v=66oNv_DJuPc&feature=youtube_gdata_player> [En ligne] Consulté le 15/08/2014.
Ce court métrage pédagogique d'animation, minimaliste sur le plan visuel, recréé avec humour une situation probablement maintes fois vécue : un chercheur souhaite accéder aux données d'un autre chercheur. Il s'agit d'une synthétique illustration des nombreuses difficultés rencontrées pour obtenir, lire puis comprendre un jeu de données constitué par autrui et qui n'a pas été convenablement géré.

[33] HUMA-NUM. Les services de conservation de données proposés par Huma-Num. Mai 2014 <http://www.huma-num.fr/sites/default/files/ressourcesdoc/dossier-thematique-mai2014.pdf> [En ligne] Consulté le 25/06/2014.
Huma-Num propose aux équipes de recherche des solutions pour assurer la sauvegarde, la pérennité et le partage des données. Ce dossier, version mise à jour d'un guide paru en mars 2013, présente l'ensemble des services proposés par la TGIR, dont NAKALA, service d'exposition de données, né en juin 2014.

[14] JEANNERET Yves, *Y a-t-il -vraiment- des technologies de l'information*, Presses Univ. Septentrion, 2000, 134 p.
Yves Jeanneret explique que le discours sur les « technologies de l'information » et la « société de l'information » se fondent sur l'illusion que l'information serait une matière première dont le traitement informatique suffirait à produire de la connaissance. Ainsi, il distingue deux types d'information : l'information1 (mathématique) et l'information² (sociale).

[28] LATOUR Bruno. Changer de société. Refaire de la sociologie. La Découverte, 2006, 402 p.
Cet important ouvrage de sociologie de la science comprend, entre autre, une définition et des utilisations du concept d'actant, utilisé dans ce mémoire.

[5] MADAY Charlotte. « Les professions archivistiques et documentaires donnent des définitions différentes des archives imprimées et numériques de la recherche. » Territorial.fr. Interview de Charlotte MADAY, mis en ligne le 28/10/2014. <http://www.territorial.fr/PAR_TPL_IDENTIFIANT/69669/TPL_CODE/TPL_ACTURES_FICHE/PAG_TITLE/Archives+de+la+recherche+%281%29++%3A+une+probl%E9matique+qui+monte+en+puissance+++/302-actu.htm> [En ligne] Consulté le 15/11/2014.
Charlotte Maday, présidente de la section AURORE (réseau des archivistes des universités, rectorats, organismes de recherche et mouvements étudiants) de l'AAF (l'Association des Archivistes Français), résume dans cette interview le point de vue

des archivistes sur le sujet des données de la recherche.

[7] OCDE. *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics*, Paris, 2007, 29 p. <http://www.oecd.org/fr/science/sci-tech/38500823.pdf> [En ligne] Consulté le 4/07/2014.
L'OCDE propose des orientations aux responsables politiques pour élaborer des politiques et des bonnes pratiques autour des données numériques de la recherche financées sur fonds publics, afin de faciliter leur accès, leur exploitation et leur gestion.

[6] OMB (Office of management and budget). USA, 1993. Circulaire A-110 amendée le 30/09/99. <http://www.whitehouse.gov/omb/circulars_a110#36> [En ligne] Consulté le 3/10/2014.
Le Bureau de la gestion et du budget du gouvernement fédéral américain a rédigé cette circulaire (en 1993, amendée en 1999) qui contient notamment la définition suivante des données de recherche : « La donnée de recherche est définie comme l'enregistrement factuel couramment considéré dans la communauté scientifique comme nécessaire à la validation des résultats de la recherche ». Adoptée par de nombreuses institutions américaines et par l'OCDE, cette définition est l'une de celles qui font référence.

[29] PERUCCA Brigitte. Présentation de l'enquête «L'usage des réseaux sociaux par les scientifiques». CNRS, 2013. <http://corist-shs.cnrs.fr/sites/default/files/evenements/brigitteperucca_reseauxsociaux.pdf> [En ligne] Consulté le 2/10/2014.
Afin de comprendre les processus d'échanges informels à l'œuvre dans les communautés de chercheurs pour mieux valoriser l'Open Access, le CNRS a conduit une enquête destinée à connaître les usages et pratiques des réseaux sociaux de la recherche (type Academia.edu) et de l'Open Access par les chercheurs. Cette présentation livre les premiers résultats sous forme essentiellement graphique

[31] RIANDEY Benoît. Centre Quetelet, deuxième anniversaire. In Courrier des statistiques n°107, sept. 2003, p 33-36. <http://www.insee.fr/fr/ffc/docs_ffc/cs107d.pdf>[En ligne] Consultés le 2/06/2014.
Cet article présente le fonctionnement, les missions et les partenaires de Quetelet, lors de ses premières années de vie, alors qu'il s'agissait d'un « centre » et non encore d'un réseau. La DDI et l'intérêt de son utilisation y sont brièvement expliqués.

[3] SILBERMAN Roxanne. Les sciences sociales et leurs données, Rapport de mission au Ministre de l'Education nationale, de la recherche et de la technologie. La Documentation Française. 1999, 193 p.
Le rapport Silberman, rédigé à la demande du MESR, est le texte fondateur de la prise en compte des données en sciences sociales en France. C'est à sa suite que seront constitués le CCDSHS puis le Centre Quetelet. Il s'agit d'un état des lieux : il liste donc les retards, problèmes et questions que les données de la recherche

posent en sciences sociales en 1999, et propose des pistes pour améliorer la situation de la recherche française.

[25] SOULE Véronique. Regroupement des universités : le haussement des pôles ; et interview de Pierre-Paul Zalio, président de l'ENS Cachan : « Le système français est complexe et illisible. » in *Libération* du 19/10/2014.
Cet article dresse un premier bilan du regroupement des établissements d'enseignement supérieur en 25 grands pôles, concrétisation de l'une des mesures de la loi du 22 juillet 2013. Ce système, qui doit permettre plus de lisibilité, est encore peu compris, voir méconnu. M. Zalio, président de l'ENS Cachan et partisan de cette réforme, expose son point de vue dans l'interview qui suit, et note que le système actuel est complexe et illisible.

[18] THESSEN Anne E., PATTERSON David J. « Data Issues in the Life Sciences ». In *ZooKeys*, novembre 2011, n° 150, p. 15-51.
Cet article passe en en revue les questions techniques et sociologiques en lien avec les sciences de la vie, au moment où celles-ci évoluent en disciplines centrées sur les données massives, leur traitement et les découvertes que cela peut engendrer.

[34] THIBAUT Françoise. La Lettre d'ATHENA, mai 2014. <http://www.allianceathena.fr/sites/default/files/telechargements/la_mai_14_a4_2.pdf> [En ligne] Consulté le 02/10/2014.
Cette lettre d'information de l'alliance ATHENA est entièrement consacrée aux deux TGIR des SHS : PROGEDO et Huma-Num. Elle contient entre autre une comparaison des eux structures sous forme de face à face.

Annexes

Annexe 1 : Cartographie des acteurs et initiatives Open research data ayant un impact en SHS en France

Ce tableau constitue une cartographie des actants rencontrés parmi les lectures effectuées durant le stage à l'IFSTTAR et la rédaction de ce mémoire. Le parti pris étant de ne pas trier ces actants à priori, et d'en faire des entrées du document au fur et à mesure que nous les avons rencontré, la liste des items fait apparaître côte à côte des projets et des initiatives aussi bien que des acteurs institutionnels ou des outils. Malgré cette large ouverture, ce travail ne prétend pas être exhaustif. Et, au vu du nombre croissant de projets concernant les données de recherche, il cessera rapidement d'être à jour. Pour le moment, soit à la fin de l'année 2014, il constitue un outil de repérage si l'on souhaite s'initier aux jeux des acteurs des données de la recherche en SHS et en France.

Catégories, Quoi, qui, pourquoi	Acronyme Site Web Source(s)	Présentation
France **Acteur** **Statut**, ce que c'est : **UMR** **Qui le fait**/le porte : **CNRS-EHESS-ENS** **Ce que ça fait** quant aux données de recherche en France : **Gestion et diffusion de données** **Catalogue de données** **Membre**	ADISP http://www.cmh.ens.fr/greco/adisp.php Source complémentaire : http://www.reseau-quetelet.cnrs.fr/	L'ADISP (Archives de Données Issues de la Statistique Publique, unité mixte de recherche CNRS-EHESS-ENS) a constitué un fonds de grandes enquêtes et de bases de données intéressant les sciences sociales et continue de l'alimenter grâce à des conventions signées avec l'INSEE, plusieurs services statistiques ministériels et d'autres institutions publiques. Il met ce fonds à disposition des chercheurs et conseille ceux qui le souhaitent sur l'utilisation des données. Au sein du Centre Maurice Halbwachs, l'ADISP met à disposition son fonds de grandes enquêtes françaises intéressant les sciences sociales. Depuis décembre 2001, le Centre Maurice Halbwachs est un des partenaires du Réseau Quetelet et participe au réseau européen des archives de données à travers le CESSDA. Le CMH-ADISP archive et diffuse essentiellement des données issues de la statistique publique (INSEE, CEREQ, Services statistiques des Ministères en charge de l'Emploi, des Affaires Sociales, de l'Education Nationale, de la Culture). Constitué il y a plus de 20 ans autour de quelques enquêtes phares pour les besoins de sociologues du CNRS, le catalogue s'est progressivement enrichi et ouvert à d'autres disciplines. Il est essentiellement constitué de données issues de la statistique publique : INSEE, CEREQ, OVE, services statistiques ministériels, etc. Le fonds d'enquêtes et bases de données réunies par le CMH-ADISP compte aujourd'hui plus de 780 références réalisées par différents **organismes publics** (INSEE, CEREQ, OVE, services statistiques des Ministères chargés de l'Emploi, des Affaires sociales, de l'Éducation nationale, de la Santé, de la Culture, des transports, etc.) et réparties en 10 grands thèmes :

Quetelet		éducation et formation travail et emploi salaires et revenus conditions de vie et société santé et protection sociale population et démographie opinions entreprises données localisées données historiques Enquêtes ponctuelles ou régulières, données de recensement, données de panel, extraits de fichiers administratifs, données localisées selon un maillage très fin du territoire, données représentatives au niveau national, données récentes ou historiques, le catalogue du CMH-ADISP s'accroit régulièrement grâce à l'acquisition des nouvelles enquêtes produites par les services statistiques français et par la recherche et la restauration d'enquêtes anciennes.
France **Initiative** **TGIR** **Ministère ESR** **CPU** **CGE** **CNRS** **Coordination nationale des politiques des données**	BSN Segment 10 http://www.bibliothequescientifiquenumerique.fr/ Source complémentaire : http://bbf.enssib.fr/consulter/bbf-2013-01-0061-014	Créée en 2009, à l'initiative du ministère de l'enseignement supérieur et de la recherche en fédérant de nombreux acteurs des universités et organismes de recherche, la bibliothèque scientifique numérique (BSN) veille à ce que tout enseignant-chercheur, chercheur et étudiant dispose d'une information scientifique pertinente et d'outils les plus performants possibles. En accord avec les orientations de la Commission européenne, la BSN privilégie l'accès ouvert aux documents scientifiques sous différentes formes reposant sur des innovations, des négociations avec les éditeurs ou le soutien aux archives ouvertes, en tenant compte des différences entre les disciplines. BSN facilite également l'accès aux ressources scientifiques documentaires en rendant plus visible le paysage. Le nouveau segment 10 de BSN (création en 2014) est consacré aux données de la recherche.
France **Acteur**	CCDSHS Sources : http://www.in	Créé dans la suite du rapport Silberman (Les sciences sociales et leurs données, Silberman, 1999), le Comité de concertation pour les données en sciences humaines et sociales (CCDSHS**)** propose des orientations nationales pour une politique publique de données pour

Ministères ESR, **Economie,** **Emploi.** **CNRS** **CPU** **Elaboration de la politique nationale des données**	see.fr/fr/ffc/docs_ffc/cs107c.pdf Courrier des statistiques n° 107, septembre 2003 http://www.reseau-quetelet.cnrs.fr/spip/article.php3?id_article=237	la recherche en sciences humaines et sociales selon trois grandes dimensions : -faciliter l'accès aux ensembles de données utiles à la recherche, -accroître l'utilisation de ces données, -soutenir la réalisation de grandes enquêtes utiles à la recherche. A cette fin, il organise la concertation entre producteurs de données au sein de la statistique publique, de la recherche ou du secteur privé et utilisateurs de données dans une finalité de recherche. Le CCDSHS a été créé par décret D2001-139 du 12/02/2001 auprès des ministères français chargés de l'économie, de l'emploi, de l'éducation nationale et de la recherche. Il est placé sous la présidence du ministre en charge de la recherche. Le CNRS et la Conférence des Présidents des Universités (CPU) sont associés au Comité. D'autres instances sont invitées de façon permanente ou plus ponctuellement telles que les Archives nationales et la CNIL. Aux termes du décret n° 2009-318 du 20 mars 2009 relatif au Conseil national de l'information statistique (CNIS) et au Comité du secret statistique et de l'arrêté pris en application du décret, le Conseil scientifique du Comité des données (CCDSHS) désigne plusieurs représentants des chercheurs dans les différentes instances du CNIS, assemblée plénière, bureau, comité du label, commission nationale des nomenclatures économiques et sociales, comité du secret statistique. - Représentants des chercheurs en sciences économiques et sociales au CNIS : Roxane Silberman et François-Charles Wolff - Bureau du CNIS : Roxane Silberman - Comité du secret statistique : Catherine Rhein - Comité du label : Laurent Lesnard
France **Acteur** **UMS 828, CNRS-Sciences Po** **Gestion et diffusion de données** **Catalogue de données** **Membre de Quetelet**	CDSP http://cdsp.sciences-po.fr/ http://www.bequali.fr/bequali/ Sources complémentaires : entretien avec l'équipe beQuali	Le Centre de Données Socio-Politiques (UMS 828, Sciences Po / CNRS) est membre du Réseau Quetelet et a pour mission de faciliter l'accès des chercheurs et des étudiants à des données socio-politiques. Ces données sont mises gratuitement à disposition à des fins de recherche. Depuis le début de l'année 2011, le CDSP est fortement impliqué dans la mise en place de l'équipement d'excellence DIME-SHS (Données Infrastructures et Méthodes d'Enquête en Sciences Humaines et Sociales). Le catalogue du CDSP comporte un nombre important des enquêtes majeures conduites en France dans le domaine socio-politique dont : -les enquêtes pré et post électorales du CEVIPOF -le Baromètre Politique Français - l'enquête Démocratie 2000 -les enquêtes conduites auprès des élus, des cadres et militants de partis politiques -les enquêtes de l'Observatoire Interrégional du Politique (OIP)

		-les enquêtes Image de la science -Les résultats électoraux : Le CDSP diffuse les résultats des élections politiques en France depuis 1958. A partir des données fournies par le Ministère de l'intérieur, un code politique a été attribué à chaque liste ou candidat selon un ordre gauche-droite. Les résultats sont présentés à tous les niveaux territoriaux. L'analyse de l'évolution des rapports de force entre les partis en présence en est facilitée. Cantonales (depuis 1985) Européennes (depuis 1984) Législatives (depuis 1958) Municipales (depuis 1983) Présidentielles (depuis 1965) Referendum (1946,1992, 2000, 2005) Régionales (depuis 1986) Projet et équipe BeQuali (équipement DIME-SHS) : L'équipe beQuali a pour mission non seulement de donner accès aux enquêtes mais de développer les outils et les méthodes permettant de les contextualiser et rendre ainsi possible leur utilisation. La mise à disposition de ces enquêtes répond à des enjeux scientifiques, déclinés en quatre objectifs : Renforcer la scientificité de la démarche qualitative Améliorer la méthodologie qualitative Faciliter les comparaisons Favoriser l'histoire des sciences Projet qui a pour modèle son équivalent anglais : ESDS Qualidata, devenue une partie de UK Data Service en 2012.
Europe **Acteur** **Infrastructure européenne** **Futur ERIC** **Nœud français :**	CESSDA http://www.cessda.net/	Le CESSDA (Council of European Social Science Data Archives) est le réseau européen des Archives de données pour la recherche en sciences sociales. Le CESSDA figure dans la feuille de route européenne comme l'un des candidats au statut de grande infrastructure européenne pour la recherche. Le Réseau Quetelet participe à la phase préparatoire de cette construction dans le cadre de l'appel à proposition de la Commission européenne (CESSDAPPP7). Il a notamment coordonné le *work package* consacré au périmètre des collections, aux questions d'accès aux données de la statistique

PROGEDO **CNRS** **Réseau d'archives** **Catalogue de données**		publique et à l'homogénéisation des licences d'accès en Europe. Le CESSDA propose un catalogue des données de recherche mises à disposition en Europe.: http://www.cessda.net/catalogue/ Celui-ci fournit une interface pour les jeux de données de sciences sociales des archives de données à travers l'Europe ; il peut être fouillé en recherche libre et consulté en neuf langues. Les éditeurs de données du CESSDA : (9496 études disponibles en sept 2014) Quetelet-ADISP (384 études) APIS (27 études) ADPSS-Sociodata (62 études) Quetelet-CDSP (63 études) CSDA (1009 études) DANS (124 études) ADP (1115 études) FORS (754 études) FSD (2010 études) GSDB (57 études) NSD (1034 études) ISSDA (97 études) LiDA (424 études) NSD Metadata (1034 études) UKDA (665 études) GESIS ZACAT (637 études) Le gouvernement français s'engage dans la participation à l'infrastructure européenne CESSDA (*Council of European Social Sciences Data Archives*). Un nouveau Consortium est formé sur la base d'un MoU (*memorandum of understanding* en anglais, soit mémorandum d'entente) signé par 14 pays membres ou associés. Une société de droit norvégien, CESSDA AS, assure provisoirement le portage juridique de l'infrastructure et le statut d'ERIC (European Research Infrastructure Consortium) prévu pour les infrastructures européennes de recherche, sera adopté au plus tard fin 2015.

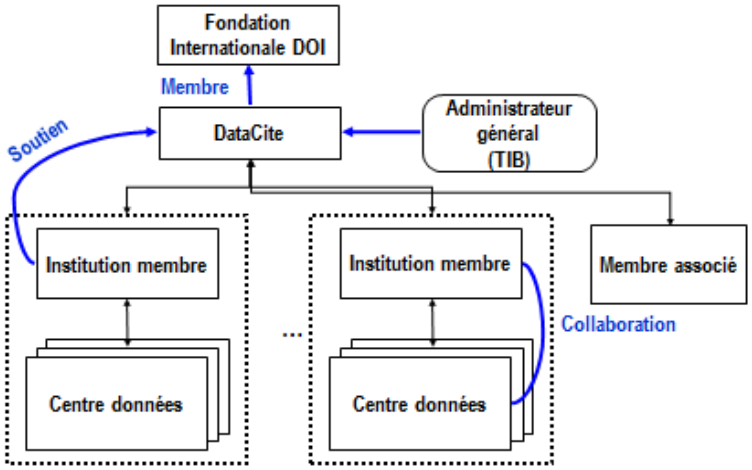
		La TGIR PROGEDO, dont fait partie le réseau Quetelet, est désignée *Service Provider* pour le nœud français, porté par le CNRS. Il reviendra à PROGEDO d'organiser le partenariat français en lien avec le Réseau Quetelet et ses partenaires de manière à assurer les services qui devront être rendus dans le cadre de l'infrastructure européenne.
France **Acteur** **EPCA** **Ministère ESR** **Services et outils pour l'archivage pérenne**	CINES https://www.cines.fr/	Le **C.I.N.E.S.** (***C**entre **I**nformatique **N**ational de l'**E**nseignement **S**upérieur*) est un Etablissement Public à caractère Administratif national, basé à Montpellier et placé sous la tutelle du Ministère chargé de l'Enseignement Supérieur et de la Recherche. Le CINES propose des moyens exceptionnels à l'ensemble de la communauté scientifique dans ses missions statutaires stratégiques nationales : - le calcul numérique intensif, - l'archivage pérenne de données électroniques. Le CINES est le site officiel d'archivage des thèses électroniques, des archives ouvertes HAL, des revues en sciences humaines et sociales de PERSEE, etc. En dehors de ces documents, un certain nombre de données scientifiques, résultats d'observations ou de simulations à destination de la recherche ou à but pédagogique, sous différentes formes (tableaux de nombres, images, sons, vidéos, …), sont potentiellement candidates à l'archivage. Cependant il est souvent difficile de sélectionner en amont celles qui doivent être archivées de façon définitive. Aussi le CINES met en place un service « d'archivage intermédiaire de données scientifiques » pour des communautés d'utilisateurs intéressées. Ce service correspond à un stockage sécurisé des données, comportant obligatoirement un jeu minimal de métadonnées descriptives associées, afin de faciliter la recherche et la compréhension, pour une période déterminée (3 ou 4 ans) et offrira une possibilité d'accès partagé à ces données pour la communauté. Le projet ISAAC : l'Information Scientifique Archivée Au CINES = un archivage intermédiaire des données de la recherche. Le CINES participe à l'initiative « Bibliothèque Scientifique Numérique » mise en place par le Ministère de l'Enseignement Supérieur et de la Recherche et pilote dans ce cadre le groupe de travail consacré à l'archivage pérenne (segment 6). Le projet PAC : archivage pérenne. Au niveau international, le CINES participe à wePreserve, APARSEN et EUDAT.

International **Acteur** **ONG** **Association** **Réseau de sensibilisation** **Edition d'un *Data journal***	CODATA http://www.codata.org/ http://codata-france.org/	CODATA (COmmittee on DATA for Science and Technology, soit Comité sur les données de la science et de la technologie), est un comité scientifique interdisciplinaire de l'ICSU (International Council for Science organisation non-gouvernementale mondiale dont les membres sont des organisations scientifiques représentant 140 pays et des Unions scientifiques, correspondant à 31 membres), créé il y a 40 ans. CODATA a été créé en 1966 par l'ICSU pour promouvoir et encourager, à l'échelle mondiale, la compilation, l'évaluation et la diffusion de données numériques fiables d'importance pour la science et à la technologie. CODATA travaille à améliorer la qualité, la fiabilité, la gestion et l'accessibilité des données d'importance pour tous les domaines de la science et de la technologie. CODATA est une ressource qui permet aux scientifiques et aux ingénieurs un accès aux activités internationales concernant les données afin d'accroitre la sensibilité à ces questions, ainsi qu'une coopération directe et de nouvelles connaissances. L'action de CODATA s'exprime à travers des *Task Groups* et *Working Groups*, les activités des membres au niveau national, la participation à des initiatives internationales, telles que : IPY (*International Polar Year*), eGY (*Electronic Geophysical Year*), l'organisation de conférences internationales, le Journal de CODATA (CODATA Data Science Journal**)**, l'organisation de *Workshops*, la publication de différentes études et rapports, mais aussi par la coopération et la liaison avec des organisations scientifiques internationales (UNESCO, OCDE, GEO, …). CODATA France est le représentant, en France, de CODATA et une Association Loi de 1901**.** Sa mission est de renforcer la contribution scientifique française au bénéfice de la société civile, par la promotion de la gestion et de l'utilisation des données scientifiques et techniques. Pour réaliser cette mission, CODATA France vise à promouvoir l'évaluation et le contrôle de la qualité des données, mais aussi l'amélioration des méthodes d'acquisition, de gestion, d'analyse, de validation, de dissémination et d'archivage de ces données. La réalisation de cette mission nécessite la coopération entre l'ensemble des acteurs utilisant et travaillant sur ces données et les traitements qui leur sont appliqués. Fondé en 1981, le comité CODATA France s'appuie sur des universitaires, des chercheurs, des professionnels du monde de l'entreprise et d'autres organisations, et conduit des actions spécifiques dans différents secteurs professionnels impliquant une interaction étroite entre Science, Technologie et Société. Il a également pour objectif d'informer sur les systèmes d'identification et d'organisation des données, dans les domaines scientifiques et techniques, mais aussi dans tous les secteurs de la vie économique, juridique et sociale.

France **Acteur** **Association d'établissements** **Réseau d'établissements** **Négociations** **Bureau de liaison français pour OpenAire**	Consortium Couperin http://www.couperin.org/ Source complémentaire : interview de JP Finance, président du CA de Couperin	Consortium unifié des établissements universitaires et de recherche pour l'accès aux publications numériques, créé en 1999 sous la forme d'une association loi 1901, par les quatre universités fondatrices (Strasbourg 1, Nancy 1, Marseille 2, Angers). Couperin s'est donné pour missions de : Recueillir et analyser les besoins documentaires de ses membres. Evaluer**,** négocier et organiser l'achat de ressources documentaires numériques au bénéfice de ses membres. Développer un réseau national de compétences et d'échanges en matière de documentation électronique notamment concernant les politiques d'acquisitions, les plans de développement de collections, les systèmes d'information, les modèles de facturation des éditeurs, l'ergonomie d'accès, les statistiques d'usage... Contribuer à clarifier et à faire évoluer les relations contractuelles Contribuer au développement d'une offre de contenu francophone. Œuvrer à l'amélioration de la communication scientifique et favoriser la mise en place de systèmes non-commerciaux de l'Information Scientifique et Technique (IST) par le développement d'outils adéquats. Développer une expertise et une évaluation des systèmes d'information documentaire et de leurs outils ainsi que des méthodes d'intégration de ceux-ci au sein des systèmes d'information des établissements, en cohérence avec les autres institutions en charge du développement et de l'implantation de systèmes d'information dans le monde de l'Enseignement Supérieur et de la Recherche. Favoriser la coopération nationale**,** européenne et internationale dans le domaine de la documentation et des publications électroniques. Couperin reconnaît le rôle positif et déterminant exercé de tous temps par les éditeurs scientifiques mais refuse le diktat de certains groupes mondiaux aux visées monopolistiques. Couperin s'engage, aux côtés d'autres acteurs internationaux de l'information, à promouvoir l'édition scientifique libre et alternative. Enfin, Couperin participe à l'élaboration d'un schéma national de préservation des données et des ressources avec les opérateurs nationaux et internationaux. Couperin coordonne l'*Open Access week* qui se déroulera en France du 13 au 26 octobre 2014. Couperin assure une mission de veille et d'expertise sur les questions techniques liées à la diffusion de la documentation électronique et vise à mutualiser les expériences et les projets dans un domaine de compétence indispensable au développement des établissements.

		Ouvert sur la base du volontariat aux personnels des établissements Couperin ou non, le consortium anime plusieurs groupes de travail, listes de discussion et assure une veille et une expertise dans les domaines suivants : les Archives ouvertes et le Libre Accès les livres électroniques au sein de la Cellule Ebook (CeB) les systèmes de gestion et d'accès aux ressources électroniques l'accès distant aux ressources électroniques l'archivage pérenne les statistiques d'utilisation les indicateurs de pilotage des politiques documentaires numériques Couperin a mis en place un site dédié aux archives ouvertes : http://www.couperin.org/archivesouvertes/ Ce site s'adresse à l'ensemble des enseignants-chercheurs et des établissements d'Enseignement Supérieur et de Recherche qui souhaitent s'associer aux initiatives d'Archives ouvertes et plus largement à l'Open Access. Il offre des outils et des conseils afin d'en favoriser la mise en œuvre, soit de manière individuelle, soit à l'échelle d'un établissement. Ces contributions sont pour l'essentiel celles du GTAO, Groupe de Travail sur les Archives Ouvertes au sein du consortium Couperin. OpenAire et Couperin : L'implication de Couperin dans OpenAIRE se situe dans le cadre de l'objectif 1- élaborer des structures de soutien pour les chercheurs devant déposer leurs publications (*Networking activities*)- et plus précisément des WP2 (*European Helpdesk* : établissement d'un système européen d'aide au dépôt) etWP3 (dissémination : dissémination dans les pays membres de l'Union Européenne). Son rôle est celui de bureau de liaison pour la France (*national liaison office*).
Europe **Initiative** **Infrastructure européenne** **ERIC** **MESR**	DARIAH ERIC www.dariah.fr http://www.huma-num.fr/international/dariah .	DARIAH, *Digital Research Infrastructure for the Arts and Humanities*, est une infrastructure numérique visant à développer et soutenir la recherche dans toutes les disciplines des sciences humaines et sociales. Elle s'attache de manière privilégiée aux objets numériques tels que les textes, les images, les sons, la vidéo, pour lesquels la collecte, les traitements, la diffusion, l'archivage, etc. mettent en jeu les technologies de l'information et de la communication (TIC). Plus précisément, il s'agit : de partager les infrastructures nationales, notamment les savoirs

CNRS **ABES** **Huma-NUM** **Réseau d'acteurs**		ayant conduit à les élaborer (c'est une infrastructure distribuée) ; de développer ces infrastructures et leur utilisation à travers la mise à disposition d'outils et de services ; de fournir aux communautés un meilleur accès aux matériaux de la recherche, y compris aux matériaux issus de l'héritage culturel européen ; de participer à la réflexion sur l'utilisation et l'appropriation du numérique par les différentes communautés de recherche. Le site de la participation française à l'infrastructure européenne de recherche DARIAH est ouvert depuis mars 2014. DARIAH ERIC est organisée par la France. L'Autriche, la Belgique, la Croatie, Chypre, le Danemark, l'Allemagne, la Grèce, l'Irlande, la France, l'Italie, le Luxembourg, les Pays-Bas, la Serbie et la Slovénie sont membres fondateurs.
International **Initiative** **Outil** **Catalogue d'entrepôts**	Databib http://databib.org/index.php Source complémentaire : https://bu.dauphine.fr/cadist-deco-gestion.html http://www.re3data.org/2014/03/datacite-re3data-org-databib-collaboration/	Databib est un outil de type catalogue ou annuaire, qui permet d'identifier et de localiser les entrepôts en ligne de données de recherche. Les utilisateurs et les bibliographes créent et organisent des enregistrements qui décrivent des entrepôts de données que les utilisateurs peuvent rechercher. Databib est un outil qui permet d'identifier et de localiser les entrepôts en ligne de données de recherche. Les utilisateurs et les bibliographes créent et organisent des enregistrements qui décrivent des entrepôts de données que les utilisateurs peuvent rechercher. Est-ce que les entrepôts sont adaptés au chercheur qui souhaite déposer ses données ? Comment trouver des entrepôts de données appropriées et découvrir des jeux de données qui répondent aux besoins des utilisateurs ? Comment les bibliothécaires peuvent-ils aider les utilisateurs à trouver et intégrer les données dans leurs recherches ou apprentissages ? Databib tente de répondre à ces besoins de la communauté de la recherche, y compris les utilisateurs de données, les producteurs de données, les éditeurs et les associations professionnelles, les bibliothécaires, les organismes de financement de la recherche. Databib est parrainé par *Sparks! Innovation National Leadership Grant* de l'*Institute of Museum and Library Services* (IMLS, USA). L'IMLS est, avec la *Purdue University* (USA) et *Penn State University* (USA), à l'origine du projet. L'hébergement est fourni par les bibliothèques de *Purdue University*. En Mars 2014, re3data.org et Databib ont annoncé fusionner leurs deux répertoires en un seul service, qui sera géré par DataCite d'ici fin 2015. Leur proposition conjointe à l'Assemblée générale DataCite a

		été approuvé, à l'avance de la 3e réunion plénière de la *Research Data Alliance* (RDA) à Dublin, Irlande.
International **Acteur** **Association** **Réseau d'institutions** **TIB** **Gestion d'identifiants pérennes de données**	DataCite http://www.datacite.org/ Source complémentaire : www.inist.fr	DataCite est un consortium international, dont l'administrateur général est la Bibliothèque Nationale de science et technologie allemande (TIB). Il opère en particulier comme agence d'enregistrement de DOI. Il est membre de la Fondation Internationale de DOI (IDF), instance dirigeante du système DOI. Depuis avril 2014, DataCite est affilié à la *Research Data Alliance* (RDA). Il s'appuie sur des institutions membres dans différents pays à travers le monde. Ces institutions collaborent avec des centres de données (institutions de recherche, bibliothèques…) pour attribuer des DOI à la recherche. L'Inist-CNRS est l'institution membre de DataCite pour la France. Les objectifs : Faciliter l'accès aux données de la recherche et leur intégration dans des services d'information Accroître la reconnaissance des données de la recherche comme entités importantes dans la production scientifique Promouvoir des normes et bonnes pratiques relatives à la citation des données Soutenir l'archivage de données rendant ainsi possible le contrôle des résultats de la recherche et leur réutilisation.  ***DataCite Metadata Store* (MDS) :** Création de DOI et enregistrement de métadonnées associées ***DataCite Metadata Search :*** Recherche des métadonnées associées aux jeux de données enregistrées dans DataCite ***DataCite Statistics :*** Statistiques d'enregistrement et de résolution de DOI ***DOI Citation Formatter :*** En collaboration avec CrossRef, création

		de différents formats de citation pour les DOI DataCite et CrossRef ***Content Negotiation :*** Possibilité d'obtenir des métadonnées dans des formats divers et/ou d'accéder automatiquement et directement à un objet plutôt que par la « landing page » ***DataCite Test Environment :*** Environnement de test DataCite ***DataCite OAI Provider :*** Exposition des métadonnées en OAI-PMH
International Initiative **IQSS-Harvard University** **Outil** **Application Open source**	Dataverse Network project http://thedata.org/ Source complémentaire : http://www.letudiant.fr/educpros/opinions/the-dataverse-network-project-l-open-data-selon-harvard.html *C. Gruson-Daniel,* 2013	Le projet de réseau Dataverse est logé à l'Institut des sciences sociales quantitatives (IQSS) à l'Université de Harvard. Le codage du logiciel Dataverse Network a commencé en 2006 sous la direction de Merce Crosas et Gary King. Il s'agit d'un référentiel pour les données de recherche qui prend soin de la conservation à long terme et de bonnes pratiques archivistiques, tandis que les chercheurs peuvent partager, garder le contrôle et obtenir la reconnaissance de leurs données. Il prend en charge le partage des données de recherche avec une citation de données persistantes, et permet la recherche reproductible. Le Réseau Dataverse est une application open source pour publier, partager, référencer, extraire et analyser les données de recherche. Il permet de rendre des données disponibles pour les autres, et de répliquer les travaux d'autres chercheurs. Les chercheurs, auteurs de données, éditeurs, distributeurs de données, et les établissements affiliés, sont assurés d'être crédités et cités convenablement. Un réseau Dataverse héberge plusieurs dataverses. Chaque dataverse contient des études ou des collections d'études, et chaque étude contient des informations de catalogage qui décrit les données, ainsi que les données réelles et les fichiers complémentaires. *The Dataverse Network* permet à n'importe quel chercheur de partager, archiver ses données mais aussi citer et réutiliser celles d'autres chercheurs. Initialement prévue pour accueillir des *data* en sciences sociales, la plate-forme intègre un processus complet pour rendre les informations non identifiables. En effet, le partage des données concernant des humains est un champ nécessitant des précautions spécifiques. Il est nécessaire de les rendre anonymes afin d'assurer qu'elles ne puissent être associées à la personne dont elles sont issues. *The Dataverse Network* intègre par ailleurs un certain nombre de fonctionnalités créant des interconnexions avec d'autres champs de l'*Open Science*. Il travaille notamment avec *l'Open Journal Systems* pour lier plus facilement les données aux articles scientifiques

		correspondants. Une telle infrastructure sécurisée et interconnectée constitue un levier vers de nouvelles pratiques, incitant les chercheurs à ouvrir l'accès à leurs données. La version 3 permet de choisir un catalogage selon différentes normes : DDI, Dublincore...
Anglais **Acteur** **Organisme public britannique** **JISC** **Fournisseur de services et d'outils**	DCC (JISC) http://www.jisc.ac.uk/dcc Source complémentaire : R. Gaillard, p. 66	*Digital Curation Center* En Grande Bretagne, les bibliothèques universitaires peuvent compter, dans la structuration de services autour des données de recherche, sur l'aide du *Digital Curation Center*, bras armé du JISC pour la création d'outils communs et l'accompagnement des initiatives institutionnelles. Le DCC fournit en effet aux bibliothèques des solutions « clé en main » pour assister leurs chercheurs, comme *DMPonline*, outil permettant la création en ligne de plans de gestion des données. Les bibliothèques n'ont ainsi plus qu'à s'approprier l'outil, former les chercheurs à son utilisation et, pourquoi pas, l'adapter aux besoins spécifiques de leur institution. C'est l'un des grands avantages de *DMPonline*, qui peut être personnalisé par les universités britanniques : elles peuvent notamment définir un « modèle » (*template*) institutionnel de plan de gestion, intégré à *DMPonline* mais adapté à leur politique de données. Le logiciel étant *open source*, les universités peuvent aussi l'intégrer à leur système d'information, ou « l'institutionnaliser » en l'adaptant à leur charte graphique, en y greffant leur logo ou en créant leur propre URL (*ex* : dmponline.southampton.ac.uk). À Northampton, la bibliothèque a travaillé avec le DCC pour créer un modèle institutionnel de plan de gestion et celle d'Oxford proposera bientôt aux chercheurs de l'université une version intégrée de *DMPonline*. Le DCC est membre de la *Research Data Alliance* (RDA).
International **Initiative** **Association** **University of Michigan (UM)**	DDI http://www.ddialliance.org/ Source complémentaire : Wikipedia	Data Documentation Initiative (DDI) est un projet international (programme de *l'University of Michigan* (UM), gérer par la *DDI Alliance*) initié en 1995, afin de créer et maintenir un standard de documentation technique pour décrire et conserver les informations statistiques et plus globalement les informations et données d'enquêtes en sciences humaines et sociales. En effet, la réexploitation des données d'enquête nécessite une documentation détaillée et fiable pour autoriser de nouveaux traitements statistiques. Cette documentation d'enquête est constituée d'une part des instruments de recueil des données (questionnaires et formulaires) et des référentiels qui ont permis de coder l'information (nomenclatures, dictionnaires de codes ou de

Standard de métadonnées pour les données SHS		variables). La standardisation de cette documentation et du format des fichiers qui la compose facilite à la fois la recherche (variété et richesse des modes de recherche dans les répertoires) et la réexploitation de ces données dans de nouvelles études grâce à la précision des données de contexte. Les 34 membres actuels de l'Alliance sont des institutions spécialisées de très nombreux pays, dont le Réseau Canadien des Centres de Données de Recherche (RCCDR) ou Statistique Canada, le Réseau Quetelet en France. Le Réseau Quetelet chargé en France de l'archivage et de la diffusion des données pour les sciences humaines et sociales, a réalisé en 2004 une traduction en français d'une des premières versions du standard (1.2.2). Aujourd'hui, chacun des partenaires du Réseau Quetelet - le CASD (Centre d'Accès Sécurisé Distant aux données), le CDSP (Centre de données socio-politiques de Sciences Po), le service de données ADISP du laboratoire du Centre Maurice-Halbwachs, CNRS et l'INED (Institut national d'études démographiques) - propose une documentation des données à la norme DDI (sous le logiciel Nesstar). Le CDSP (membre de Quetelet) utilise également DDI pour la description des enquêtes, tant quantitatives que qualitatives. Pour les données quanti, l'équipe projette en 2014 de passer à la version 3, qui permet la prise en compte du cycle de vie des données.
France **Initiative** **EQUIPEX** **Sciences Po** **3 instruments** **Outils de production et réutilisation de données SHS**	DIME-SHS http://www.sciencespo.fr/dime-shs/	L'équipement DIME-SHS (Données Infrastructures et Méthodes d'Enquête en Sciences Humaines et Sociales : c'est un EQUIPEX de la vague 2010) est porté par Sciences Po et réunit des partenaires de différents pans de la recherche et de l'enseignement supérieur dans des disciplines variées (sociologie, démographie, sciences économiques, science politique) : Grandes écoles : Sciences Po, le Groupe des Ecoles Nationales d'Economie et de Statistique (GENES) et Télécom ParisTech ; Institut de recherche : Institut national d'études démographiques (INED) ; Université : Université Paris Descartes ; Recherche privée : EDF R&D ; Infrastructure de recherche : GIS Réseau Quetelet. DIME-SHS vise à doter la France d'une nouvelle structure en matière de collecte, d'enrichissement et de diffusion des données pour la recherche en sciences humaines et sociales. DIME-SHS propose des ressources aux chercheurs pour produire ou réutiliser des données dont la qualité repose sur une grande rigueur

		méthodologique. DIME-SHS / Quanti : Un instrument pour les données quantitatives qui prend la forme d'un panel internet, ELIPSS (Etude Longitudinale par Internet pour les sciences humaines et sociales) ; Au sein du consortium DIME-SHS, le CDSP de Sciences Po est responsable de la mise en place du panel ELIPSS, en partenariat avec le Service des enquêtes de l'INED. DIME-SHS / Quali : Un instrument pour les données qualitatives qui prend la forme d'un site web, BeQuali (banque d'enquêtes qualitatives). Porté par l'équipe BeQuali du CDSP, il comprend plusieurs projets et partenariats : beQuali - le portail : équipe qualitative du CDSP enQuêtes - l'instrument de DIME-SHS : la banque d'enquêtes qualitatives archiPolis - le consortium de la TGIR Huma-Num : Archives des sciences sociales du politique reAnalyse - le projet ANR : projets d'expérimentation de l'analyse secondaire des données qualitatives Quali² - le site ressource : documents et débats sur les méthodes qualitatives DIME-SHS / Web : Un instrument pour les données du web qui offrira des outils pour constituer des corpus et pour les analyser. L'instrument *DIME Web* comprend une série d'outils, proposés par une équipe –celle du médialab de Sciences Po- dont le savoir-faire porte sur les méthodes numériques dédiées aux sciences sociales. Son horizon scientifique est l'étude des traces numériques, et en particulier le web, à titre de terrain d'investigation. Comme instrument, il permet d'étudier, dans un contexte maîtrisé, des phénomènes qui nous sont d'ordinaire inaccessibles en raison de leur taille, de leur dynamicité ou de leur complexité.
International Initiative **Association IDF** **Outil** **Système d'identifiant pérenne**	DOI http://www.doi.org/ Sources complémentaires : Wikipedia http://www.inist.fr/?Attribution-de-DOI	*Digital object identifier* (DOI, littéralement « identifiant d'objet numérique ») est un mécanisme d'identification de ressources numériques, comme un film, un rapport, des articles scientifiques, etc, porté par *l'International DOI Foundation* (IDF), dont DataCite et CrossRef sont membres et *registration agencies* (organismes d'enregistrement). Depuis février 2010, l'Institut de l'information scientifique et technique (INIST, du CNRS), est doté d'un statut « agence DOI » pour les données de recherche, faisant partie du consortium DataCite. Les DOI représentent une alternative aux URI. Et depuis 2012, le DOI est une norme ISO. Le DOI d'un document permet notamment une identification pérenne de celui-ci. Par exemple, il permet de retrouver l'emplacement d'un document en ligne si son URL a changé. Les DOI permettent ainsi de faciliter l'utilisation des bases de données bibliographiques, des logiciels

		de gestion bibliographique, et de produire des citations plus fiables et plus pérennes. Un DOI est un cas particulier d'identifiant Handle. C'est à la fois le mécanisme de nommage des ressources et un protocole de résolution des identifiants en adresses plus concrètes. La motivation principale pour tenter de remplacer les URI était apparemment leur manque de permanence (un URL change trop facilement et il est trop concret, trop lié à une localisation) et la motivation principale pour tenter de remplacer le DNS semble avoir été le désir d'inventer un nouveau protocole, qui n'aurait pas à supporter l'héritage, notamment administratif (le système de l'ICANN et des registres actuels) du DNS. Un DOI est dès lors unique et permanent. Le protocole de résolution, concurrent du DNS, est décrit dans la RFC 3652. La 3651 décrit le mécanisme de nommage et la 3650 l'architecture. Exemple d'identifiant Handle : **hdl:cnri.dlib/december95** Exemple d'identifiant DOI : **doi:10.1340/309registries** Selon certaines conventions, l'étiquette doi: peut être omise, ce qui donne : **10.1340/309registries** **10.XXXX** est le préfixe. Il identifie le registre ou *Naming Authority*. Le suffixe, c'est-à-dire tout ce qui est après la barre oblique / dépend du registre. DOI a donc une infrastructure sociale (registres et bureaux d'enregistrement - *Registration Agencies*) propre. Le but est, par exemple, d'assurer la persistance des identificateurs. Au bout d'un DOI, on trouve : les métadonnées (restrictions d'usage ou bien droit d'auteur, par exemple), décrites par un modèle de données commun à tous les DOI, l'*indecs Data Dictionary*, une adresse ou localisation physique (en général un URL), le traducteur cité plus haut redirige vers cet URL, diverses informations, comme l'autorité de nommage.
International Initiative **Association** **Outil** **Entrepôt de**	DRYAD http://datadryad.org/ source complémentaire : http://www.gfii.fr/uploads/docs/GFII_Hodson_Dryad.pd	Il s'agit d'un organisme à but non lucratif, de type 501(3c) US, ce qui est l'équivalent américain d'une Association Loi 1901. Il est gouverné par des membres associés (associations savantes, revues scientifiques, institutions de recherche) et un Conseil des Directeurs. La mission de DRYAD est de maintenir une archive de données qui viennent à l'appui des conclusions scientifiques des articles de recherche de qualité (revues scientifiques arbitrés). L'entrepôt DRYAD est donc ouvert aux données de tous domaines de recherche, sous conditions d'être associées avec un article de revue à comité de lecture.

données	f	
Europe **Initiative** **Cadre juridique** **CE**	ERIC Sources : http://www.dgdr.cnrs.fr/dsfim/fiscalite/Outils%20TVA/M%C3%A9mento_TVA%20.pdf http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:206:0001:0008:FR:PDF http://europa.eu/legislation_summaries/research_innovation/general_framework/ri0005_fr.htm	*European Research Infrastructure Consortium*, soit Consortium européen pour les Infrastructures de Recherche. Les progrès de la science dans le monde et le coût des investissements dans les nouvelles technologies incitent les États membres à renforcer leur coopération pour la création de grandes infrastructures de recherche. Cependant pour développer l'Espace européen de la recherche (EER) et la compétitivité, un cadre juridique applicable à ce type d'infrastructures est nécessaire afin de pallier l'absence de règles nationales ou internationales appropriées. En réponse aux demandes des États membres et de la communauté scientifique, la Commission adopte le présent règlement dont l'objectif est de faciliter l'établissement et l'utilisation communes d'installations de recherche d'intérêt européen par plusieurs États membres et pays associés au 7e programme-cadre de la Communauté pour la recherche et le développement. Le règlement (CE) n° 723/2009 du Conseil du 25 juin 2009 relatif à un cadre juridique communautaire applicable à un Consortium pour une infrastructure européenne de recherche (ERIC).instaure une base juridique unique destinée à faciliter la création et l'exploitation d'un Consortium pour une infrastructure européennes de recherche (ERIC) par les États membres et les pays associés aux programmes-cadre de recherche de la Communauté. Les États membres restent cependant les seuls responsables de l'élaboration d'un projet d'infrastructure et de la définition des aspects principaux comme les statuts, le siège, etc. Le présent cadre juridique ne s'applique qu'aux infrastructures ayant un intérêt paneuropéen. Les activités d'un ERIC n'ont pas de but lucratif afin de prévenir toute distorsion de concurrence. Toutefois, un ERIC peut exercer des activités économiques restreintes pour autant qu'elles n'entravent pas le but principal de l'infrastructure de recherche.
Europe **Initiative** **CE** **Projet FP7** **Outils**	EUDAT http://www.eudat.eu/ Source complémentaire : CINES https://www.cines.fr/wp-content/uploads/2014/02/CI	Le projet EUDAT est une initiative européenne cofinancée par le 7e Plan-cadre (Framework Programme 7 ou FP7) de la Communauté Européenne. L'objectif de ce projet est de répondre aux besoins futurs des chercheurs en matière d'accès et de préservation de données scientifiques. Pour atteindre cet objectif, tous les partenaires vont recenser les besoins et exigences des communautés de recherche afin de développer, déployer et sécuriser une infrastructure commune et persistante de données, la « Collaborative Data Infrastructure » (CDI). Face au nombre croissant d'infrastructures et à leur distribution, EUDAT apportera des services communs pour la gestion et la valorisation des données de la recherche à l'ensemble des communautés.

Services **Grille de partage, diffusion et archivage de données** **Embryon d'une e-infrastructure européenne de préservation des données**	NES_GAZETTE_Special_Archivage-3-pdfA3b-2.pdf	EUDAT comprend actuellement 25 partenaires. Il travaille particulièrement avec 5 communautés scientifiques (CLARIN, EPOS, ENES, Life Watch, VPH) et 12 grands centres de données, dont le CINES. Le projet EUDAT (d'une durée de trois ans) a démarré en 2011 (fin prévue en septembre 2014) et prévoit de fournir les services communs suivants : La réplication sécurisée des données (via iRODS avec des identifiants pérennes basés sur le système Handle/EPIC) ; Le transfert des données vers des centres de traitement ; Un service d'archivage de petits fichiers ; Un catalogue de métadonnées ; Une infrastructure d'authentification et d'autorisation (AAI). Tous ces services seront accessibles via le portail EUDAT. La philosophie d'EUDAT est de faire travailler les différents acteurs ensemble et de faire participer les utilisateurs finaux à l'élaboration des services. Le CINES participe au projet européen comme centre de ressources. Depuis 2004, les compétences et l'expertise du CINES sont reconnues dans l'archivage pérenne au niveau national. Il est considéré comme l'un des spécialistes nationaux dans les projets d'archivage à long terme pour les documents numériques de tous types. C'est à ce titre qu'il est devenu l'unique nœud français de l'infrastructure de données européenne mise en place dans le cadre du projet EUDAT. Son expérience dans l'archivage de données pérennes est un atout incontestable dans la mise en place de systèmes de stockage et d'échange de grandes masses de données pour les systèmes d'informations répartis dans le cadre du projet européen. EUDAT est membre de la *Research Data Alliance* (RDA).
France **Initiative** **CCSD, UMS** **CNRS** **INRIA**	HAL http://hal.archives-ouvertes.fr/ Source complémentaire : Wikipedia http://cache.media.enseignementsup-	La présence de cette archive ouverte dans ce glossaire a une fonction de témoin, car HAL n'est pas prévu pour héberger des jeux de données mais des publications scientifiques. Hyper articles en ligne (ou HAL) est une archive ouverte, permettant à des chercheurs de déposer leurs articles et manuscrits (pas de dépôt de jeux de données) dans une base à accès ouvert, hébergée et gérée par le Centre pour la communication scientifique directe (CCSD) du CNRS. L'accès aux données est libre, mais pas nécessairement leur utilisation ou réutilisation. Cette initiative concerne toutes les disciplines, mais les articles, dont le niveau scientifique est vérifié, ne sont pas tous évalués. Cet outil vient donc en complément de publications dans des revues à comité de

Université de Lyon **Archive ouverte française interdisciplinaire** **Publications**	recherche.gouv.fr/file/HAL/93/3/01_Convention_HAL_246933.pdf	lecture. Le 6 septembre 2005, le CNRS, l'Inserm, l'Inria, l'Inra et la Conférence des présidents d'université se sont entendus pour démarrer une phase préparatoire à un portail commun de publications scientifiques basé sur la plate-forme HAL et développé par le CCSD. HAL propose en effet un ensemble de services et d'outils aux chercheurs : transfert automatique des documents vers une archive ouverte internationale telle qu'ArXiv ou prochainement Pubmed Central lorsqu'ils appartiennent aux disciplines concernées, ce qui en augmente la visibilité internationale et l'impact ; coopération avec The European Physical Journal – EDP Sciences - système d'alerte personnalisable selon un profil défini par l'utilisateur lui-même ; création aisée d'interfaces paramétrables permettant à des institutions ou des communautés scientifiques de créer leur propre environnement ; extraction automatique de la production des laboratoires et des institutions grâce à une structure cohérente des données déposées (association auteur-laboratoire-institution) ; constitution de collections grâce à des « tampons » permettant d'authentifier la production d'un laboratoire, d'une équipe, les articles d'un journal, etc. ; extraction automatique de listes de publications pour des chercheurs ou des laboratoires sous des formats divers. Pour la physique, les mathématiques et l'informatique, certains documents placés sur HAL peuvent aussi être automatiquement déposés sur arXiv. Depuis avril 2013, une convention inter-établissements a permis de mutualiser la gouvernance de HAL. Le CCSD, jusqu'alors Unité propre de service et relevant uniquement du CNRS, devient une unité mixte de service (UMS 3368). Ses tutelles sont dorénavant le CNRS, l'INRIA qui représente les EPST et l'Université de Lyon au titre des universités françaises. Le comité scientifique et technique, que prévoit aussi le texte de la convention, a été constitué. Il permet « d'associer l'ensemble des partenaires signataires », soit 22 établissements, l'AMUE, la CGE et la CPU, liste qui peut d'ailleurs s'allonger. Il s'est réuni pour la première fois le 10 avril 2014.

International **Initiative** **Association** **CNRI** **Outil** **Système d'identifiant pérenne**	Handle http://www.handle.net/ Source complémentaire : Wikipedia Defense Advanced Research Projects Agency http://www.cnri.reston.va.us/	Le *Handle System*, ou Système Handle, est une spécification technique pour attribuer, gérer et résoudre des identifiants persistants attribués à des objets numériques et à d'autres ressources Internet. Les protocoles spécifiés permettent à un système informatique distribué de stocker des identifiants (appelés *handle* , littéralement anse) de ressources numériques et de les résoudre pour fournir l'information nécessaires à la localisation, à l'accès et à l'usage. Cette information peut être modifiée au besoin pour refléter l'état et / ou l'emplacement de la ressource identifiée sans changer le *handle*. Le *Handle System* a été développé par Bob Kahn, co-inventeur du protocole TCP/IP d'Internet, avec le soutien initial de la *Defense Advanced Research Projects Agency* (DARPA). La *Corporation for National Research Initiatives* (CNRI) gère et continue à développer le système. Le *Handle System* est actuellement utilisé dans plusieurs applications. Le système est conçu pour être fonctionnel quel que soit le nombre d'entités sans dégradation des performances. Il permet une administration distribuée, et la résolution des données courantes en multiples parties (dont chacune peut être gérée séparément). La résolution correspond au processus dans lequel une requête est posée à un réseau et reçoit en retour une réponse spécifique composée d'un ou de plusieurs éléments contenant l'information courante en relation avec l'entité identifiée: une localisation (URL) par exemple. Le *Domain Name System* (DNS) résout les noms de domaines humainement compréhensibles en adresses IP (localisation des serveurs de fichiers). Le *Système Handle* est compatible avec le DNS, mais ne l'exige pas, à la différence des systèmes d'identifiants persistants tels que PURL ou *Archival Resource Key* (ARK). D'autres différences significatives comprennent la gestion administrative distribuée, possible avec le *Système Handle* (plusieurs administrateurs gèrent différents *handles*, plusieurs administrateurs peuvent gérer un seul *handle*) et la possibilité de gérer des données de différents types. Le DNS a bien reconnu les problèmes de sécurité et de mise à jour, ce qui laisse penser que la technologie existante pourrait ne pas faire face à de nouvelles exigences. En séparant explicitement les noms de domaine de toutes les données associées, y compris la localisation, le *Système Handle* répond à une exigence clé de l'architecture Internet du futur : « il est possible de séparer la localisation et l'identification, toutes deux représentées par l'adresse IP dans l'Internet actuel, ... l'architecture résultante facilite la mobilité ainsi que la résolution d'autres problèmes du réseau actuel » d'après D. Clark, K. Sollin, et al., 2003, disponible en ligne : http://www.isi.edu/newarch/iDOCS/final.finalreport.pdf)
Europe **Initiative**	Horizon 2020 http://www.horizon2020.go	Horizon 2020 (souvent abrégé en H2020) est le nouveau programme de financement de la recherche et de l'innovation, démarré le 1er janvier 2014 pour 7 ans. Doté de 79 milliards d'euros (en euros courants, Euratom compris) pour

CE **Programme cadre** **Incitation au partage des données**	uv.fr/ http://ec.europa.eu/programmes/horizon2020/	la période 2014-2020, il rassemble les programmes de recherche et d'innovation de l'Union européenne. - C'est le successeur du 7e programme-cadre de recherche et développement technologique (P.C.R.D.T.) ou en abrégé de l'anglais FP7 (*Framework Programme 7*); - il regroupe l'actuel programme-cadre de recherche et développement technologique (7e P.C.R.D.T.), Euratom, le programme-cadre pour la compétitivité et l'innovation C.I.P., ainsi que l'Institut européen d'innovation et de technologie (I.E.T.) ; - avec ce programme, l'Union européenne financera des projets interdisciplinaires susceptibles de répondre aux grands défis économiques et sociaux. Il couvrira l'ensemble de la chaîne de l'innovation, depuis l'idée jusqu'au marché, et renforcera le soutien à la commercialisation des résultats de la recherche et à la créativité des entreprises ; - Horizon 2020 concentre ses financements sur la réalisation de trois priorités : l'excellence scientifique, la primauté industrielle et les défis sociétaux. Nouveauté du programme Horizon 2020, la Commission européenne généralise l'accès libre aux publications de recherche et introduit dans certains cas l'accès libre aux données de recherche issues des recherches qu'il aura contribuées à financer, sous peine de sanctions financières. Tous les bénéficiaires doivent assurer un accès gratuit en ligne des publications scientifiques sur les résultats du projet à tout utilisateur. Périmètre de l'obligation Toutes les publications revues par les pairs (peer-reviewed) et relatives aux résultats générés par le bénéficiaire. Les bénéficiaires sont également encouragés, dans la mesure du possible, à diffuser en Open Access toutes les données nécessaires à la validation des résultats présentés dans la publication ou le projet de publication. Ce n'est pas une obligation, contrairement aux données qui s'inscrivent dans les projets du pilote Open Research Data. Les bénéficiaires sont également encouragés à diffuser en Open Access les monographies, livres, actes de conférence, etc., publiés de manière informelle et non contrôlés par des journaux. Open Access ne signifie pas obligation de publier. La décision de publier ou non revient à l'auteur. Open Access ne signifie pas non plus renoncer à exploiter les résultats sur lesquels portent la publication.

		La protection du résultat aura lieu avant la publication. Objectif : Il s'agit de fournir un accès en ligne, large et gratuit, à toutes informations scientifiques réutilisables pour tous les utilisateurs. droit des utilisateurs : a minima droit de lecture, téléchargement et impression; droits additionnels potentiels : droit de copier, distribuer, rechercher, renvoyer vers des liens, indexer (non exhaustif). Protection des auteurs : Les auteurs ont droit au respect de l'intégrité de leur travail. Ils doivent être correctement reconnus et cités selon les standards habituels. La Commission européenne encourage les auteurs à conserver leur droit d'auteur et à recourir à des licences de type "*Creative Commons*" par exemple. Le pilote de libre accès aux données de recherche Open Research Data Il s'agit d'une opération pilote tendant à rendre accessible au plus grand nombre d'utilisateurs les données de recherche générées dans des projets financés dans le cadre du programme Horizon 2020. Les bénéficiaires qui y sont tenus doivent rendre accessibles gratuitement les données de recherche issues des projets financés. Le Work Programme définit les domaines dans lesquels le pilote est applicable. De quelles données s'agit-il ? données et métadonnées nécessaires à la validation des publications : obligatoire ; autres données et métadonnées que le bénéficiaire a choisi de diffuser en accès ouvert : spécifiées dans le plan de gestion des données ou DMP **-** **"***Data Management Plan***".** Certaines données ne pourront être rendues accessibles : cela devra être justifié dans le DMP (risque de compromettre le projet, raisons éthiques, règlementation relative aux données personnelles, propriété intellectuelle, sécurité...). Cela doit être fait au sein d'une base de données de recherche - "*research data repository*" - permettant de garantir gratuitement à tout tiers au projet : un accès, une extraction, une exploitation, une reproduction et une dissémination.

		Les utilisateurs devront être informés des outils utilisés par les bénéficiaires pour valider les résultats (les logiciels utilisés par exemple). Les données validant une publication doivent être déposées dès que possible sur la base de données de recherche choisie. Les autres données doivent être déposées selon ce qui est prévu dans le DMP. Cela concerne, en principe, les projets dans le champ d'application du pilote et tous les projets qui souhaitent y souscrire sur la base du volontariat (ce principe est qualifié "*d'opt in*"). Exception : certains projets peuvent se désengager du pilote au démarrage ou en cours de projet, sur justification. Ces raisons doivent être explicitées dans le plan de gestion des données.
France **Acteur** **TGIR** **UMS** **CNRS** **Université d'Aix-Marseille** **Campus Condorcet** **Grille de services** **Portail d'accès** **Archivage**	Huma-Num http://www.huma-num.fr/	Née en mars 2013 de la fusion du TGE Adonis et de Corpus IR, Huma-Num est une très grande infrastructure (TGIR) visant à faciliter le tournant numérique de la recherche en sciences humaines et sociales. Pour remplir cette mission, la TGIR Huma-Num est bâtie sur une organisation originale consistant à mettre en œuvre un dispositif humain (concertation collective) et technologique (services numériques pérennes) à l'échelle nationale et européenne en s'appuyant sur un important réseau de partenaires et d'opérateurs. La TGIR Huma-Num favorise ainsi, par l'intermédiaire de consortiums regroupant des acteurs des communautés scientifiques, la coordination de la production raisonnée et collective de corpus de sources (recommandations scientifiques, bonnes pratiques technologiques). Elle développe également un dispositif technologique unique permettant le traitement, la conservation, l'accès et l'interopérabilité des données de la recherche. Ce dispositif est composé d'une grille de services dédiés, d'une plateforme d'accès unifié (ISIDORE) et d'une procédure d'archivage à long terme. La TGIR Huma-Num propose en outre des guides de bonnes pratiques technologiques généralistes à destination des chercheurs. Elle peut mener ponctuellement des actions d'expertise et de formation. Elle porte la participation de la France dans le projet DARIAH en coordonnant les contributions nationales. La TGIR Huma-Num est portée par l'Unité Mixte de Services 3598 associant le CNRS, l'Université d'Aix-Marseille et le Campus Condorcet. La TGIR Huma-Num a établi un partenariat avec le CINES sur l'archivage à long terme des données numériques, en collaboration avec le centre de calcul de l'IN2P3.
France **Acteur**	INED http://www.ined.fr/	L'Institut national d'études démographiques (INED) est un établissement public à caractère scientifique et technologique. Producteur **d'**enquêtes socio-démographiques depuis sa création en 1945, l'Ined a entrepris au début des années 2000 de mettre à disposition ses enquêtes auprès de la communauté scientifique.

EPST **MESR** **Ministère des Affaires Sociales** **Catalogue de données** **Membre Réseau Quetelet**	Source complémentaire : http://www.reseau-quetelet.cnrs.fr/spip/rubrique.php3?id_rubrique=58	Les enquêtes de l'Ined traitent de thématiques variées telles que la fécondité et son évolution, les structures familiales, les relations de genre, la mobilité géographique et les migrations, les sans-abri, la santé et la mortalité, ou encore les trajectoires biographiques. Depuis juin 2012, la mise à disposition des enquêtes de l'Ined est assurée par le biais d'un serveur Nesstar, outil commun au réseau Quetelet. En outre, l'Ined assure la rediffusion de l'ensemble des enquêtes réalisées dans le cadre du projet européen GGP (Generation and Gender Programme) en plus de son propre catalogue d'enquêtes.
France **Initiative** **Huma-Num** **Outil** **Service** **Moissonneur de données**	ISIDORE Source : Huma-Num	ISIDORE (*« Intégration de services, interconnexion de données de la recherche et de l'enseignement »*) est un service de la TGIR Huma-num qui collecte, enrichit et offre un accès unifié aux documents et données numériques des sciences humaines et sociales. ISIDORE « moissonne » - c'est le terme consacré- les notices, les métadonnées et le texte intégral issus des publications électroniques, des corpus, des bases de données et des actualités scientifiques, accessibles sur le web et proposés dans des standards ouverts d'interopérabilité. ISIDORE moissonne principalement des données francophones, mais intègre aussi des données en langues étrangères produites en France ou dans le monde francophones. Enfin ISIDORE valorise les données en libre accès (*open access*). Une fois moissonnées, ces informations sont enrichies par croisement avec des référentiels métiers (listes de vocabulaires, thésaurus, référentiels) produits soit par la communauté scientifique soit par les grandes institutions du domaine de l'enseignement supérieur et de la recherche. Ces informations constituent des points d'entrée vers le texte intégral qui est lui aussi indexé quand cela est possible. Utilisant les méthodes et principes du web de données (modèle RDF) et du *linked data* (URIs), ISIDORE est le plus gros projet d'*open data* scientifique en France. Il propose l'accès à plus de trois millions de documents numériques. ISIDORE est utilisable via un site web dédié, www.rechercheisidore.fr mais il peut aussi être intégré à des portails thématiques, disciplinaires ou universitaires (utilisation d'une API et d'un SPARQL *endpoint*). ISIDORE n'est donc pas un moteur de recherche classique : c'est une plateforme de recherche modulable qui a vocation à fédérer l'accès aux données numériques de la recherche en SHS et d'offrir un accès unifié pour les enseignants, chercheurs, doctorants et étudiants.

Anglais **Acteur** **Organisme public britannique** **Politique britannique des données** **Edition de guides et d'articles scientifiques accessibles en lignes sur la gestion de données**	JISC http://www.jisc.ac.uk/ Source complémentaire : R. Gaillard, p. 31 http://www.ambafrance-uk.org/Les-publications-en-acces-libre	Fondé en 1993 par les agences de financements pour l'éducation des plus de 16 ans au Royaume-Uni, le JISC *-Joint Information Systems Committee-* est un organisme public, financé par les *Research Councils* et chargé de promouvoir auprès des universités britanniques l'utilisation des technologies de l'information et de la communication. Dans ce cadre, le JISC a mis en place un programme pour l'accès libre (« open access ») aux publications scientifiques. Devant le succès rencontré par la première année de ce programme, le JISC a décidé de le prolonger. Début janvier 2005, le comité a annoncé le nom des éditeurs qu'il a décidé de soutenir financièrement pour favoriser l'accès libre à leurs journaux. Ce financement a permis de supprimer une partie des coûts de publication pour les auteurs travaillant dans les institutions d'enseignement supérieur britanniques éligibles. S'appuyant en particulier sur son *Digital Curation Center* (DCC), fondé en 2004, le JISC pilote depuis 2009 le programme de financement *Managing Research Data*. Entré dans sa seconde phase en 2011, ce programme a permis de soutenir 17 projets d'infrastructures et de services autour de la gestion des données de la recherche à l'échelle des universités britanniques. Le rapport de la *Royal Society* de Londres, *Science as an open enterprise*, a préconisé en juin 2012 que ce programme soit prolongé et étendu au-delà des 17 premières institutions-pilotes durant cinq années, afin de favoriser l'adoption à l'échelle nationale de politiques institutionnelles de gestion des données.
France **Initiative** **Huma-Num** **Outil** **Service** **Exposition de données** **Attribution d'identifiants pérennes** **Accès** **Sécurisation**	NAKALA Source : Huma-Num	Partant du constat que de nombreux producteurs de données scientifiques ne disposent pas de l'infrastructure numérique nécessaire qui permettrait un accès persistant et interopérable à leurs données, la TGIR Huma-Num a mis en œuvre un nouveau service d'exposition de données appelé NAKALA. NAKALA propose deux grands types de services : des services d'accès aux données elles-mêmes et des services de présentation des métadonnées. Les producteurs de données numériques ainsi soulagés de la gestion purement technique, peuvent ainsi se consacrer à la valorisation scientifique de leurs données. **Les services d'accès aux données** *Un identifiant pérenne* Un identifiant unique est associé à chaque donnée et permet ainsi de citer les données indépendamment de leur localisation réelle : la technologie proposée est basée sur l'attribution d'identifiants de type *handle*, qui possèdent un mécanisme d'adressage générique, sans lien avec l'institution qui porte le service. C'est ce qui rend pérenne l'accès à la donnée, même en cas de changement ou d'évolution de l'institution qui porte le service.

		Un accès permanent L'accessibilité permanente aux données est assurée par l'utilisation de l'infrastructure de la grille de la TGIR. *La sécurisation des données* Les données sont stockées sur l'infrastructure gérée par la TGIR et sont ainsi sauvegardées régulièrement. La TGIR possède ses propres serveurs sécurisés au sein du centre de calcul de l'IN2P3-CNRS, partenaire des SHS depuis 2003. **Les services de présentation des métadonnées** *Une exposition basée sur les technologies du Web de données* NAKALA dispose d'un entrepôt RDF (*Resource Description Framework)* de type *Triple Store*, qui permet de partager les informations en utilisant les principes, méthodes et technologies du Web de données. L'utilisation de ces technologies standardisées permet de bâtir des applications de valorisation de ces données, par exemple des cartes interactives. Il sera aussi envisageable de les connecter à d'autres entrepôts existants comme DBPedia *(http://fr.dbpedia.org/)*, la version adaptée au Web de données de Wikipedia. *Un accès interopérable utilisant le protocole OAI-PMH* NAKALA permet à chaque producteur de données de disposer de son propre entrepôt OAI-PMH, le standard d'interopérabilité des métadonnées utilisé depuis 1999 dans le domaine de la recherche et de l'enseignement supérieur. Les données peuvent ainsi être « moissonnées » par des services spécialisés, comme par exemple ISIDORE, mais aussi Europeana, Gallica, etc., ce qui permettra d'en accroître la visibilité. En pratique les métadonnées descriptives seront exprimées classiquement, en utilisant le format standard Dublin Core étendu (dcterms). **Ce que ne propose pas NAKALA** NAKALA ne propose pas de moteur de recherche, ni de site Web d'éditorialisation scientifique, ni de dispositif d'enrichissement de données : en revanche, il est possible de bâtir ces outils en s'appuyant sur les services offerts par NAKALA ou bien d'utiliser ISIDORE pour enrichir ces données avec des référentiels scientifiques.
International Initiative	OAIS Sources : http://www.archivesdefranc	Il s'agit de la norme OAIS (ISO 14721:2003). Sa mise au point de l'OAIS a été pilotée par le *Consultative Committee for Space Data Systems*. Le modèle de référence pour un OAIS (*Open Archival Information System*, soit système ouvert d'archivage de l'information) décrit les

CCSDS **Outil** **Norme ISO** **Modèle conceptuel d'archivage électronique**	e.culture.gouv.fr/gerer/archives-electroniques/standard/norme-oais-iso-14721/ Sources complémentaires : http://pin.association-aristote.fr/lib/exe/fetch.php/public/documents/norme_oais_version_francaise.pdf (V1, traduite en français en 2005) http://www.archivesdefrance.culture.gouv.fr/static/4940 Wikipédia	responsabilités, les fonctions et les rapports avec son environnement d'un système d'archivage électronique pour assurer la pérennisation de l'information numérique. Il s'agit d'une norme fonctionnelle généraliste sur laquelle s'appuient de nombreuses autres normes et standards. Sa première version est parue en 2002 et est enregistrée comme norme ISO sous le numéro 2003 :17421. Une nouvelle version est en phase d'acceptation par l'ISO pour laquelle la traduction est déjà en cours. La terminologie et les fonctions définies par la norme OAIS servent de cadre conceptuel et rédactionnel aux contenus de ce site. En effet, le modèle général de cette norme répond à un besoin primordial pour tout SAE, avant même toute considération technique : l'organisation. La norme OAIS fait partie des recommandations du Référentiel général d'interopérabilité pour les systèmes d'archivage électronique. Le modèle OAIS constitue une référence décrivant dans les grandes lignes les fonctions, les responsabilités et l'organisation d'un système qui voudrait préserver de l'information, en particulier des données numériques, sur le long terme, pour en garantir l'accès à une communauté d'utilisateurs identifiés. Le long terme est défini comme suffisamment long pour être soumis à l'impact des évolutions technologiques. La norme OAIS est essentiellement constituée d'un glossaire et d'une définition des principaux concepts, des responsabilités liées à la mise en place d'une archive OAIS, de deux modèles détaillés — modèle fonctionnel et modèle d'information — ; des perspectives de pérennisation et de l'interopérabilité entre les archives OAIS. L'environnement d'un OAIS est constitué de producteurs, d'utilisateurs et de décideurs (le Management) s'échangeant de l'information. Un « Paquet d'information » contient les informations à archiver, à conserver ou à communiquer aux utilisateurs. Le paquet d'information contient toujours l'objet que l'on veut conserver, et les informations (métadonnées) nécessaires à sa préservation. Il en existe trois types, le paquet d'information à verser (SIP), le paquet d'information archivé (AIP) et le paquet d'information diffusé (DIP).
International/ Europe **Initiative** **CERN** **British Library** **ORCID**	ODIN http://odin-project.eu/ Source complémentaire : ORCID	ODIN - *ORCID and DataCite Interoperability Network* - est un projet prévu sur deux ans et initié en septembre 2012, financé par la Commission européenne au titre du 7e programme-cadre (FP7). Les partenaires du projet ODIN sont des institutions porteuses d'innovations en sciences et sciences de l'information, et des membres de l'industrie de l'édition : le CERN, la British Library, ORCID, DataCite, Dryade, arXiv et le Service national australien de données. Il s'agit de permettre l'interopérabilité des réseaux ORCID et DataCite. ODIN s'appuiera sur les initiatives ORCID et DataCite pour identifier de façon unique des scientifiques et des jeux de données, et relier ces

DataCite **Dryade** **ArXiv** **SNAD** **Outil** **Connecteur d'identifiants pérennes auteur/jeux de données**		informations au travers de multiples services et infrastructures de communication savante. Il abordera les questions centrales et encore ouvertes du sujet : Le référencement d'un objet constitué par des données Le suivi de l'utilisation et de la réutilisation Les liens entre des jeux de données, des sous-ensembles de données, des articles, des droits, et des personnes impliquées dans le cycle de vie de ces données.
Europe **Initiative** **CE** **ERC** **Projet de type CPCSA FP7** Combination of Collaborative Projects and Coordination and Support Actions **Incitation au dépôt en accès libre** **Portail web européen pour le dépôt de jeux de données et lien avec publi**	OpenAIRE et OpenAIRE+ https://www.openaire.eu/ Sources complémentaires : http://www.couperin.org/groupes-de-travail-et-projets-deap/open-access/open-aire/125-open-aire8/737-open-aire53 http://eudesk.haifa.ac.il/index.php?option=com_content&view=article&id=39&Itemid=37	OpenAIRE (*Open Access Infrastructure for Research in Europe*) est un projet européen dont le but est d'accompagner l'obligation de dépôt en accès libre (*deposit mandates*) décidée par la Commission Européenne et le Conseil Européen de la Recherche (ERC). En Europe, la décision de la Commission Européenne de rendre obligatoire le dépôt pour 20 % des recherches financées par le 7e PCRD constitue une étape très importante. Elle est venue renforcer l'obligation de dépôt rendue publique fin 2007 par l'European Research Council (ERC) pour les recherches qu'il finance. OpenAIRE a pour but d'accompagner concrètement la mise en œuvre de ces deux décisions : la publication, le 17 décembre 2007, de recommandations de l'European Research Council (ERC) demandant la mise en accès libre des résultats des recherches financées par l'ERC au plus tard six mois après leur publication ; l'annonce à l'été 2008, par la commission européenne (CE) de l'obligation de diffusion en accès libre dans un délai de 6 ou 12 mois des publications issues de 20 % des projets financés par le 7e Programme Cadre pour la Recherche et le Développement (PCRD). Baptisé Open Access Pilot, cette décision importante est un premier pas vers une obligation de dépôt plus large. Elle concerne de manière spécifique sept disciplines : santé, énergie, environnement robotique, sciences socio-économiques et humaines, programmes « infrastructures électroniques » et « science dans la société ». Afin d'accompagner ces deux décisions, la Commission a lancé au début de l'année 2009 un appel d'offre dans le cadre du 7e PCRD (partie Infrastructures) afin de mettre en œuvre « une infrastructure électronique et des mécanismes de soutien permettant l'identification, le dépôt, la consultation et la gestion des articles financés par l'ERC et le 7e PCRD ». Le projet OpenAIRE a été retenu au mois de juillet 2009. La Commission y attachait une importance particulière car de son succès

		dépendait l'extension de l'obligation de dépôt à l'ensemble des recherches qui financées par le 8e PCRD (H2020). OpenAIRE regroupe 38 institutions représentant 26 des 27 pays de l'UE (manque le Luxembourg). Le coordonnateur du projet est l'université d'Athènes. L'université de Göttingen est également particulièrement impliquée. Le projet a débuté le 1er décembre 2009 pour une durée de 36 mois (jusqu'au 1er décembre 2012). Il a été prolongé par le projet openAIREplus qui élargit son périmètre aux données de la recherche, jusque fin 2014. La Commission travaille à la pérennisation de son infrastructure. OpenAIREplus utilise l'infrastructure du projet OpenAIRE qui donne la possibilité aux scientifiques de déposer en Open Access leurs publications et leurs résultats de recherches financés par la Commission Européenne en Open Access. Grâce à OpenAIREplus, les données de recherches peuvent dorénavant être archivées comme des banques de données ou des fichiers audiovisuels, et peuvent être liées avec les publications qui y sont associées. Les chercheurs peuvent déposer leurs données dans le répertoire Open Access de leur choix ou dans l'archive orpheline qui se trouve sur le site web d'OpenAIRE. Le helpdesk OpenAIRE et les NOADs (National Open Access Desks) sont disponibles pour assister et informer tous ceux qui s'intéressent au projet ou à l'Open Access en général. Le Dr Norbert Lossau, coordonnateur scientifique d'OpenAIREplus et directeur de la bibliothèque nationale et universitaire de Göttingen en Allemagne déclare : *« La structure participative d'OpenAIREplus va guider de façon transparente le chercheur vers l'open access des données de recherches[…]».*
International **Initiative** **Association** **Outil** **Répertoire de chercheurs**	ORCID http://orcid.org/	ORCID (pour *Open Researcher and Contributor ID*) est une entreprise ouverte (statut américain 501 (3c), équivalent à une association loi 1901), participative et à but non lucratif, dont l'objectif est d'offrir un registre d'identifiants de chercheurs, ainsi qu'une méthode transparente pour relier les activités et les résultats de la recherche à ces identifiants. ORCID inclut toutes les disciplines, tous les secteurs de recherche et tous les territoires au-delà des frontières nationales, et coopère avec d'autres systèmes d'identification. C'est une plate-forme qui connecte les chercheurs et la recherche en intégrant les identifiants ORCID dans des flux de travail clés, tels que le maintien du profil de la recherche, la soumission de manuscrit, la demande de subvention et la demande de brevet. ORCID remplit deux fonctions essentielles : (1) un registre pour obtenir un identifiant unique et gérer un fichier d'activités, et (2) des API qui permettent une communication et une authentification de système à système. ORCID rend son code disponible sous une licence libre et publiera chaque année un fichier de données sous une licence CC0

		destiné à être téléchargé gratuitement. ORCID est dirigé par un conseil d'administration élu, la majorité étant à but non lucratif, composé de quatorze membres de la communauté de la recherche universitaire provenant du monde entier. Les membres du conseil d'administration sont issus et sont représentatifs des organismes adhérents à ORCID. Le conseil est responsable de l'établissement de règles générales pour la gouvernance d'ORCID, sur la base de principes essentiels parmi lesquels figurent l'ouverture et la transparence. Les identifiants ORCID sont une sorte de DOI pour les chercheurs et acteurs de la recherche.
International **Acteur** **Association/fondation d'institutions internationales et de chercheurs** **CE** **US NSF-NIST** **AGDI** **Lobbying** **Elaboration/révision d'outils internationaux** **Codes informatiques** **Guide de bonnes pratiques** **Normes**	RDA https://rd-alliance.org/ Sources complémentaires : http://hackyourphd.org/2014/04/research-data-alliance-interview-avec-odile-hologne/ http://www.capdigital.com/evenements/reunion-dinformation-research-data-alliance/ http://www.dlib.org/dlib/january14/plale/01plale.html http://www.dlib.org/dlib/january14/parsons/01parsons.html	*Research Data Alliance (*RDA) est un projet international qui doit permettre le partage des données de la recherche. RDA souhaite bâtir des ponts politiques, sociologiques et techniques pour partager les données, quelles que soient les disciplines scientifiques. RDA a été officiellement lancée en mars 2013 avec une première réunion plénière à Götteborg (Suède), où Neelie Kroes, vice-présidente de la commission européenne, a insisté sur les enjeux d'une science ouverte pour le progrès et l'innovation. La *Research Data Alliance* a été créée à l'initiative de trois institutions de pays différents (Commission européenne, *US National Science Foundation and National Institute of Standards and Technology*, et *Australian Government's Department of InnovationAustralie*). Il s'agit d'une communauté "*open science*" sans frontière de pays ou de disciplines, partageant le même objectif "*data sharing without barriers*" et une compréhension des enjeux du partage des données. RDA est basée sur la participation des individus. Les difficultés et les défis à relever sont liés notamment à ce mode de fonctionnement : il n'y a pas de moyens budgétaires alloués. Au sein de la Research Data Alliance, le travail s'organise en gouvernance internationale, avec un conseil qui oriente les actions politiques, et une petite équipe de permanent qui suit les actions. Les participants proposent les pistes de travail qui peuvent se concrétiser dans des groupes d'intérêt, des groupes de réflexions (BoF : Birds Of a Feather) ou des groupes de travail (GT) . Les 2 premiers sont des instances de réflexions. Les groupes de travail rédigent des projets (case statement) qui sont soumis à l'évaluation d'un comité technique (TAB : Technical Advisory Board). Seuls les GT ont des objectifs de résultats concrets sur une période de 18 mois. Les groupes de travail et les groupes d'intérêt sont au cœur du fonctionnement de la RDA. Les groupes de travail mènent à court terme (12-18 mois) des efforts pour élaborer et mettre en œuvre des outils spécifiques, du code, des bonnes pratiques, des normes, etc, dans de multiples institutions. Les groupes d'intérêts ont une portée plus large et une durée de vie plus

		longue. Ils travaillent à définir les enjeux et intérêts communs, qui conduisent à la création de groupes de travail plus ciblées. À l'automne de 2013, il y avait environ trois douzaines de groupes d'intérêts et de groupes de travail constitués pour examiner un large éventail de sujets, depuis les types d'identifiants pérennes à l' interopérabilité des données d'agriculture, en passant par les données toxico-génomique. Le nombre de groupes de travail et d'intérêt continue de croître rapidement.
International **(origine allemande)** **Initiative** **Association** **German Research Foundation DFG** **Outil** **Répertoire d'entrepôts de données**	Re3data.org http://www.re3data.org/ Source complémentaire : https://bu.dauphine.fr/cadist-deco-gestion.html	*Registry of research data repositories* soit un registre ou catalogue d'entrepôts de données de la recherche. Re3data.org est un registre mondial des entrepôts de données de recherche qui couvre les entrepôts des différentes disciplines universitaires. Il présente des lieux de dépôts pour le stockage permanent et l'accès à des jeux de données pour les chercheurs, les organismes de financement, les éditeurs et institutions scientifiques. Re3data.org favorise une culture de partage, un accès accru et une meilleure visibilité des données de recherche. Le registre a été créé à l'automne 2012 et est financé par la German Research Foundation (Fondation allemande pour la recherche). Les partenaires du projet re3data.org sont la *Berlin School of Library and Information Science* de la Humboldt-Universität, la bibliothèque et le service informatique (LIS) du *German Research Centre for Geosciences* (GFZ), et la bibliothèque du *Karlsruhe Institute of Technology* (KIT). Ces partenaires sont activement impliqués dans la *German Initiative for Network Information* (DINI). Certains éditeurs et revues comme Copernic Publications, PeerJ, Springer et *Nature's Scientific Data* se réfèrent à re3data.org dans leurs politiques éditoriales comme un outil pour l'identification facile des entrepôts de données appropriées pour stocker des données de recherche. L'utilisation de re3data.org est également recommandé par la Commission européenne dans ses "*Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020*" (Lignes directrices sur le libre accès aux publications et données scientifiques et de recherche pour Horizon 2020). En Mars 2013, re3data.org et Databib ont annoncé fusionner leurs deux répertoires en un seul service, qui sera géré par DataCite d'ici fin 2015. L'objectif de cette fusion est de réduire la duplication des efforts et de mieux servir la communauté de la recherche avec un registre unique et durable des entrepôts de données de recherche qui intègre les meilleures fonctionnalités des deux projets. OpenAIRE et re3data ont signé en octobre 2013 un protocole d'accord

		qui prévoit l'échange de métadonnées concernant les entrepôts de données de la recherche entre re3data.org et OpenAIREplus. Ce dernier intègrera les entrepôts indexés par re3data.org et fournira en retour des informations concernant les statistiques sur l'utilisation des jeux de données et les liens entre les jeux de données et les publications. Les échanges porteront également sur les bonnes pratiques et les normes et directives à appliquer dans le cadre de l'archivage des données de recherche.
France **Acteur** **Composant de TGIR** **UMS 3558** **CNRS-EHESS** **Réseau** **GIS** **CNRS** **EHESS** **INED** **Universités** **CCDSHS** **Gestion (dépôt archivage documentation) et diffusion de données SHS**	Réseau Quetelet http://www.reseau-quetelet.cnrs.fr/spip/ Source complémentaire : entretien avec des documentalistes de l'ADISP	Le Réseau Quetelet est une des composantes de l'infrastructure PROGEDO (PROduction et GEstion des DOnnées), la très grande infrastructure de recherche des sciences humaines et sociales, qui assure la collecte, l'archivage, la documentation et de l'accès des chercheurs à ces données. Dans le cadre de la politique du Comité des données (CCDSHS), le Réseau Quetelet coordonne les activités d'archivage, de documentation et de diffusion des données en sciences humaines et sociales du Centre Maurice Halbwachs (ADISP), du Centre de données socio-politiques et du service des enquêtes de l'Ined. L'action du réseau est organisée par Roxane Silberman, secrétaire générale du CCDSHS Il établit un lien direct entre les missions fixées par le décret D2001-139 du 12 février 2001 pour les données en sciences humaines et sociales, les données archivées par les unités partenaires et le milieu de la recherche. Plus de 1100 jeux de données sont actuellement accessibles via le Réseau Quetelet et ses partenaires. Le GIS « Réseau Quetelet » est un groupement d'Intérêt Scientifique fondé par le CNRS, l'Institut National d'Etudes Démographiques (INED), l'Ecole des Hautes Etudes en Sciences Sociales (EHESS), l'université de Caen-Basse Normandie, l'Université des sciences et technologies de Lille 1, la Fondation nationale des sciences politiques, l'Ecole d'Economie de Paris (EEP) et L'Institut de Recherche et Documentation en Economie de la Sante (IRDES), afin d'**a**ssurer la coordination des orientations proposées par le CCDSHS. Il a pour mission l'appui à la collecte, à la documentation, à la préservation, à la promotion d'un vaste ensemble de données françaises nécessaires aux disciplines des sciences humaines et sociales dans un cadre européen et international. Le GIS gère notamment les modalités de coordination et de coopération entre les établissements partenaires pour les aspects scientifiques (archivage, documentation et diffusion des données), pour l'administration du portail national ainsi que pour les représentations internationales de ces activités dans les structures européennes ou internationales du domaine.
France **Acteur**	TGIR PROGEDO http://www.progedo.fr/	L'infrastructure PROGEDO (PROduction et GEstion des DOnnées), très grande infrastructure de recherche des sciences humaines et sociales, assure la collecte, l'archivage, la documentation et l'accès des chercheurs à ces données. Cette infrastructure, inscrite dans la feuille de route française sur les TGIR depuis décembre 2008, est portée officiellement par le CNRS pour

TGIR **UMS 3558** **CNRS** **Ministère ESR** production et mise à disposition de données pour les SHS **Réseau Quetelet** **CASD** **Enquêtes**	Source complémentaire : http://www.reseau-quetelet.cnrs.fr/	le compte du Ministère depuis 2012. Destiné à doter la France d'une infrastructure comparable à ses équivalents européens, PROGEDO est impliqué dans trois consortium européens (statut ERIC) constitués ou en cours de constitution autour des banques de données, (CESSDA - Council of European Social Science Data Archives) et des enquêtes européennes (ESS et SHARE). PROGEDO assure deux missions principales : le développement d'un dispositif d'archivage, de documentation et de mise à disposition des données pour la recherche en sciences humaines et sociales (réseau Quetelet), comprenant notamment un dispositif d'accès sécurisé à distance pour les données confidentielles (CASD), l'organisation et la contribution au financement de grandes enquêtes pluridisciplinaire d'intérêt national, notamment aux grandes enquêtes retenues sur la feuille de route des infrastructures européennes (ESFRI). Pour ses fonctions d'archivage, de documentation, et de diffusion contrôlée des données, PROGEDO s'appuie sur le réseau Quetelet et ses unités partenaires. 3 EQUIPEX (CASD, D-FIH et DIME-SHS) contribuent également au développement de certaines composantes de l'infrastructure. Le TGIR PROGEDO appuyé sur une Unité Mixte de Service Quetelet-PROGEDO (placée sous la responsabilité conjointe du CNRS et de l'EHESS), assure la mise en œuvre d'une politique publique de données pour les sciences humaines et sociales dont les grandes orientations sont proposées par le Comité de Concertation pour les Données en SHS (CCDSHS) et coordonnées au sein d'un GIS « réseau Quetelet », en partenariat avec plusieurs grands établissements.
Europe **Initiative** **CERN** **OpenAire** **CE** **Outil** **Entrepôt de publi et de données**	ZENODO http://zenodo.org/	Entrepôt hébergé par le CERN ; créé par le CERN et OpenAire, financé par la CE. ZENODO construit et exploite un service simple et innovant qui permet aux chercheurs, scientifiques, projets et institutions de l'UE de partager et de mettre en valeur les résultats de recherches multidisciplinaires (de données et publications) qui ne font pas partie de dépôts institutionnels ou des différentes matières existantes des communautés de recherche. ZENODO permet aux chercheurs, scientifiques, projets de l'UE et institutions de : -Facilement partager la "longue traine" des résultats de recherches modestes dans une grande variété de formats, y compris texte, feuilles de calcul, audio, vidéo, et images dans tous les domaines de la science. -Afficher les résultats de leur recherche et se créditer en rendant les résultats de la recherche citables et en les intégrants dans des lignes de *reporting* existants au sein d'organismes comme la Commission

		européenne. -Accéder facilement et réutiliser les résultats de recherche communs. Les livrables : -Un entrepôt numérique ouvert à tous, pour tout ce qui n'est pas desservi par un service dédié ; ce que l'on appelle la «longue traine» des résultats de recherche. -L'intégration avec l'infrastructure OpenAIRE, et l'inclusion assurée dans le corpus OpenAIRE. -Un téléchargement facile et l'achèvement de métadonnées semi-automatique par la communication avec les services en ligne existants, tels que DropBox pour le téléchargement, Mendeley / ORCID / CrossRef / OpenAIRE pour le téléchargement et les métadonnées pré-renseignées. -Un accès facile aux résultats de la recherche par l'intermédiaire de visualisation innovante, ainsi que des API ouvertes et l'intégration avec les services en ligne existants et la préservation des formats de données indépendants. -Un service sûr et fiable fournissant une curation judicieuse, des archives et stratégie de conservation numérique selon les meilleures pratiques. -La création d'identifiant pérennes (comme par exemple les DOI) pour les résultats de recherche partagés. -Un service d'hébergement selon les meilleures pratiques de l'industrie des centres de données professionnels du CERN. -Des moyens de lier les résultats de recherche avec d'autres résultats, les sources de financement, les instituts et les licences.

Annexe 2 : Modèle de Data Management Plan (DMP) ou Plan de gestion des données (PGD).

Ce document est la reproduction d'une trame diffusée par le site CoopIST, elle-même adaptée de la *Checklist for a Data Management Plan* du DCC britanique (disponible à cette adresse : http://www.dcc.ac.uk/resources/data-management-plans), et de l'outil en ligne de cette institution (mis à disposition à cette adresse : (https://dmponline.dcc.ac.uk/). Il s'agit donc du récapitulatif des informations que doit contenir un DMP, à adapter en fonction de chaque cas particulier de projet de recherche.

Exemple de trame d'un plan de gestion de données (PGD)

1) Informations administratives

- nom et identifiant du projet
- description du projet- agence(s) de financement
- nom et identifiant éventuel du responsable principal de projet
- contact pour les données de projet
- date de la 1re version
- date de la dernière mise à jour
- politiques associées au projet, incluant les instructions ou recommandations de l'agence de financement et de l'institution

2) Collection de données

- description des données, incluant le type de données, le format et le volume
- jeux de données existants qui seront utilisés
- méthodes de collecte et de création des données
- système d'organisation, de nommage et de gestion des répertoires et des fichiers
- processus d'assurance qualité mis en œuvre

3) Documentation et métadonnées

- informations nécessaires pour lire et interpréter les données
- organisation de la collecte et de la documentation
- standards de métadonnées adoptés

4) Ethique, cadre légal

41- Ethique

- détails de l'accord de conservation et de partage des données
- étapes pour la protection de l'identité des participants
- étapes pour la sécurité du stockage et du transfert de données sensibles

42- Droits de propriété intellectuelle et copyright

- nom de(s) propriétaire(s) des données
- licence(s) pour la réutilisation des données (par exemple, une des licences *Creative Commons* ou *Open Data Commons*)
- restrictions d'utilisation par les tierces parties
- délai requis pour le partage de données (embargo lié à la publication dans une revue ou délai d'application d'un brevet)

5) Stockage, sauvegarde, et sécurité

51- Stockage et sauvegarde

- lieu de stockage des données
- plan de sauvegarde
- personne ou équipe responsable de la sauvegarde
- procédures de récupération

52- Sécurité

- risques et leur gestion
- dispositif d'accès
- dispositif éventuel pour le transfert sûr et intègre des données collectées sur le terrain

6) Sélection et conservation

- informations détaillées sur les données qui seront retenues, partagées et/ou conservées, et référence aux obligations contractuelles, légales ou réglementaires
- utilisations de recherche prévues des données
- durée de conservation des données au-delà du projet
- entrepôt ou archive de conservation des données et responsabilités associées
- temps et effort nécessaires à la préparation des données pour leur conservation et leur partage

7) Partage des données

- étapes à mener pour faciliter la prise de connaissance (*discovery*) des données par les autres
- conditions de restriction du partage des données et détails de leur application dans l'accord de partage de donnés
- mécanisme de partage de données (via un entrepôt, sur demande expresse ou tout autre processus)
- délai de publication
- procédure éventuelle d'obtention d'un identifiant persistant pour les données

8) Responsabilités et moyens

- nom de la personne responsable de la mise en œuvre du plan de gestion de données
- nom de la personne responsable de chaque activité de gestion des données
- équipements et logiciels requis (en addition à ceux existants fournis par l'institution)
- besoins additionnels d'expertise ou de formation
- charges imposées par les entrepôts de donnée

Annexe 3 : Fiche 7 du DMST. Déposer vos publications et vos rapports de littérature grise sur le portail MADIS

Ce document est une reproduction de l'outil que le service documentation de l'IFSTTAR (le DMST) met à disposition, dans l'Intranet de l'institut. Il s'agit d'un mode d'emploi permettant aux chercheurs de cet établissement de déposer leurs travaux (publications et rapports) dans le fonds documentaire. Les documentalistes utilisent ces informations pour alimenter le catalogue informatisé de l'IFSTTAR nommé MADIS. C'est de ce document que nous nous sommes inspiré pour créer deux « guides du déposant de jeu de données », disponibles en annexe 6 et 7.

Afin que vos publications et vos rapports internes soient référencés dans MADIS vous devez IMPERATIVEMENT en faire le dépôt

Voir la note de la Direction scientifique du 29 mars 2011 :

http://communication.ifsttar.fr/fileadmin/Direction_Scientifique/Documentation/Depot_des_productions_scientifiques_mars_2011.pdf

Vous devez pour chaque dépôt, remplir plusieurs champs de description :

- type de production : publication / littérature grise
- type de document : ex. compte rendu d'essai ; article à comité de lecture,…
- le titre
- le résumé
- vos mots-clés
- numéro d'opération de recherche le cas échéant
- pour les publications, cocher la case indiquant si le document attaché à la notice peut être versé dans HAL
  - ⇨ Il faut pour cela vous assurer que vous pouvez légalement déposer votre publication sur un serveur d'archive ouverte (cliquez sur le lien qui vous renverra une page d'aide et le site officiel des droits concédés par chaque éditeur : SHERPA/RoMEO)
- cocher la case « confidentiel » si nécessaire

Puis attacher obligatoirement le document (qui pourra ainsi à minima être consulté en interne)
Les notices notées confidentielles (et le document attaché) ne seront visibles que par les documentalistes

La notice sera contrôlée et complétée par le service documentation et également versée, pour les publications, dans l'archive nationale HAL
Une notice incomplète vous sera renvoyée pour modification et retour au service documentation.

Pour plus d'information sur la procédure à suivre, merci de vous reporter au guide du dépôt disponible sur Madis, rubrique « déposer » (attention de bien vous identifier auparavant)
et sur les pages intranet du service DMST
http://intranet.ifsttar.fr/fonctions-transversales/documentationist/madis-mise-a-disposition-de-linformation-scientifique/

Contact : **http://madis.ifsttar.fr** ; rubrique à l'accueil « votre service documentation »

ou

http://intranet.ifsttar.fr/fonctions-transversales/documentationist/organisation/

Annexe 4 : Ebauche guide du déposant

Voici une première version du « guide du déposant de jeu de données », élaborée à partir de plusieurs documents du même type : le guide du dépôt de publications et rapports en usage à l'IFSTTAR (annexe 3), le guide du déposant de données actuellement utilisé par le CDSP, et le *Codebook* de la norme de métadonnées DDI.

C'est cette version qui a été testée par les trois chercheurs de l'IFSTTAR qui ont bien voulus participer à ce projet. Nous les avons observés lors de ces tests, en notant leurs réflexions comme leurs réponses aux questions posées par le document. Ces réponses et réflexions ont été consignées et comparées dans un tableau disponible en annexe 5.

Guide du déposant de données IFSTTAR pour le Dataverse

Attention : Ce guide est destiné à la description de données de type statistique. Pour tout autre type de données, un autre formulaire de description est disponible. Néanmoins, certaines questions ne sont pertinentes que pour un type très spécifique de données : seuls les champs en rouge sont à renseigner pour tout jeu de données.

Une banque de données vit par la qualité et la quantité des apports faits par les chercheurs. Le DMST vous propose à cet effet un guide du déposant pour faciliter le dépôt de vos données.

Le formulaire permettra aux documentalistes de préparer la description de l'enquête.

Le guide du déposant se décompose en 2 grandes parties :

1/ Description de l'étude

Titre / publication / auteur / production / distribution / séries / bibliographie

Mots-clefs / résumé / aspects chronologiques et géographiques

Collecte / méthode / échantillonnage poids / nettoyages / évaluation

Données / accès / confidentialité / restrictions

2/ Description des fichiers

1/Description de l'étude

Titre/auteur/production/distribution/séries/bibliographie

Titre : Titre complet du jeu de données. Le titre complet doit indiquer la couverture géographique aussi bien que la période temporelle couverte.

.....
.....
.....

.....
.....

Sous titre : Titre secondaire utilisé pour souligner certaines limitations du titre principal. Il peut répéter des informations déjà présentes dans le titre principal.

.....

.....
.....

Publication : Publication(s) à laquelle le jeu de données a donné lieu, basé sur les données. Cela peut prendre la forme de références bibliographiques. Au minimum, indiquer le titre, l'auteur, la date, le producteur/diffuseur de la (ou des) publication(s).

.....
.....
.....
.....

Auteur : Personne, morale ou physique, responsable du contenu du jeu de données (possibilité de plusieurs auteurs). Le patronyme doit être indiqué en premier.

.....

.....
.....
.....

Producteur : Personne ou organisme qui a assuré la responsabilité financière ou administrative pour la réalisation matérielle du jeu de données (Possibilité de mentionner ici les différents types de participation dans le processus de production).

.....
.....
.....

Licence/copyright : Droits de reproduction attachés au jeu de données.

.....
.....

Date de production : Date de production des données- pas la date de distribution ni d'archivage -. Format AAAA-MM-JJ.

.....

Lieu de production : Adresse de l'organisme qui a produit les données.

.....
.....
.....

Logiciel de création : Identifie le logiciel utilisé pour créer ou pour archiver les données. Si besoin mentionner le numéro de version.

.....
.....

Source de financement : Source(s) de financement pour la production des données.

.....
.....
.....

Distribution : Organisme(s) désigné(s) par l'auteur ou par le producteur pour effectuer des copies des données.

.....
.....
.....

Contact : Nom et adresse de la personne à contacter comme personne ressource en cas de problème ou de question soulevés par les utilisateurs.

.....

Dépositaire : Nom de la personne (ou de l'institution) qui a déposé le jeu de données.

.....
.....
.....

Date de dépôt: Date du dépôt du jeu de données. Format AAAA-MM-JJ

.....

Série : Nom de la série à laquelle le jeu de données appartient

.....
.....
.....

Informations sur la série : Historique de la série et récapitulatif des informations qui s'appliquent à la série entière.

.....
.....
.....
.....
.....

Citation bibliographique : Eléments permettant de constituer une citation bibliographique complète du jeu de données. Ex. : INSEE –Enquête emploi, 1965

.....
.....
.....

Mots-clefs / sujet / aspects chronologiques et géographiques

Thèmes principaux : Mots ou phrases décrivant le contenu des données

.....
.....
.....

Classification : Mots décrivant les thèmes plus généraux couvrant les données

.....
.....
.....

Résumé/sujet : Le résumé décrit le but, la nature et les limites de la collection de données, les caractéristiques de contenu, les principaux sujets couverts, à quelles questions on a tenté d'apporter des réponses. Une liste des principales variables peut être apportée ici.

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

Période d'étude : Période de recueil ou d'observation couverte par les données.

.....
.....
.....

Date (s) de collecte : Date(s) de collecte des données. Spécifier le cas échéant, date de début, date de fin, ou encore le cycle ou la vague. Format AAAA-MM-JJ.

.....
.....

Pays de l'étude : Indique le(s) pays couvert(s) dans le fichier.

.....
.....
.....

Couverture géographique : Information sur la couverture géographique : totalité des limites géographiques des données ainsi que tous les niveaux géographiques complémentaires pouvant être codés dans les variables. Exemple : la région Rhône-Alpes, le département de la Savoie, l'Europe.

.....
.....
.....

Unité géographique : Plus petit niveau géographique couvert par les données. Exemple : commune, canton, département, région, pays, etc.

.....
.....
.....

Individus : Unité de base pour les analyses ou les observations décrites : individu, famille/ménage, groupe, vélo, station, institution/organisation, unité administrative, etc.

.....
.....
.....

Univers: Description de la population couverte, des groupes de personnes ou autres éléments qui constituent l'objet de l'étude ou de l'enquête et auxquels les résultats se réfèrent.

.....
.....
.....
.....

Types de données : Type de données contenues dans le fichier : données d'enquête, données de recensement, données agrégées, code source de programme, texte lisible en machine, données administratives, données d'expérience, test psychologique, données textuelles, texte codé, document codé, agenda d'activité, données d'observation, données produites par traitement etc.
.....
.....

Collecte / méthode / échantillonnage / poids / nettoyages / évaluation

Méthode temps : Dimension temporelle de l'enquête. Exemples : enquête panel, étude de tendance, séries de temps.

.....
.....
.....

Collecteur : Entité (individu, organisme ou institution) responsable de l'administration du questionnaire, de l'entretien ou de la compilation des données.

.....
.....
.....
.....

Fréquence de collecte : Si les données collectées comprennent plus d'un point dans le temps, indiquer la fréquence à laquelle les données sont collectées.

.....
.....
.....
.....

Echantillonnage: Type d'échantillon et plan de sondage utilisés pour sélectionner les enquêtés représentatifs de la population étudiée. Des indications sur la taille de l'échantillon cible et sur des fractions de l'échantillon peuvent être mentionnées :

-quotas : préciser la nature des quotas et leur qualité.

-aléatoire : préciser ici la base ayant servi au tirage, la méthode aléatoire retenue.

-spécifique : préciser le type de méthode retenu.

-panel : mêmes individus interrogés dans le temps.

-autre méthode (préciser le plus possible)

.....
.....
.....
.....
.....
.....

Explication de biais : Montre la correspondance aussi bien que les différences entre l'échantillon obtenu et les statistiques disponibles pour une population entière (âge, sex-ratio, statut matrimonial etc.).

.....
.....
.....
.....
.....

Mode de collecte : Méthode utilisée pour collecter les données ; caractéristiques d'instrumentation :

-dans le cas où un seul mode d'administration du questionnaire a été utilisé, préciser parmi : face à face / téléphonique / auto-administré sans enquêteur / auto-administré avec enquêteur présent / auto-administré par journal / autre.

-Dans le cas où plusieurs modes d'administration du questionnaire ont été utilisés au cours de la même enquête, veuillez préciser ici le plus possible. Ex : les questions 1 à 75 ont été administrées en face à face.

.....
.....
.....
.....
.....

Type instrument de collecte : Type de questionnaire ou d'outil utilisé.

Ex : "Directif", "Semi-directif", "Structuré", "Semi-Structuré", "Non-structuré" …

.....
.....
.....

Environnement de collecte : Utilisé pour décrire des aspects notables de la collecte des données.....
.....
.....
.....

Contrôles de collectes : Méthodes employées pour faciliter le contrôle des données opéré par l'enquêteur primaire ou par le centre d'archivage des données. Préciser les programmes spécifiques utilisés pour l'opération.

.....
.....
.....
.....

Poids : Décrire ici les critères d'utilisation des pondérations employées dans l'analyse d'un ensemble de données

.....
.....
.....
.....

Préparation / Nettoyage : Méthode utilisée pour "nettoyer" les données, c'est à dire vérifications de cohérence, vérifications de codes aberrants, etc.

.....
.....
.....
.....

Informations supplémentaires méthode : Informations complémentaires concernant la méthodologie et les traitements impliqués, ou le taux de réponse, l’estimation d’erreur d’échantillonage, ou un autre problème concernant l’évaluation des données.

.....
.....
.....

Données / accès / confidentialité / restrictions

Archive d’origine : Centre d’archivage auprès duquel les données ont été obtenues; centre d’archivage d’origine.

.....
.....

Types de fichiers disponibles : Récapitule le nombre de fichiers existant pour une collection, le nombre de fichiers contenant des données et si la collection contient une documentation électronique et/ou d’autres fichiers et informations supplémentaires tels que dictionnaire de données, définitions de données, ou instrument de collecte.

.....
.....
.....
.....
.....

Nombre de fichiers : Nombre total de fichiers associés à la collection.

.....

Déclaration de confidentialité : Utilisé pour préciser s'il est nécessaire de signer un engagement de confidentialité pour accéder aux données. .

.....
.....
.....

Permission spéciales : Cet élément est utilisé pour indiquer si une permission spéciale est nécessaire pour accéder à la ressource.

.....
.....
.....

Restrictions : N'importe quelle restriction d'accès ou d'usage des données telles que certification privée ou restrictions de diffusion doivent être mentionnées ici. Les restrictions peuvent être exigées par l'auteur, le producteur ou le diffuseur des données. Si la restriction ne porte que sur une sorte d'utilisateur, préciser le type.

.....
.....
.....

Contact : Personne ou organisation (avec l'adresse complète et le n° de téléphone) qui contrôle l'accès aux données si elle est différente de celle du diffuseur des données. Possibilité de mentionner l'adresse mail du contact.

.....
.....

Conditions d'accès : Indiquer toute information complémentaire pouvant aider l'utilisateur à comprendre les conditions d'accès et d'utilisation des données.

.....
.....
.....
.....

Dénégation de responsabilité : Information concernant la responsabilité d'utilisation des données.

.....
.....
.....
.....

Notes sur l'accès aux données : Indications complémentaires sur l'accès et l'utilisation des données.

.....
.....
.....

Etudes apparentées : Information sur les relations entre l'étude décrite et les autres (précédentes, suivantes, autres vagues ou autres cycles) ou avec d'autres éditions du même fichier. Doivent être inclus les noms des ensembles additionnels de données engendrées ainsi que les autres collections sur le même thème. Cela peut prendre la forme de références bibliographiques).

.....
.....
.....
.....

Autres références : Indications sur les autres références pertinentes. Cela peut prendre la forme de références bibliographiques

.....
.....
.....
.....

2/ Description des fichiers

Nom : Contient un titre court qui sera utilisé pour distinguer chaque fichier des autres fichiers de la même collection. Au format NOM.EXTENSION

.....
.....
.....
.....

Contenu : Résumé ou description du fichier. Description sommaire du but, de la nature et du champ du fichier, des caractéristiques spéciales de son contenu, des principaux thèmes couverts, et à quelles questions les investigateurs ont essayé de répondre en créant le fichier. Une liste des principales variables est importante ici.

.....
.....
.....
.....
.....

Observations : Nombre d’observations du fichier

.....

Variables : Nombres de variables du fichier

.....

Type de fichier : Type de fichier (ASCII, EBCDIC, etc.) et le logiciel associé comme les 'fichiers de données SAS', les 'fichiers export SPSS', etc.

.....
.....
.....

Format de fichier : format physique du fichier : format de la longueur de l'enregistrement logique, format délimité, format libre, etc. : délimité (, / ; / | ...) / non-délimité / tabulé

.....
.....
.....

Vérifications : Décrire ici au niveau de chaque fichier, quels ont été les types de contrôles et les opérations faites sur le fichier de données.

.....
.....
.....

Logiciel de génération du fichier : Logiciel utilisé pour créer le fichier. L'attribut "version" permet de spécifier le numéro de la version du logiciel

.....

Annexe 5 : Grille de lecture des tests pour le format issu de DDI

Ce tableau constitue un document de travail. Les lignes représentent les questions posées par le guide du déposant, et les colonnes figurent les trois chercheurs ayant répondus à ces questions.

C'est à partir de ce tableau qui permet de comparer, questions par questions, les difficultés et les usages rencontrés, que nous avons pu établir une version épurée et finale du guide, disponible en annexe 6.

Modèle de départ	**Chercheur 1**	**Chercheur 2**	**Chercheur 3**	**Bilan**
Titre	Soucis avec la mention de couverture géographique dans le titre, quand c'est national. Difficulté car données qui sont un réajustement d'une enquête nationale transport. Que mettre en premier dans le titre ? Globalement, trouve difficile de donner un titre à un jeu de données. J'interviens pour proposer un titre qui indique le travail effectué sur les données, ici un nettoyage et une « procédure d'imputation », et pas seulement les données « brutes » ou de l'enquête initiale.	OK	OK	Modifier les consignes : non pas « doit indiquer » mais « peut indiquer »
Sous-titre	Pas utilisé mais pourrait servir à mentionner que le jeu de données a été produit dans le cadre d'un projet ? Rendre ça obligatoire ? Envisager un champs multivalué ou répétable.	OK	Pas utilisé	Garder
Publication	OK	OK	OK. Mentionner qu'il peut y en avoir plusieurs.	Modifier les consignes pour signaler qu'il peut y en avoir plusieurs

Modèle de départ	Chercheur 1	Chercheur 2	Chercheur 3	Bilan
Auteur	Règles d'écriture ?	Règles d'écriture ?	OK. Petite difficulté pour savoir si l'auteur est le producteur des données brutes ou le chercheur qui retravaille les données. Après discussion, décision de mettre les deux fichiers et donc les deux auteurs.	Etablir des règles d'écriture (nom de famille en premier, en majuscules)
Producteur	OK	OK	OK	Garder
Licence/copyright	Proposer un menu déroulant ?	Vérifier si les données sources sont « libres », ou si licences. Propose la CC0, de base, pour nous.	OK. Attention, données uniquement diffusable en interne : accord avec le producteur des données brutes.	Faire un choix de licence ouverte « par defaut » Hélas, menu déroulant pas possible dans Dataverse. Mais liste possible dans le doc « Guide du déposant ». Rapprocher des champs concernant les droits.
Date de production	OK mais propose Année uniquement	OK mais propose Année uniquement	OK	Noter dans les consignes que c'est à minima l'année, en format AAAA
Lieu de production	OK	Pas utilisé. Propose adresse IFSTTAR par défaut	OK	Mettre IFSTTAR par défaut dans le Dataverse, mais pas dans le « Guide du déposant »
Logiciel de création	OK	Pas utilisé. Dit que redondant avec partie description fichier.	OK	C'est redondant : garder ça plutôt pour la partie fichier

Modèle de départ	Chercheur 1	Chercheur 2	Chercheur 3	Bilan
Financement	Pas utilisé mais a du sens pour d'autres jeux de données. Propose de le renommer « opération de recherche » ou « projet de recherche », de le placer plus tôt dans le guide car le projet doit être un nœud.	Pas utilisé	Pas utilisé	Garder car le « projet » pourrait devenir une notion importante
Distribution	OK. Ça pourrait être IFSTTAR par défaut ?	Pas utilisé	Pas utilisé	Supprimer ou mettre IFSTTAR par défaut ?
Contact	OK : Le chercheur. Adresse électronique. Doit-on y mettre une adresse plus générique, celle du labo ? Discussion car que faire quand le chercheur s'en va ?	Pas utilisé. Propose qu'on rajoute « si différent du dépositaire ».	OK Le chercheur, nom et mail	Préciser qu'il s'agit du chercheur, son nom puis son mail. Regrouper si on demande d'autres contacts.
Dépositaire	Pas utilisé	OK Le chercheur. Nom prénom. Propose qu'on demande le mail là. Et qu'on rende le champ obligatoire.	OK Le chercheur. Prénom Nom.	Garder mais préciser les consignes (ça a du sens surtout quand plusieurs auteurs et que l'un d'eux dépose, ou si c'est un documentaliste qui effectue le dépot). Règles d'écriture.
Date de dépôt	Date du jour	Date du jour. Propose qu'on rende le champ obligatoire.	OK Date du jour	Supprimer du « guide » car c'est automatiquement la date du jour de catalogage dans le Dataverse ? ou garder comme date de dépôt mais du questionnaire ?
Série	Pas utilisé	OK	Pas utilisé Mais est-ce collection ou série ?	Fusionner avec « info sur la série » et inclure la notion de collection.
Info sur la série	Pas utilisé	OK mais bref, et peut être redondant avec le champ « série »	Pas utilisé	A fusionner avec le précédent

Modèle de départ	Chercheur 1	Chercheur 2	Chercheur 3	Bilan
Citation	Pas utilisé	OK mais question : citation du jeu déposé ou du jeu utilisé comme source ?	Pas utilisé	Etablir règles d'écriture de citation et bien spécifier dans les consignes qu'il s'agit du jeu de données qu'on est en train de déposer. Rapprocher ce champ de celui qui concerne d'éventuels autres jeux de données ou publi utilisés comme source.
Thèmes principaux	OK	OK	OK	Echangé l'ordre avec champ suivant et intituler ça « Mots-clés »
Classification	Pas utilisé	OK mais plus difficile	Pas utilisé	Echanger l'ordre avec champ précédent et intituler ça « Thématique principale » ou « Thème »
Résumé	OK	OK. Mais propose de mettre ce champ plus haut.	OK	Garder
Période d'étude	Redite puisque c'est comme pour l'étude nationale transport que ça retravaille	OK	OK	Garder. Fusionner avec date de collecte ?
Date de collecte	OK	Pas utilisé	Pas utilisé	Supprimer et fusionner avec période d'étude ?
Pays de l'étude	Redite puisque c'est comme pour l'étude nationale transport que ça retravaille	OK	Pas utilisé	Supprimer : redondant avec « couverture géographique »
Couverture géographique	Redite puisque c'est comme pour l'étude nationale transport que ça retravaille	OK	OK	Garder

Modèle de départ	Chercheur 1	Chercheur 2	Chercheur 3	Bilan
Unité géographique	Redite puisque c'est comme pour l'étude nationale transport que ça retravaille	OK	Pas utilisé	Supprimer, « couverture géographique » suffit
Individus	OK	Pas utilisé	OK	Garder
Univers	OK	OK	OK	Garder
Type de données	OK	OK	OK mais malentendu : note le type de fichier.	Garder. Déplacer, soit très tôt dans le guide, soit avec la description des fichiers
Méthode temps	Pas utilisé	Pas utilisé, aucun intérêt.	Pas utilisé	Supprimer
Collecteur	Pas utilisé	Pas utilisé. Pour ces champs méthodo, propose qu'on dise que si le jeu de données a été construit par retraitement de données existantes, on passe au chapitre suivant.	Pas utilisé	Supprimer
Fréquence de collecte	Pas utilisé	Pas utilisé. Propose « dates de collecte ».	Pas utilisé	Supprimer
Echantillonnage	Pas utilisé	Pas utilisé	Pas utilisé	Supprimer
Explication de biais	Pas utilisé	Pas utilisé. Propose qu'on supprime.	Pas utilisé	Supprimer
Mode de collecte	Pas utilisé	Pas utilisé	Pas utilisé	Supprimer
Type instrument de collecte	Pas utilisé	Pas utilisé. Propose qu'on regroupe avec « Mode de collecte ».	Pas utilisé	Supprimer
Environnement de collecte	Pas utilisé	Pas utilisé.	Pas utilisé	Supprimer
Contrôle de collecte	Pas utilisé	Pas utilisé	Pas utilisé	Supprimer
Poids	Pas utilisé	Pas utilisé. Propose qu'on dise « Pondération »	Pas utilisé	Supprimer
Préparation/nettoyage	Pas utilisé	Pas utilisé. Propose qu'on enlève Préparation.	OK	Garder

Modèle de départ	Chercheur 1	Chercheur 2	Chercheur 3	Bilan
Info supplémentaires méthode	Pas utilisé	Pas utilisé	Pas utilisé	Supprimer
Archives d'origine	Utilisé pour décrire l'archive d'origine des données d'origine, c'est-à-dire de l'enquête transport 2008	Pas utilisé. Trouve ça pas clair.	OK	Préciser dans les consignes ? ou supprimer ? ou garder pour les données/études sources ?
Types de fichiers disponibles	Juste un nombre de fichier, du coup c'est redondant avec le champ suivant	Pas utilisé. Propose déplacement vers description des fichiers.	OK	Supprimer car redondant (utile seulement si on fournit les mêmes données sous plusieurs formats)
Nombre de fichiers	OK	Pas utilisé. Propose déplacement vers description des fichiers.	OK	Déplacer vers description des fichiers
Déclaration de confidentialité	Redondant avec permissions spéciales et restrictions ?	OK. Propose que ce champ soit obligatoire. Déplacement vers description des droits : licence ?	Pas utilisé	Supprimer
Permissions spéciales	OK	Pas utilisé. Déplacement vers description des droits : licence ? ce champ spécifierait les modes d'accès possibles	OK	Supprimer
Restrictions d'usages	OK	Pas utilisé. Déplacement vers description des droits : licence ? ce champ spécifierait les usages possibles	Pas utilisé	Garder pour définir les usages possibles : tous, ou seulement pour recherche, par exemple
Contact	Pas utilisé	Pas utilisé car redondant mais propose que soit précisé si c'est le labo ou le chercheur qui assure ce contact.	OK	Garder, utiliser pour le cas ou les droits dépendent de quelqu'un/une institution qui n'a pas déposer le jeu de données

Modèle de départ	Chercheur 1	Chercheur 2	Chercheur 3	Bilan
Conditions d'accès	Pas utilisé	Pas utilisé. Propose qu'on supprime.	Pas utilisé	Garder pour définir les accès possibles : pour tout public, ou IFSTTAR seulement, par exemple
Déclaration de responsabilité	Pas utilisé	Pas utilisé. Propose qu'on supprime.	Pas utilisé	Supprimer du guide mais garder dans Dataverse avec une phrase type : Le producteur et le diffuseur des données ne sont pas responsables des interprétations ou déductions faites à partir de l'utilisation des données.
Notes sur l'accès aux données	Pas utilisé	Pas utilisé. Propose qu'on supprime.	Pas utilisé	Supprimer
Etudes apparentées	OK	Pas utilisé. Propose qu'on supprime.	Pas utilisé	Fusionner avec « autres références »
Autres références	Pas utilisé	Pas utilisé. Propose qu'on supprime.	OK	Fusionner avec « études apparentées ». Réécrire les consignes en conséquence.
Nom de fichier	OK	OK	OK	Garder
Contenu du fichier	Pas utilisé (mais note que utile si plusieurs fichiers composent le jeu)	OK	OK	Garder
Nombre d'observations du fichier	Pas utilisé	Pas utilisé	OK	Supprimer
Nombre de variables du fichier	Pas utilisé	Pas utilisé	OK. Mais c'est redondant avec ce qui est demandé avec le résumé.	Supprimer
Type de fichier	Redondant avec logiciel et format ?	OK	OK	Garder celui là
Format du fichier	Pas utilisé	Pas utilisé. Propose de regrouper type et format de fichier	OK	Regrouper ?

Modèle de départ	Chercheur 1	Chercheur 2	Chercheur 3	Bilan
Vérifications	Pas utilisé	OK	Pas utilisé	Supprimer
Logiciel de génération du fichier	OK	OK	OK mais trouve ça redondant	Garder celui là
Commentaires généraux du testé	Rajouter champs labo ou département d'appartenance de l'auteur	Autoriser les données méthodo et donc sans publi « officielle ». Rajouter champs labo ou département d'appartenance de l'auteur	Changer l'ordre des champs pour mettre ceux qui concernent tout type de données au début	

Annexe 6 : Guide du déposant

Voici la version finale du guide du déposant de données, élaborées pour les *datasets* de type enquête, à partir de la norme de description DDI et des résultats des tests effectués par trois chercheurs (voir annexe 5). Nous n'y retenons que les questions correspondant à des besoins qui sont apparus durant cette phase de test.

Guide du dépôt de données IFSTTAR dans le Dataverse

Une banque de données vit par la qualité et la quantité des apports faits par les chercheurs. Voici à cet effet un guide du déposant pour faciliter le dépôt de vos données. Ce formulaire permet de préparer la description des données.

Attention : Ce guide est destiné à la description de données d'enquête ou statistiques. Pour tout autre type de données, un autre formulaire de description est disponible. Néanmoins, certaines questions ne sont pertinentes que pour un type très spécifique de données : seuls les champs en rouge sont à renseigner pour tout jeu de données.

Date : .....

1/Description du jeu de données

Titre : Titre complet du jeu de données. Le titre complet peut indiquer la couverture géographique aussi bien que la période temporelle couverte.
.....
.....
.....

Sous titre : Titre secondaire utilisé pour souligner certaines limitations du titre principal. Il peut répéter des informations déjà présentes dans le titre principal.
.....
.....
Types de données : Type de données contenues dans le fichier : données d'enquête, données de recensement, données agrégées, code source de programme, texte lisible en machine, données administratives, données d'expérience, test psychologique, données textuelles, texte codé, document codé, agenda d'activité, données d'observation, données produites par traitement etc.
.....
.....

Publication(s) : Publication(s), rapport ou note, à laquelle le jeu de données a donné lieu , basé sur les données. Cela peut prendre la forme de références bibliographiques. Au minimum, indiquer le titre, l'auteur, la date, le producteur/diffuseur de la (ou des) publication(s).

.....
.....
.....
.....

Auteur : Personne, morale ou physique, responsable du contenu du jeu de données (possibilité de plusieurs auteurs) sous le format NOM DE FAMILLE Prénom, Laboratoire ou département d'appartenance pour les personnes physique, et Nom complet (Acronyme) pour les institutions.

.....
.....
.....

Producteur : Personne ou organisme qui a assuré la responsabilité financière ou administrative pour la réalisation matérielle du jeu de données (Possibilité de mentionner ici les différents types de participation dans le processus de production).

.....
.....

Date de production : Date de production des données- pas la date de distribution ni d'archivage -. Format AAAA-MM-JJ ou AAAA au minimum.

.....

Lieu de production : Adresse de l'organisme qui a produit les données.

.....
.....

Financement : Source(s) de financement pour la production des données ou projet financé dans le cadre duquel les données ont été produites.

.....
.....

Distribution : Organisme(s) désigné(s) par l'auteur ou par le producteur pour effectuer des copies des données.

.....
.....

Contact : Nom et adresse de la personne à contacter comme personne ressource en cas de problème ou de question soulevés par les utilisateurs. En général, c'est le chercheur qui dépose les données. Noter son nom, son mail et son laboratoire.

.....

Dépositaire : Nom de la personne (ou de l'institution) qui a déposé le jeu de données. Ne pas renseigner si c'est la même personne qui est mentionnée en « contact ».

.....

Série ou collection : Nom de la série ou de collection à laquelle le jeu de données appartient. Eventuellement, historique de la série, récapitulatif des informations qui s'appliquent à l'ensemble.

.....
.....
.....

Citation bibliographique : Eléments permettant de constituer une citation bibliographique complète du jeu de données. Ex. : INSEE –Enquête emploi, 1965. Au minimum, Titre – Producteur, année de production.

.....
.....
.....

Classification : Mots décrivant les thèmes généraux couvrant les données

.....
.....
.....

Mots-clefs: Mots ou phrases décrivant le contenu des données

.....
.....
.....

Résumé/sujet : Le résumé décrit le but, la nature et les limites de la collection de données, les caractéristiques de contenu, les principaux sujets couverts, à quelles questions on a tenté d'apporter des réponses. Une liste des principales variables peut être apportée ici.

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

Période d'étude : Période de recueil ou d'observation couverte par les données.

.....
.....

Couverture géographique : Information sur la couverture géographique : totalité des limites géographiques des données ainsi que tous les niveaux géographiques complémentaires pouvant être codés dans les variables. Exemple : la région Rhône-Alpes, le département de la Savoie, l'Europe.

.....
.....
.....

Individus : Unité de base pour les analyses ou les observations décrites : individu, famille/ménage, groupe, vélo, station, institution/organisation, unité administrative, etc.

.....
.....
.....

Univers: Description de la population couverte, des groupes de personnes ou autres éléments qui constituent l'objet de l'étude ou de l'enquête et auxquels les résultats se réfèrent.

.....
.....
.....
.....

Préparation / Nettoyage : Méthode utilisée pour "nettoyer" les données, c'est à dire vérifications de cohérence, vérifications de codes aberrants, etc.

.....
.....
.....
.....

Archive d’origine : Centre d’archivage auprès duquel les données ont été obtenues ; centre d’archivage d’origine.

.....

Licence/copyright : Nom de la licence ouverte régissant les droits de reproduction attachés au jeu de données.

.....
.....

Restrictions d’usage : Définit les usages autorisés/interdits des données. Par exemple, certification privée ou restrictions de diffusion. Les restrictions peuvent être exigées par l'auteur, le producteur ou le diffuseur des données. Si la restriction ne porte que sur une sorte d'utilisateur, préciser le type.

.....
.....
.....

Contact pour les droits : Personne ou organisation (avec l'adresse complète et le n° de téléphone si possible) qui contrôle l'accès aux données si elle est différente de celle du diffuseur des données. Possibilité de mentionner l'adresse mail du contact.

.....

.....

Conditions d’accès : Indiquer toute information complémentaire pouvant aider l'utilisateur à comprendre les conditions d'accès et d'utilisation des données. Par exemple, une diffusion restreinte au personnel de l’IFSTTAR.

.....
.....
.....

Etudes apparentées : Information sur les toute autre étude en lien avec le jeu de données décrit. Mentionner le type de lien.

.....
.....
.....
.....

2/ Description des fichiers

Nombre de fichiers : Nombre total de fichiers associés à la collection.

.....

Nom(s) : Contient un titre court qui sera utilisé pour distinguer chaque fichier des autres fichiers de la même collection. Au format NOM.EXTENSION

.....
.....
.....

Contenu : Résumé ou description du fichier. Description sommaire du but, de la nature et du champ du fichier, des caractéristiques spéciales de son contenu, des principaux thèmes couverts, et à quelles questions les investigateurs ont essayé de répondre en créant le fichier. Par exemple, une liste des principales variables, ou le nombre d'observations.

.....
.....
.....
.....
.....

Type de fichier : Type et format de fichier (ASCII, EBCDIC, etc.) et le logiciel associé comme les 'fichiers de données SAS', les 'fichiers export SPSS', etc.

.....
.....
.....

Logiciel de génération du fichier : Logiciel utilisé pour créer le fichier. L'attribut "version" permet de spécifier le numéro de la version du logiciel

.....

Annexe 7 : Guide du déposant issu du format DataCite

Voici un second guide du dépôt de données, destiné aux *datasets* issus de recherches et d'études qui ne sont pas des enquêtes. Il s'appuie sur le modèle de description et de métadonnées de jeux de données proposé par DataCite, association internationale qui œuvre pour la « citabilité » des données de recherche. Nous l'avons testé auprès d'un chercheur unique, et c'est pourquoi il n'a pas été nécessaire de comparer les résultats dans un tableau. Le document reproduit ici est la version finale du guide du déposant pour ce type de données.

Guide du dépôt de données dans le Dataverse IFSTTAR

Pour les données d'essais ou autre type de données non issues d'enquête

Attention : Ce guide est basé sur le modèle proposé par DataCite. Pour les jeux de données statistiques ou issus d'enquêtes, un autre formulaire de description est disponible, basé sur la norme DDI.

Néanmoins, certaines questions ne sont pertinentes que pour un type très spécifique de données : seuls les champs en rouge sont à renseigner obligatoirement.

Une banque de données vit par la qualité et la quantité des apports faits par les chercheurs. Ce formulaire constitue un guide pour faciliter le dépôt de vos données et leur description. Il se décompose en 2 parties :

1/ Description de l'étude

2/ Description des fichiers

1/Description de l'étude

Auteur(s) : Personne, morale ou physique, responsable du contenu du jeu de données (possibilité de plusieurs auteurs).

.....

Affiliation de l'auteur : son laboratoire ou son département.

.....

Titre : Titre complet du jeu de données. Si possible, le titre complet doit indiquer la couverture géographique aussi bien que la période temporelle couverte.

.....

.....

.....

Sous-titre : Titre en anglais pour les données en français, titre en français pour les jeux en anglais.

.....

.....

Editeur(s) ou producteur(s) : Personne ou organisme qui a assuré la responsabilité financière et/ou administrative pour la réalisation matérielle du jeu de données (Possibilité de mentionner ici les différents types de participation dans le processus de production).

.....

.....

.....

Date d'édition ou de production : Date de production des données- pas la date de distribution ni d'archivage -. Format AAAA/MM/JJ ou AAAA au minimum.
.....

Mots-clefs : Mots-clefs définissant le sujet du jeu de données.
.....
.....
.....

Description : Il peut s'agir d'un résumé, d'une description de la méthode, ou d'une table des matières. La description explique le but, la nature et les limites de la collection de données, les caractéristiques de contenu, les principaux sujets couverts, à quelles questions on a tenté d'apporter des réponses. Une liste des principales variables peut être apportée ici.
.....
.....
.....
.....
.....
.....
.....
.....
.....

Date(s) : Toute date pertinente pour décrire le jeu de données. Spécifier à quoi elle correspond.
.....

Publication liée : Publication(s) à laquelle le jeu de données a donné lieu. Cela peut prendre la forme de références bibliographiques. Au minimum, indiquer le titre, l'auteur, la date, le producteur/diffuseur de la (ou des) publication(s). Si la publication possède un identifiant pérenne (DOI, Handle, URI…), le noter ici.
.....
.....
.....
.....
.....

Types de données : Type de données contenues dans le fichier, par exemple données d'essai, données agrégées, code source de programme, texte lisible en machine, données administratives, données d'expérience, test psychologique, données textuelles, texte codé, document codé, agenda d'activité, données d'observation, données produites par traitement etc.
.....
.....

Version : numéro de version du jeu de données.
.....

Droits : Licence, copyright ou tout autre droit de reproduction attachés au jeu de données.
.....
.....

Source de financement : Source(s) de financement pour la production des données, projet dans le cadre duquel les données ont été recueillies etc.
.....
.....
.....

Contact : Nom et adresse de la personne à contacter comme personne ressource en cas de problème ou de question soulevés par les utilisateurs.
.....

Localisation géographique : nom du lieu où les données ont été recueillies ou que les données concernent.
.....

2/ Description des fichiers

Nom : Contient un titre court qui sera utilisé pour distinguer chaque fichier des autres fichiers de la même collection. Au format NOM.EXTENSION
.....
.....
Format de fichier : format physique du fichier
.....
.....

Logiciel de génération du fichier : Logiciel utilisé pour créer le fichier. L'attribut "version" permet de spécifier le numéro de la version du logiciel
.....

Date de dépôt : Date du dépôt du jeu de données. Format AAAA-MM-JJ
.....

Annexe 8 : Synthèse sur les licences et les waivers pour les données de la recherche.

Le document suivant est la synthèse que nous avons élaborées à destination des chercheurs de l'IFSTTAR qui portent le projet Belgrand-GEBD, afin qu'ils puissent choisir en connaissance de cause une licence appropriée à leurs besoins. Elle ne prend en compte que les dispositifs juridiques les intéressant directement, et exclue donc les licences ne permettant pas de réutilisation des données, ou les licences ne concernant que les logiciels.

Licences et *waivers*.

Quelle réutilisation autoriser pour des données de recherche ?

Des licences ouvertes, pourquoi faire ?

Dans le droit français, les données « brutes » ne sont pas protégées par le droit d'auteur, mais la constitution d'une collection « originale » peut relever du droit *sui generis* des BDD.

Mais l'enregistrement d'une donnée de recherche relève la plupart du temps d'une opération scientifique : comment, dès lors, identifier la donnée libre de droit, et la distinguer de la donnée « créée » par un auteur-chercheur ?

En l'état actuel, il n'existe pas en France de réponse car il n'y a pas de définition claire de la donnée brute libre de droit. Cette indétermination juridique nationale se double de la multiplicité des points de vu européens et internationaux.

Dans l'attente d'une clarification européenne du cadre légal, l'utilisation de licence demeure le mécanisme le plus adapté à un projet d'ouverture, même partielle, de données de la recherche, et le meilleur moyen de définir les obligations juridiques attachées à une donnée.

Comment choisir une licence ouverte ?

Dans le cadre d'un projet d'ouverture, même contrôlée, qui doit permettre la réutilisation des données, il est inutile d'englober dans la comparaison les licences qui empêchent la modification des données (clause de type ND) et/ou leurs usages à des fins commerciales (clause NC). Ces deux clauses sont globalement incompatibles avec tout objectif de réutilisation de données.

Les licences qui traitent spécifiquement des logiciels ou centrées sur le droit d'auteur ont été écartées.

Deux types de clauses nous intéressent tout particulièrement :

- La mention de paternité (ou attribution ou BY) : c'est la mention de la source. La clause BY oblige à toujours mentionner le créateur du jeu de données lorsce qu'on utilise ce dernier.

Cette mention –ou son absence- ne concerne pas la citation du jeu de données dans une publication, comme une simple référence bibliographique, mais bien le crédit lors d'une réutilisation.

Avantage : cette clause assure au créateur d'un jeu de données d'être crédité lors de sa réutilisation.

Inconvénient : le souci du *cascading attribution* ou *attribution stacking*, c'est-à-dire l'empilage des attributions. L'agrégation d'un nombre important de *datasets* peut se révéler compliqué s'il faut créditer individuellement chaque contributeur.

- Le partage à l'identique (ou Share-Alike ou SA ou Copyleft) : C'est une clause qui oblige à redistribuer les modifications sous les mêmes conditions que celles apposées sur le jeu « initial ». Cela force une réutilisation « open ». Une licence qui pose ce principe est appelée contributive (à l'inverse, lorsqu'une licence n'impose pas de partage à l'identique, elle est dite permissive).

  Avantage : cela impose le fait d'ouvrir à nouveau, et à tous, ce qui a été produit, en reversant cette valeur ajoutée dans une sorte de « pot commun ». Les contributions réalisées sous cette clause participent à un enrichissement collectif.

  Inconvénient : il s'agit d'une forte limitation aux possibilités de réutilisations. Un *dataset* ainsi protégé ne peut être réutilisé que pour une production elle-même partageable sous les mêmes conditions.

Licence ou waiver ?

Certaines item présentés dans le tableau ci-dessous ne font mention ni de la clause BY, ni de la clause SA (ou de leurs équivalents). Ce ne sont pas *stricto sensu* des licences, mais des *waivers*, ou renonciations, ou licences du domaine public. Il s'agit d'outils juridiques permettant au détenteur des droits de propriété intellectuelle sur une donnée de les abandonner, faisant basculer celle-ci dans le domaine public.

Les partisans de l'ouverture des données privilégient ce type d'outil. Formellement, les licences du domaine public n'imposent pas au réutilisateur de mentionner l'origine et le créateur des données. Mais les normes internationales de la recherche et l'évaluation par les pairs sont généralement suffisantes pour concilier cet abandon des droits sur les données et une mention légitime de leur producteur.

Les licences permissives					
Licences	**Editeurs**	**Contenus**	**Exemples d'utilisations**	**BY ou paternité**	**SA ou copyleft**
LO	Etalab	La licence ouverte (LO) d'Etalab est dite permissive : la seule contrainte est de mentionner la paternité et la date de la dernière mise à jour. Pour le reste, il est possible de reproduire, redistribuer, modifier, exploiter à titre commercial. Elle assure ainsi un maximum de compatibilité avec les licences libres existantes.	Portail data.gouv.fr Nombreuses collectivités : Bordeaux, Montpellier, l'Auvergne, Les Hauts de Seine	Oui	Non
ODC-By	Open Knowledge Foundation	*L'open data common by* (ODC by) est permissive : elle autorise toutes les utilisations à condition que la paternité (*by*) soit indiquée. Elle est donc proche de la licence ouverte d'Etalab.		Oui	Non
CC-BY	Creative Commons	Depuis fin 2013, CC a publié une version de ses licences compatible avec le droit des BDD.	Les licences Creative Commons v4.0 sont privilégiées par OpenAIRE	Oui	Non
Les licences collaboratives					
ODbL	Open Knowledge Foundation	Licence de style *copyleft* qui permet de copier, modifier, de faire un usage commercial, sous trois conditions : citer la source ; redistribuer sous des conditions de partage identiques les modifications ; maintenir ouverte techniquement la base de données que vous redistribuez, qu'elle soit modifiée ou non. Toutefois, il est possible de déroger au *share alike* moyennant contre-partie.	Portail Paris data, Toulouse Métropole, de nombreux offices du tourisme, Communauté urbaine de Bordeaux	Oui	Oui
CC-BY-SA	Creative Commons	Depuis fin 2013, CC a publié une version de ses licences compatible avec le droit des BDD.	Les licences Creative Commons v4.0 sont privilégiées par	Oui	Oui

			OpenAIRE		
Domaine public					
PDDL	Open Knowledge Foundation	La licence *Public Domain Dedication and License* (PDDL) revient à renoncer à tout droit puisque la base de données est placée dans le domaine public. Elle se rapproche de la licence CC0.		Non	Non
CC0	Creative Commons	La CC0 est une renonciation : placer ses données sous CC0 revient à les mettre dans le domaine public. Avant la version 4.0, c'était la seule licence CC compatible avec le droit *sui generis* des bases de données.	Archive ouverte Dryad Editeurs Pensoft et BioMed Central	Non	Non

Sources :

Ball, A. (2012). « How to License Research Data'. DCC How-to Guides. Edinburgh: Digital Curation Centre. » Disponible en ligne : http://www.dcc.ac.uk/resources/how-guides

Blanc, S. (2013). « Le fouillis des licences open data s'éclaircit » [Fiche pratique]. *La Gazette des Communes*, 25/11/2013. Disponible en ligne :

http://www.lagazettedescommunes.com/208893/le-fouilli-des-licences-open-data-seclaircit-fiche-pratique/

Gaillard, R. (2013). « De l'Open data à l'Open research data : quelle(s) politique(s) pour les données de recherche ? » Mémoire ENSSIB.

Disponible en ligne : http://www.enssib.fr/bibliotheque-numerique/documents/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche.pdf

Annexe 9 : Fiche de poste du CDSP pour le recrutement d'un ingénieur d'études

Ce document correspond à un poste mis au concours ITRF 2014.

Description du poste

BAP : D
Corps : Ingénieur d'études
Emploi-type : Traitement et analyse de bases de données

Mission :

- L'ingénieur d'études sera chargé de l'enrichissement du catalogue de données quantitatives diffusées par le centre (données d'enquêtes d'opinion, données électorales, enquêtes ELIPSS).
- Il sera en contact avec les producteurs de données pour leur documentation et les utilisateurs (chercheurs et étudiants) pour leur diffusion.
- Il contribuera à la production de données (Enquête sociale européenne, enquêtes et bases de données produites par des équipes de recherche).
- Il participera à des réseaux scientifiques nationaux et internationaux en matière d'archivage, de documentation et de diffusion de données (Réseau Quetelet, CESSDA et IASSIST notamment).

Activités :

- Documenter les données en vue de leur diffusion :
* Décrire les objectifs et le protocole de l'enquête
* Vérifier la cohérence des données
* Enrichir les fichiers de données (description des variables, questions et modalités de réponse…)
* Publier les données et leur documentation sur le serveur Nesstar du CDSP et sur le portail du Réseau Quetelet
- Analyser les données et évaluer leur qualité
- Aider les producteurs de données pour la documentation de leurs enquêtes
- Identifier les statuts de diffusion des données et mettre en oeuvre les protocoles de diffusion correspondants
- Informer la communauté scientifique de la nature et de l'intérêt des données du catalogue du CDSP
- Assister les chercheurs pour produire des données
- Participer aux différentes phases de la partie française de l'Enquête sociale européenne (préparation du questionnaire, préparation et suivi de la collecte et préparation des données)
- Assurer une aide technologique aux utilisateurs de données
- Participer aux activités du réseau Quetelet

Compétences :

- Maîtrise des outils statistiques et informatiques (SPSS, SAS, STATA ou R)
- Bonnes connaissances en sciences sociales
- Expérience dans le traitement statistique des enquêtes par questionnaire
- Grande rigueur dans la documentation et dans l'analyse de la qualité des données
- Capacités rédactionnelles
- Bonne compréhension écrite et orale de l'anglais

DRH-Pôle académique
27 rue Saint-Guillaume
75007 Paris
Isabelle DELACROIX Catherine TANAKA
isabelle.delacroix@sciencespo.fr catherine.tanaka@sciencespo.fr
Tél. : 01 45 49 52 79 Tél : 01 45 49 50 15. Site internet du Ministère pour le suivi des candidatures :
http://www.itrf.education.gouv.fr