Comment passer de l'entrepôt de données aux data papers ? Retour sur l'expérience de data.sciencespo : difficultés rencontrées et pistes de solutions

Alina Danciu, Anna Egea, Guillaume Garcia, Cyril Heude

▶ **To cite this version:**

Alina Danciu, Anna Egea, Guillaume Garcia, Cyril Heude. Comment passer de l'entrepôt de données aux data papers ? Retour sur l'expérience de data.sciencespo : difficultés rencontrées et pistes de solutions. #dhnord2021 - Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux, Maison Européenne des Sciences de l'Homme et de la Société, Nov 2021, Lille, France. hal-03445731

HAL Id: hal-03445731

https://sciencespo.hal.science/hal-03445731

Submitted on 30 Nov 2021

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

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

Passer de l'entrepôt de données aux data papers ? Retour sur l'expérience de Data Sciences Po

Alina Danciu, Sciences Po, Centre de données sociopolitiques (CDSP), CNRS, Paris, France
Anna Egea, Sciences Po, Centre de sociologie des organisations (CSO), CNRS, Paris, France
Guillaume Garcia, Sciences Po, Centre de données sociopolitiques (CDSP), CNRS, Paris, France
Cyril Heude, Sciences Po, Direction des Ressources et de l'Information Scientifique (DRIS), Paris, France

Cette contribution restitue la manière dont la mise en œuvre d'un entrepôt Dataverse mutualisé au sein des laboratoires de Sciences Po - "data.sciencespo.fr" - nous a poussés à investir la question des data papers. Nous détaillerons le cheminement qui nous a amenés à nous saisir de ce format pour encourager les dépôts de jeux de données par les chercheurs et chercheuses, et faciliter leur travail pour documenter les données. Nous montrerons comment les obstacles auxquels nous faisons face - la faible stabilisation de ce nouveau genre rédactionnel et sa faible visibilité dans les communautés de recherche - nous ont poussés à explorer certaines pistes pour tenter de dépasser ces écueils.

Ce compte rendu sera polyphonique, puisqu'il croise les expériences des différents acteurs et métiers concernés par cette démarche : un laboratoire de recherche (le Centre de Sociologie des Organisations - CSO), un service transverse (la Direction des Ressources et de l'Information Scientifique - DRIS) et un centre de données (le Centre de Données Socio-Politiques - CDSP) ; une documentaliste en soutien à la recherche, des data librarians, des data managers.

Nous remettrons d'abord en contexte les dispositifs qui ont précédé la mise en place de Data Sciences Po. Nous concentrerons toutefois notre propos sur cet entrepôt, qui abrite deux collections différentes, l'une destinée aux chercheurs et chercheuses de Sciences Po, et l'autre destinée à la communauté nationale et internationale de recherche en sciences sociales.

Nous reviendrons ensuite sur les problèmes que nous rencontrons pour favoriser la documentation des données, quelles que soient leurs modalités de dépôt. Cet entrepôt propose en effet deux formes de service : un service d'auto-dépôt accompagné ; un service de curation des données qui est assuré pour le compte des déposants. Inciter les chercheurs et chercheuses à auto-documenter leurs jeux de données, tout comme prendre en charge cette activité pour eux, présentent des limites, et c'est pourquoi nous nous dirigeons vers la solution des data papers.

Nous discuterons enfin les solutions que nous commençons à mettre en place pour tenter de développer la pratique de rédaction de data papers, en faisant de ce genre rédactionnel nouveau une véritable opportunité pour les communautés de recherche.

En guise de conclusion, nous dirons quelques mots de la nécessité qu'il y a, selon nous, de repenser la question du positionnement des data papers via celle de leur lectorat.

Retour sur notre expérience en matière de documentation des données

Sciences Po, précurseur en matière de valorisation des données de la recherche

Pendant longtemps, à Sciences Po, le dépôt et la diffusion des données en sciences sociales était assurés par le CDSP - depuis 2006 avec la création de cette unité mixte de service du CNRS et de Sciences Po, qui a la particularité d'accueillir essentiellement des ingénieurs issus de différents métiers. La démarche a débuté avec des données quantitatives - i.e. le plus souvent obtenues par la passation de questionnaires sur des échantillons représentatifs de la population générale ou de certains groupes sociaux spécifiques. L'outil qui permettait d'explorer en ligne ces données, NESSTAR, a été utilisé jusqu'en 2020 (il est désormais obsolète). Un peu plus tard a été conçue la base de question Quetelet. Cet outil a été développé par le CDSP et mis à disposition de l'ancien réseau Quetelet, qui est aujourd'hui Quetelet PROGEDO-Diffusion. L'outil permet de faire des recherches dans le texte des questions, les codes et étiquettes des modalités de réponse, les noms et les étiquettes de variable. A partir de 2013, le CDSP a commencé à diffuser des données d'enquêtes qualitatives - i.e. obtenues par entretiens approfondis ou par observations[1]. L'outil utilisé pour explorer les données en ligne s'appelle beQuali. Ce service de dépôt et de diffusion[2] a comme particularité de s'adresser aux chercheurs et équipes de recherche situés aussi bien à Sciences Po qu'à l'extérieur de Sciences Po. Son autre particularité est que le CDSP prend en charge en interne la curation des données[3] pour le compte des déposants.

La première étape de l'évolution vers Data Sciences Po s'est opérée via les réalisations faites pour les besoins d'un projet aujourd'hui clos, Archipolis. Archipolis était un réseau de laboratoires de Sciences Po et d'autres universités en France - Lille, Grenoble, Lyon, Bordeaux - organisé comme un consortium d'Huma-Num entre 2012 et 2016. Le projet visait à créer une dynamique de préservation et de valorisation d'archives d'enquêtes de terrain. Seules étaient concernées les archives d'enquêtes qualitatives - principalement par entretiens - situées dans un domaine disciplinaire hybride, les sciences sociales du politique - essentiellement en science politique et en sociologie (Duchesne et al., 2014).
Archipolis a notamment permis de réaliser des inventaires d'enquêtes dans les différents laboratoires membres[4] et de développer, en 2015, une collection dédiée sous le logiciel libre Dataverse. L'objectif était de diffuser des notices d'enquêtes, comme une première étape destinée à faire évoluer les mentalités des ingénieurs comme des chercheurs des laboratoires, en prenant l'habitude d'inventorier et de décrire les jeux de données qui y étaient produits. La diffusion des données elles-mêmes était prévue dans un second temps de développement du projet, mais l'arrêt du réseau en 2016 n'a pas permis de franchir cette étape.

[1] Une vingtaine d'enquêtes qualitatives ont été collectées, qui sont déjà diffusées qui vont l'être prochainement.
[2] Le téléchargement des données s'effectue depuis de nombreuses années depuis le portail Quetelet Progedo Diffusion
[3] Par curation, on entend des opérations comme l'anonymisation des données, l'ajout de métadonnées, la migration des fichiers des données en formats libres.
[4] 238 notices ont ainsi été élaborées, sur les 8 laboratoires membres du réseau en 2016.

Figure 1 : Catalogue Archipolis

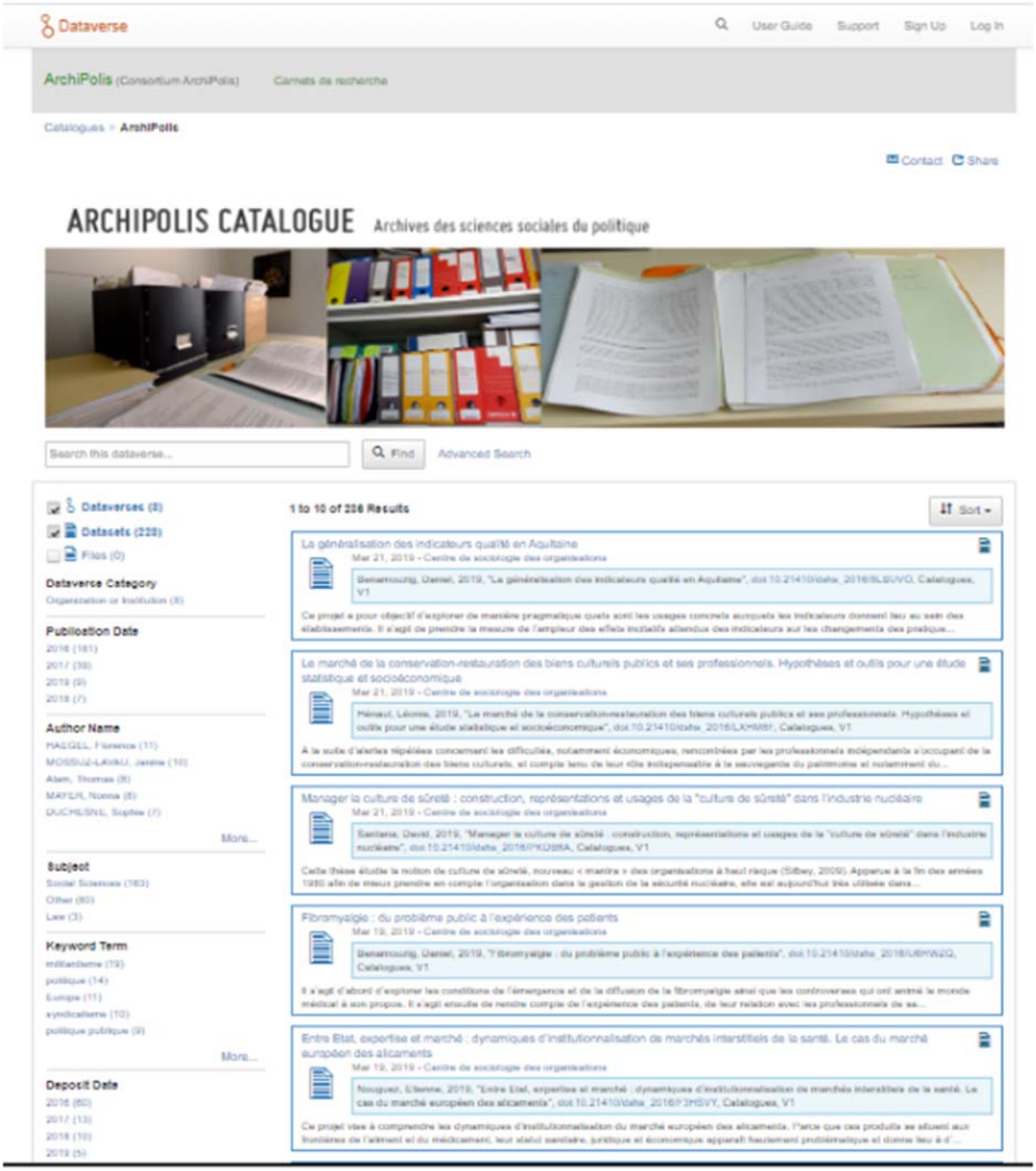


Image produite par l'auteur

Ces notices étaient organisées comme un ensemble de métadonnées relevant du standard international Data Documentation Initiative (DDI)[5].

[5] Le CDSP est fortement impliqué dans la communauté en charge du maintien et de la formation à ce standard de métadonnées en SHS. Il est par ailleurs membre de la DDI Alliance, organisme qui fait évoluer ce standard.

Figure 2 : Exemple d'une notice d'enquête Archipolis

Image produite par l'auteur

Les notices étaient renseignées par les ingénieurs des laboratoires du réseau ou des contractuels recrutés sur le projet, en lien avec les producteurs des enquêtes. Avec Archipolis nous installions ainsi l'idée d'un circuit reliant projets de recherche, publications et données, chercheurs et ingénieurs, et commencions à questionner la relation strictement personnelle entre les données et leurs producteurs qui prévalait jusqu'alors.

Le passage à l'entrepôt de données de la recherche Data Sciences Po

Vers 2016, l'idée a émergé de doter Sciences Po d'un entrepôt institutionnel de données de recherche, Data Sciences Po. À Sciences Po, la politique de la science ouverte est portée par la Direction des ressources et de l'information scientifique (DRIS) conjointement avec la Direction Scientifique (DS). Cette politique a abouti, entre autres, à la mise en place d'une archive ouverte institutionnelle, à la rédaction d'un texte-cadre et à une expertise développée par les bibliothécaires[6] (accompagnement à la rédaction de PGD, mise en place d'un groupe inter-labos d'entraide, résolution collective de cas pratiques et de dialogue entre laboratoires et services transverses, rédaction d'un guide en ligne bilingue avec système d'options selon le niveau d'information et l'appétence, formations doctorales et des personnels des laboratoires, conseils aux chercheurs sur l'ensemble du cycle de vie des données, guichet unique d'information, archivage numérique pérenne, engagement dans des réseaux nationaux et européens de professionnels des données). De nombreux acteurs ont été mobilisés pour développer Data Sciences Po et plusieurs mondes professionnels ont été articulés pour faire vivre le projet : des data managers, représenté par le CDSP, des data librarians, représenté par la Direction des ressources et de l'Information Scientifique de Sciences Po, la DSI, la DPO de

[6] Pour en savoir plus, voir une vidéo de présentation : https://drive.google.com/file/d/1ZyTWJx1HRTQjtUfq3RZMObfUj4bABz1b

Sciences Po, la Direction des affaires juridiques et des marchés (DAJAM), ainsi que les laboratoires de recherche, qu'il s'agisse des chercheurs ou des personnels IST. C'est le CDSP qui a coordonné sa mise en place, en se basant sur l'expérience accumulée l'entrepôt Dataverse construit dans le cadre du projet Archipolis.

Les déposants ont à disposition deux collections.

Figure 3 : Page d'accueil de Data Sciences Po

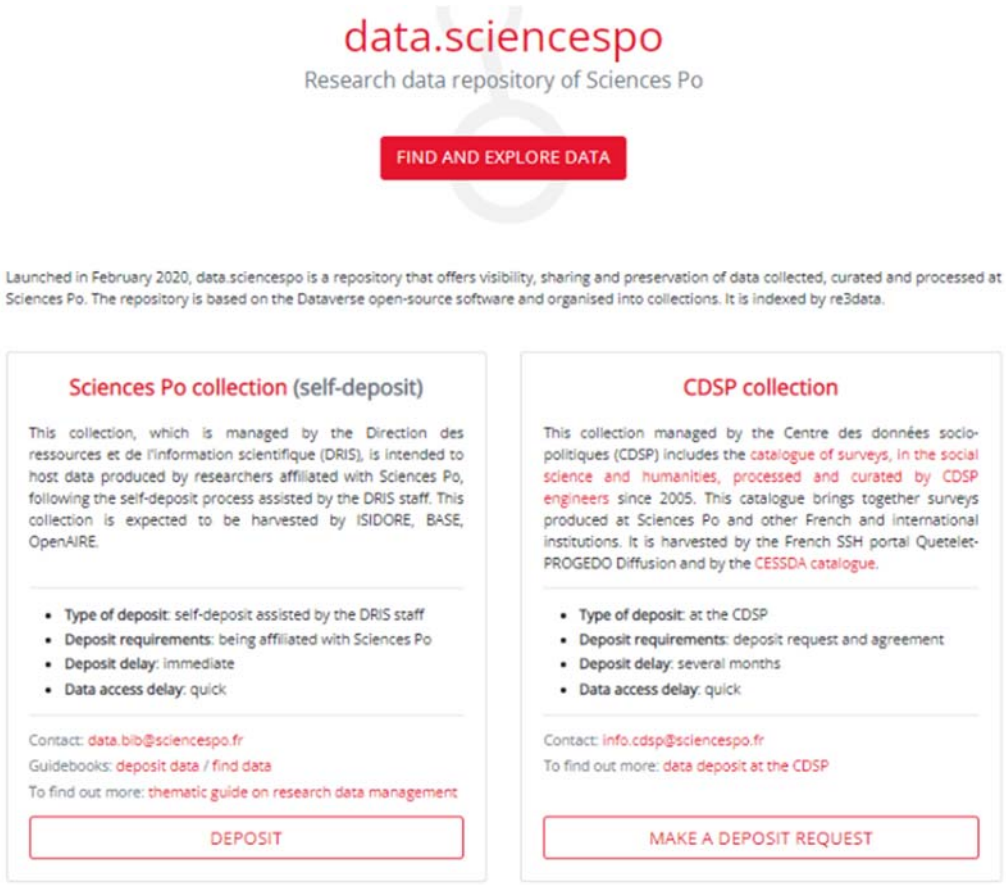


Image produite par l'auteur

La première accueille les données produites par les chercheurs et chercheuses affiliés à Sciences Po (données d'enquête, bases de données, etc.), sur le principe d'auto-dépôt accompagné par la DRIS, en collaboration avec les personnels des laboratoires.

Figure 4 : Catalogue de la collection “au-dépôt” de Data Sciences Po

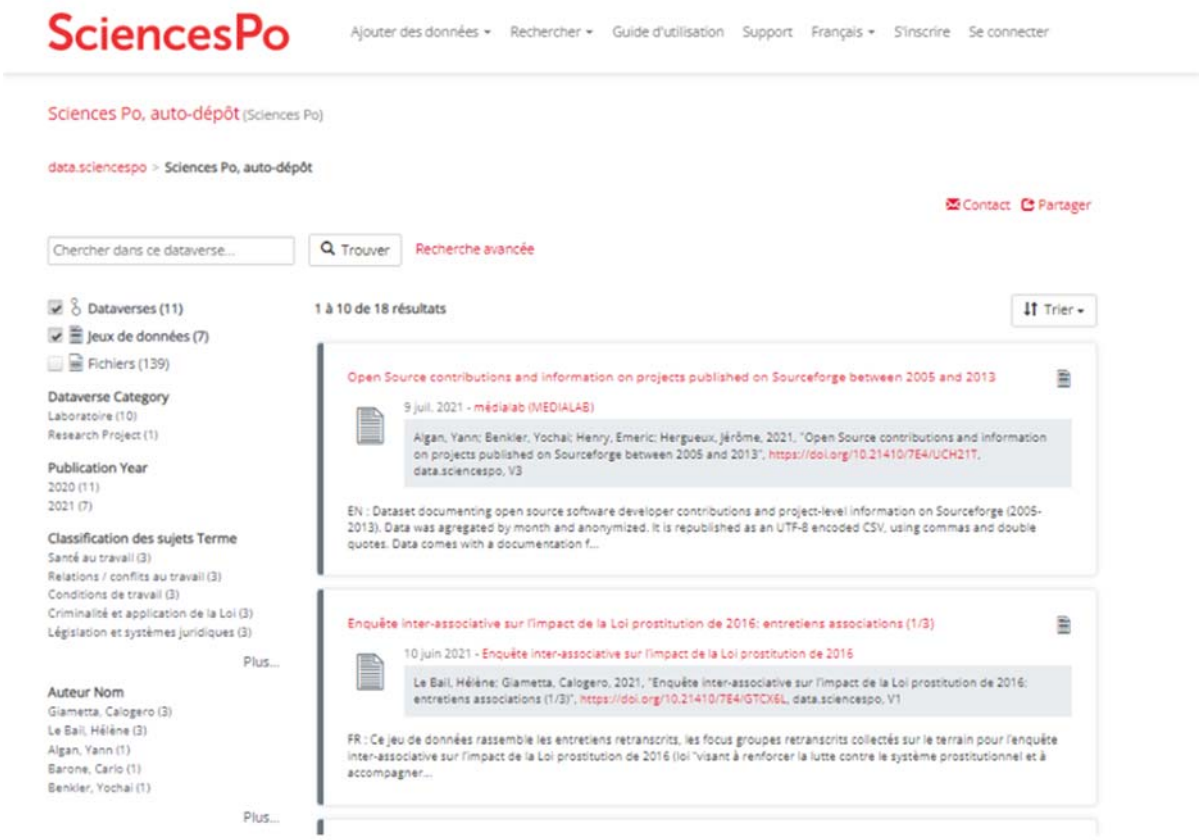


Image produite par l’auteur

La deuxième, administrée par le CDSP, comprend le catalogue d’enquêtes en sciences humaines et sociales traitées et contextualisées par les ingénieurs du CDSP depuis 2006[7]. Ce catalogue rassemble des enquêtes produites à Sciences Po ou dans d’autres institutions françaises et internationales.

[7] Catalogue composé de 350 enquêtes et bases de données.

Figure 5 : Catalogue de la collection “banque de données du CDSP” de Data Sciences Po

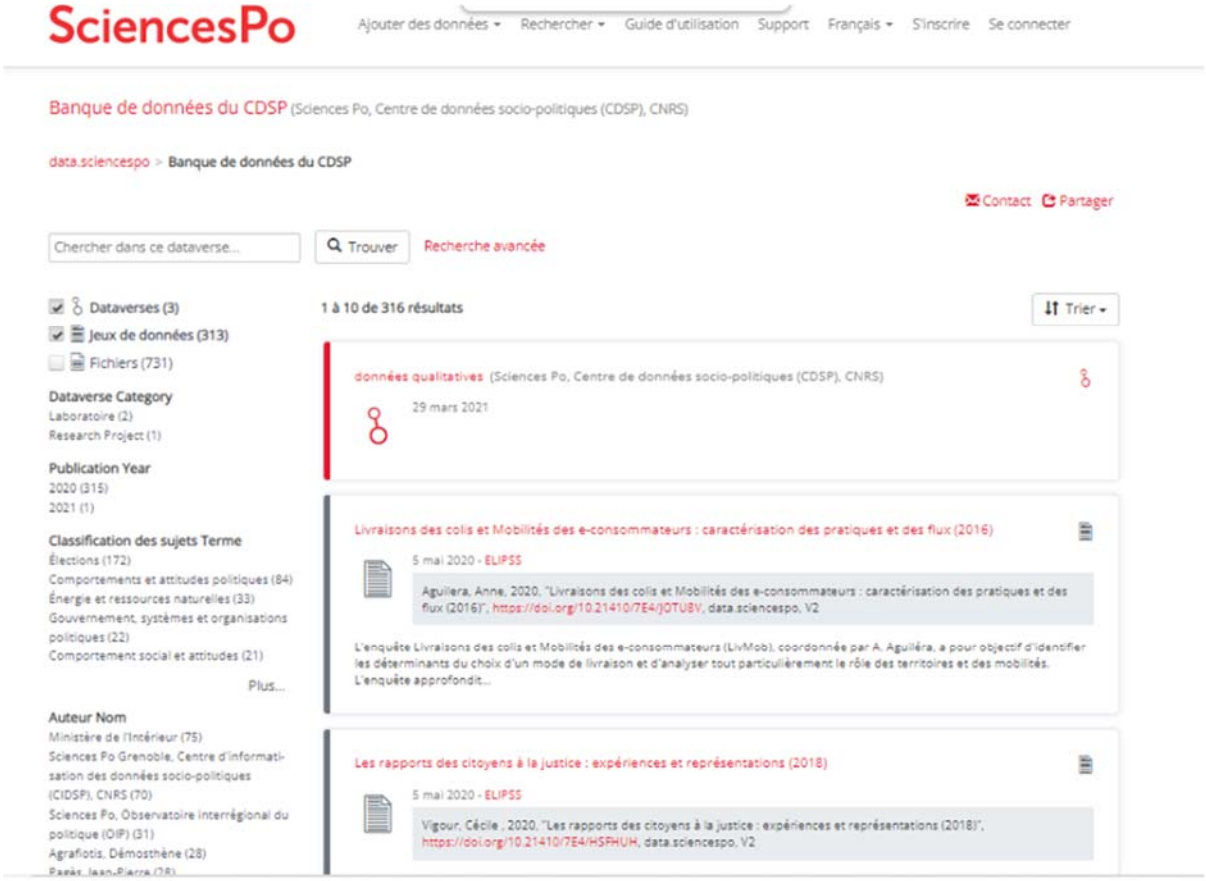


Image produite par l’auteur

Dans les deux cas, les données sont contextualisées en respectant les normes internationales en vigueur, notamment le standard DDI et les principes FAIR. Un DOI est assigné à chaque jeu de données, favorisant ainsi la citabilité et le crédit académique.
Les avantages de cet entrepôt sont multiples. Outre le respect des principes FAIR et les aspects déjà mentionnés, il est moissonnable par d’autres plates-formes et référencé dans les moteurs de recherche, ce qui permet d’accroître significativement la visibilité des données pour les réutilisateurs[8]. Par ailleurs, l’entrepôt répond aux besoins de la communauté des producteurs et utilisateurs de données de recherche, plus spécifiquement au regard des exigences en matière de plans de gestion de données (projets ANR, H2020...). L’entrepôt accepte tous types de données et de formats. Outre l’intégration de la plupart des fonctionnalités génériques de Dataverse destiné à valoriser la visibilité, la citation et la réutilisation des jeux de données des collections par projet/équipe de recherche sont possibles, tout comme les différents types d’accès (ouvert, sur demande, restreint). Il a été développé grâce au même logiciel que le futur entrepôt national de la recherche, ce qui facilitera à terme les liens avec ce dernier.
L’outil technique comme les dispositifs de sensibilisation pour (auto)déposer les données existent, tout comme les ressources et l’expertise pour les documenter ; néanmoins la documentation reste une activité complexe et coûteuse.

[8] Data Sciences Po est référencé par exemple par le registre des données de la recherche re3Data.

Documenter des jeux de données, une activité complexe et coûteuse

Pour clarifier notre propos, nous distinguerons la prise en charge de la curation des données pour le compte des déposants et l'auto-dépôt accompagné.

Gérer la curation pour les déposants : une documentation enrichie des jeux de données

S'agissant des jeux de données dont la curation est d'emblée prise en charge par le CDSP[9], la situation est relativement confortable pour les producteurs déposants, même si ces derniers sont associés, à des degrés variables selon les cas de figure, à tout ou partie des composantes de cette curation. L'équipe dont c'est le cœur de métier est dotée de moyens spécialisés, tant en termes de nombre de personnes que d'expertises.
Avant toute chose, les ingénieurs du CDSP s'assurent que la confidentialité des "répondants" ou des "enquêtés" est respectée. Ensuite, les données sont documentées de manière fine à l'aide du standard de métadonnées DDI. Le CDSP utilise ici le modèle de métadonnées du CESSDA (réseau des centre de données européens en données SHS), grâce auquel on peut renseigner des informations au niveau de l'enquête elle-même (et donc accéder à un certain niveau de connaissance du protocole de recherche) mais aussi documenter plus finalement les données, en particulier les données "quantitatives". Ces dernières sont documentées à l'aide de logiciels tels que Nesstar, Colectica ou R, et leur description détaillée permet à la fois la recherche dans les questions et variables et la ré-exploitation des données. Par ailleurs, le CDSP utilise des vocabulaires contrôlés comme ELSST et ceux de la DDI Alliance, les seconds étant traduits et maintenus en français par le laboratoire. Les données "qualitatives" sont, elles, documentées avec des outils *in-house*.

[9] Les données mises à disposition par le CDSP servent régulièrement dans des projets d'analyse secondaire qui donnent lieu à des publications ou qui sont utilisés dans des cours d'enseignement des méthodes quantitatives ou qualitatives. Un exemple récent d'ouvrage est Controlling the Electoral Marketplace. How Established Parties Ward Off Competition (doi.10.1007/978-3-319-58202-3), écrit par Joost van Spanje, en utilisant des données à thématique électorale diffusées par le CDSP.

Figure 6 : Exemple d'une notice d'enquête de Data Sciences Po

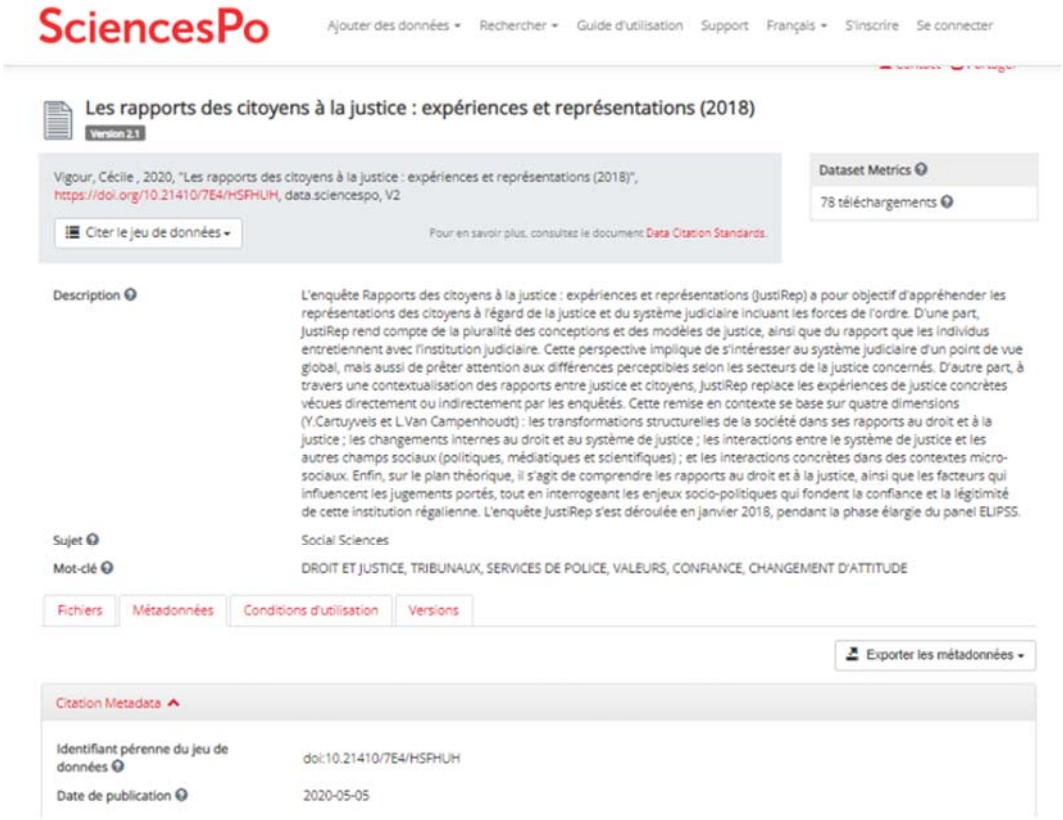


Image produite par l'auteur

Figure 7 : Exemple d'une documentation très fine au niveau des variables quantitatives (source : base de questions Quetelet[10])

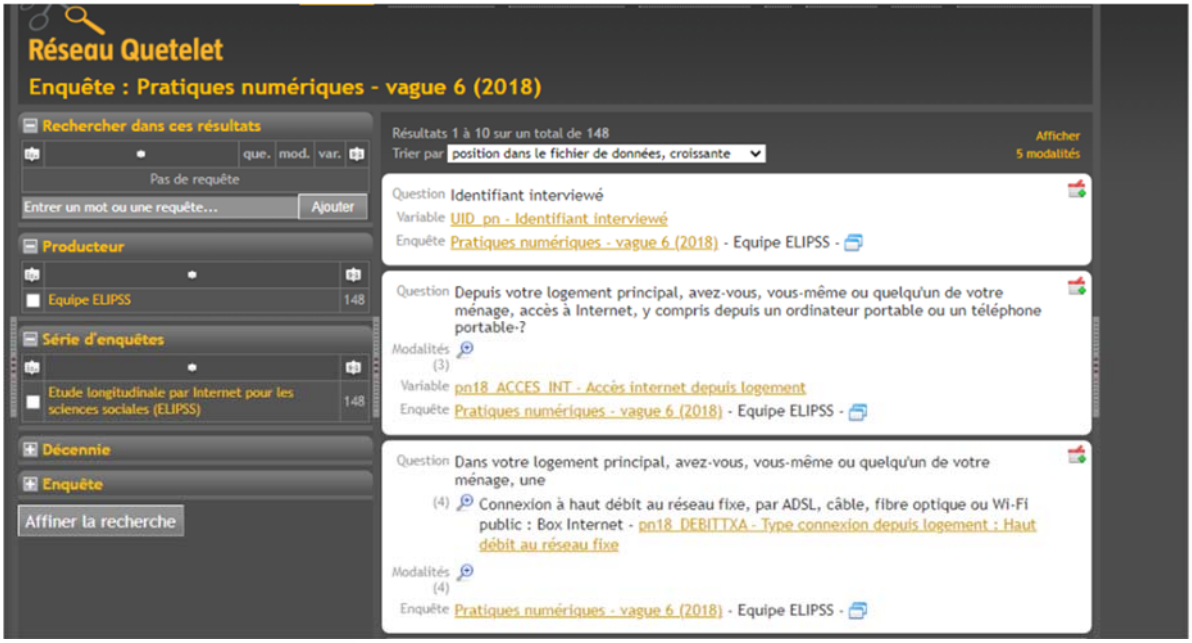
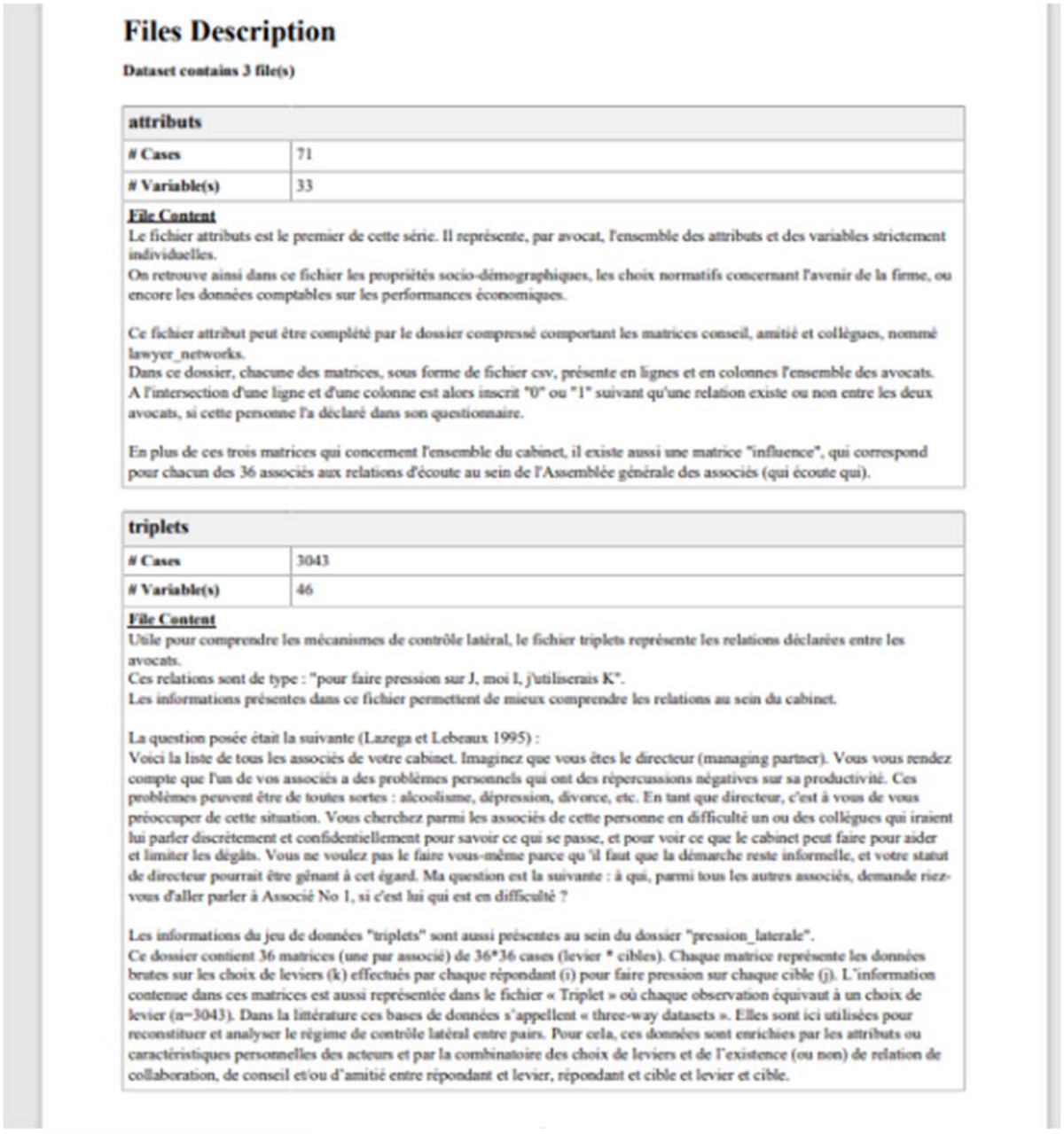


Image produite par l'auteur

[10] http://bdq.quetelet.progedo.fr/fr/Questions_et_variables_d_une_enquete/FR/1998/.

A côté de cette première forme de documentation, les ingénieurs du CDSP procèdent à la rédaction ou la relecture de rapports permettant d'approfondir la compréhension des protocoles d'enquête. S'agissant des bases de données quantitatives, le but est de donner des précisions ou des "explications de texte" concernant le contexte de l'enquête (description de l'enquête, mode de collecte, méthode d'échantillonnage, pondérations, etc.) ainsi que le dictionnaire des variables (texte de la question, univers, instructions enquêteurs, etc), etc.

Figure 8 : Exemple de documentation des variables de l'enquête "Le phénomène collégial"[11]



Files Description

Dataset contains 3 file(s)

attributs	
# Cases	71
# Variable(s)	33

File Content

Le fichier attributs est le premier de cette série. Il représente, par avocat, l'ensemble des attributs et des variables strictement individuelles.
On retrouve ainsi dans ce fichier les propriétés socio-démographiques, les choix normatifs concernant l'avenir de la firme, ou encore les données comptables sur les performances économiques.

Ce fichier attribut peut être complété par le dossier compressé comportant les matrices conseil, amitié et collègues, nommé lawyer_networks.
Dans ce dossier, chacune des matrices, sous forme de fichier csv, présente en lignes et en colonnes l'ensemble des avocats. A l'intersection d'une ligne et d'une colonne est alors inscrit "0" ou "1" suivant qu'une relation existe ou non entre les deux avocats, si cette personne l'a déclaré dans son questionnaire.

En plus de ces trois matrices qui concernent l'ensemble du cabinet, il existe aussi une matrice "influence", qui correspond pour chacun des 36 associés aux relations d'écoute au sein de l'Assemblée générale des associés (qui écoute qui).

triplets	
# Cases	3043
# Variable(s)	46

File Content

Utile pour comprendre les mécanismes de contrôle latéral, le fichier triplets représente les relations déclarées entre les avocats.
Ces relations sont de type : "pour faire pression sur J, moi I, j'utiliserais K".
Les informations présentes dans ce fichier permettent de mieux comprendre les relations au sein du cabinet.

La question posée était la suivante (Lazega et Lebeaux 1995) :
Voici la liste de tous les associés de votre cabinet. Imaginez que vous êtes le directeur (managing partner). Vous vous rendez compte que l'un de vos associés a des problèmes personnels qui ont des répercussions négatives sur sa productivité. Ces problèmes peuvent être de toutes sortes : alcoolisme, dépression, divorce, etc. En tant que directeur, c'est à vous de vous préoccuper de cette situation. Vous cherchez parmi les associés de cette personne en difficulté un ou des collègues qui iraient lui parler discrètement et confidentiellement pour savoir ce qui se passe, et pour voir ce que le cabinet peut faire pour aider et limiter les dégâts. Vous ne voulez pas le faire vous-même parce qu'il faut que la démarche reste informelle, et votre statut de directeur pourrait être gênant à cet égard. Ma question est la suivante : à qui, parmi tous les autres associés, demande riez-vous d'aller parler à Associé No 1, si c'est lui qui est en difficulté ?

Les informations du jeu de données "triplets" sont aussi présentes au sein du dossier "pression_laterale".
Ce dossier contient 36 matrices (une par associé) de 36*36 cases (levier * cibles). Chaque matrice représente les données brutes sur les choix de leviers (k) effectués par chaque répondant (i) pour faire pression sur chaque cible (j). L'information contenue dans ces matrices est aussi représentée dans le fichier « Triplet » où chaque observation équivaut à un choix de levier (n=3043). Dans la littérature ces bases de données s'appellent « three-way datasets ». Elles sont ici utilisées pour reconstituer et analyser le régime de contrôle latéral entre pairs. Pour cela, ces données sont enrichies par les attributs ou caractéristiques personnelles des acteurs et par la combinatoire des choix de leviers et de l'existence (ou non) de relation de collaboration, de conseil et/ou d'amitié entre répondant et levier, répondant et cible et levier et cible.

Image produite par l'auteur

[11] https://cdsp.sciences-po.fr/fr/ressources-en-ligne/ressource/fr.cdsp.ddi.phenomenecollegial1991/.

S'agissant des enquêtes qualitatives, il s'agit de prendre en charge la rédaction d'une "enquête sur l'enquête", en collaboration avec les chercheurs déposants, qui consiste principalement en un rapport de plusieurs dizaines de pages recontextualisant les dimensions du processus de recherche qui restent obscures ou peu éclairées par les archives elles-mêmes ou par les publications tirées de l'enquête. Le document retrace la genèse de l'enquête, ses ancrages théoriques, la réalisation du terrain, ou encore l'analyse des données.

Figure 9 : Sommaire d'un rapport portant sur l'enquête "Choisir son école"[12]

Enquête sur l'enquête « Choisir son école », beQuali, 2016.

Sommaire

Image produite par l'auteur

L'ensemble de ces activités de documentation sont destinées à favoriser le potentiel de réutilisation des jeux de données, mais restent lourdes à gérer. Même si cela dépend des enquêtes et du degré de contribution des déposants, une telle charge est difficilement soutenable sur un nombre important de jeux de données et à long terme, compte tenu de l'incertitude qui pèse sur nos ressources humaines - comme cela est plus généralement le cas actuellement dans nos univers professionnels. De plus, les rapports que nous produisons, notamment l'enquête sur l'enquête, s'ils sont régulièrement publiés sur HAL, restent de la

[12] https://bequali.fr/fr/les-enquetes/lenquete-sur-lenquete/cdsp_bequali_s1/.

“littérature grise”, et sont donc peu valorisables selon les canons de la publication scientifique.

Retour d'expérience de curation des données par les Data librarians

Une tournée de promotion de l’outil dans tous les laboratoires de Sciences Po a été organisée conjointement, au premier semestre 2021, en profitant du réseau initié par le groupe inter-labos et services transverses animé par la DRIS. Les interventions ont eu lieu en français ou en anglais dans divers cadres : assemblée générale de laboratoires (efficace), séminaires de recherche, réunion du personnel, réunion ad hoc (moins convaincante). Des guides d’aide au dépôt et à la recherche de données dans l'entrepôt ont été mis à jour et traduits en anglais pour l’occasion.
Effet de la tournée : des chercheurs pionniers ont été identifiés. Certains argumentent en faveur de l’ouverture des données collectées auprès de leurs collègues pendant les démos de l’outil, et ce dans des disciplines a priori peu favorables à l’open science pour des raisons économiques (droit). D’autres souhaitent faire apparaître dans l’entrepôt leurs nombreux jeux de données visibles dans d’autres entrepôts internationaux afin de servir d’hameçon à leurs collègues. A titre d’exemple, le dépôt d’une chercheuse, Hélène Le Bail (CERI), en amont de l’intervention dans son laboratoire a permis de nourrir des échanges entre chercheurs qui ont conduit 8 de ses collègues sur 4 projets différents à faire de même dans les mois qui ont suivi, y compris sur des volumes de données qui atteignent 2 TB pour un seul projet. 13 chercheurs de Sciences po se sont lancés dans l’aventure avant l’été ainsi que 3 chercheurs extérieurs - collaboration oblige. Les premiers dépôts ont permis de constater que parmi la centaine de fichiers disponibles, 84% sont des données, et 16% des éléments de contextualisation (fichiers “lisez-moi” par exemple). Les dépôts suivent les normes DDI, ELSST, Loterre, Unesco Thesaurus[13] Les données quantitatives sont majoritaires en nombre de jeux de données et les données qualitatives, majoritaires en nombre de fichiers. Le lien-rebond entre données d’appui et publications dans l’archive ouverte (institutionnelle Spire, nationale HAL) est pensé. Des dépôts s’enrichissent au fil des mois et des vagues d’entretien (2 fichiers en mars 2021, 21 fichiers en mai par exemple). Certains jeux de données sont très téléchargés ne serait-ce que parce qu’ils s’enrichissent au fil des vagues d’entretiens : 2 fichiers en mars 2021, 21 fichiers en mai et 1200 téléchargements [14]. Des collections sont dédiées à des projets de recherche[15] : par exemple l’*Enquête inter-associative sur l’impact de la Loi prostitution menée par Hélène Le Bail, chargée de recherche CNRS au CERI (UMR Sciences Po/CNRS).* L’objectif de cette enquête est de documenter l’impact de la loi de 2016 “visant à renforcer la lutte contre le système prostitutionnel et à accompagner les personnes prostituées” sur leurs conditions de vie et de travail : abrogation du délit de racolage, instauration d’une contravention pour l’achat de services sexuels, mise en place d’un parcours de sortie de prostitution. Au sein de l’entrepôt, l’enquête fait l’objet d’une enveloppe dédiée de 3 jeux de données : 70 entretiens et témoignages courts de travailleurs et travailleuses du sexe en 5 langues ; 25 entretiens et focus groups avec des associations de terrain ; documentation de l’enquête (protocole d’enquête, grille d’enquête, grille de questionnaire, charte d’utilisation). Tous les niveaux d’accès (ouvert, sur demande, restreint) sont ici représentés.

[13] Barone, Carlo, 2021, "Relative risk aversion models: how plausible are their assumptions ? Review of top-cited articles", https://doi.org/10.21410/7E4/UA0UKM
[14] Brouard, Sylvain; Foucault, Martial, 2020, "Citizens' Attitudes Under COVID-19 Pandemic", https://doi.org/10.21410/7E4/EATFBW
[15] Le Bail, Hélène, Giametta, Calogero, 2021, "Enquête inter-associative sur l’impact de la Loi prostitution de 2016: entretiens associations (1/3)", https://data.sciencespo.fr/dataverse/elp2016

Figure 10 : Retour d'expérience d'une chercheuse ayant pratiqué l'auto-dépôt

Hélène Le Bail à propos de sa démarche de dépôt : "Je trouve cela intéressant de rendre accessible à tout le monde les documents de mise en place de l'enquête. Cela peut permettre, d'une part, à des personnes de s'inspirer du protocole d'enquête, voire de reproduire certaines questions dans le cadre d'enquêtes similaires. Sur un sujet polémique comme celui de la législation sur le travail du sexe et sur un terrain difficile d'accès, il me semble important de rendre visible comment on peut arriver à mettre en place une enquête de terrain. Par ailleurs, mettre les données dans Data Sciences Po permet de porter à la connaissance d'équipes de recherche sur le sujet l'existence de ces données et leur composition et de leur proposer d'entrer en contact avec ceux qui ont produit ces données pour d'éventuels accords de revalorisation".

Par ailleurs, une réflexion pour inclure les DMP dans l'entrepôt est en cours : ils pourraient servir de fichiers de contextualisation à d'éventuels dépôts, voire nourrir le contenu de futurs data papers et faire gagner du temps aux chercheurs et aux chercheuses. L'accompagnement à la rédaction de DMP constitue un moment fort de sensibilisation au dépôt et aux data papers et des chercheurs de Sciences Po s'engagent à écrire des data papers dès cette phase. En outre, un réseau de correspondants Data Sciences Po dans les laboratoires est constitué et formé : le travail d'intégration des données (cf. exemples ci-dessus) est l'objet d'une collaboration entre le "data librarian" et les personnels des laboratoires : sélection des données, hiérarchisation des dossiers, nommage des fichiers, migration de formats le cas échéant, métadonnées de dépôt pertinentes, anonymisation éventuelle, conseils sur les licences de diffusion, récupération d'informations pertinentes sur les DMP. Ce réseau est alimenté par le groupe inter-laboratoires et services transverses animé par la DRIS. Ce groupe qui réunit des métiers différents (archiviste, documentaliste, statisticien, développeur, webmaster, cartographe, secrétaire générale) vise la résolution collective de problèmes rencontrés sur le terrain et a permis de thésauriser des exemples de chercheurs écrivant des data papers dans des revues à comité de lecture "sans le savoir" (Rovny et al., 2010 ; Rovny et al., 2015). Les objectifs sont là : informer sur la disponibilité du jeu de données, montrer l'originalité, la fiabilité et le potentiel de réutilisation des données, les questions de recherche à l'origine de la collecte et la plus-value de cette collecte, rendre les données intelligibles, décrire le protocole de recherche, le contexte d'obtention des données. Pourtant, ces premiers frémissements ne sont pas complètement des data papers car les données ne sont pas déposées dans un entrepôt assermenté, donc la question de l'identifiant pérenne de type DOI apposé au jeu de données ou la question des métadonnées de description ne sont pas abordées dans les articles. Il n'y a pas d'accès aux données via des hyperliens pérennes dans l'article[16].

La proposition d'un entrepôt de données au sein de l'établissement a pour beaucoup devancé les attentes. L'existence d'un réseau d'ingénieurs, de bibliothécaires travaillant sur les données et le travail de la DPO autour du RGPD ont permis de développer une curiosité pour ces questions ; aujourd'hui encore pourtant, de nombreux chercheurs se sentent éloignés de ces enjeux et technologies. Un premier frein est le manque d'intérêt, indépendamment des incitations des financeurs, pour entreprendre ces démarches ; quel public si on ouvre les données ? La réutilisation de nos matériaux, notamment qualitatifs, est-elle possible et souhaitable ? Cela demande également des moyens parfois très élevés pour traiter les matériaux. Aujourd'hui, il apparaît d'autant plus difficile de faire de la documentation de qualité que nos ressources en interne pour ce faire restent limitées, ce qui implique d'associer

[16] Quelques freins identifiés : publication, chargement de formats ouverts, volumétrie, affichage en front office non conformes au back office, messages d'erreurs erronés. Un conseil : s'assurer l'appui de sa DSI, réfléchir au lien avec l'entrepôt national et EOSC.

davantage les chercheurs. D'où notre mouvement actuel en direction des data papers[17].

Faire des data papers une opportunité pour documenter les jeux de données, un véritable défi

Si de plus en plus de chercheurs se montrent intéressés par une démarche de partage de leurs jeux de données, un point problématique reste la documentation des données : comment favoriser cette dernière sans que cela soit trop contraignant à la fois pour les chercheurs déposants et pour nous-mêmes ? Une piste que nous cherchons actuellement à explorer consiste à favoriser la pratique de rédaction de data papers.
Les avantages du format data paper, véritable publication valorisable dans un dossier de carrière, qui permet d'ancrer la citation du jeu de données dans les canons du champ scientifique, sont connus. A la différence de ce qui se passe lorsqu'on demande aux chercheurs de fournir des informations sur leurs données pour créer l'enquête sur l'enquête à beQuali ou remplir des notices DDI, la perspective est désormais celle d'une publication en propre, avec un meilleur retour sur investissement, plus immédiat - ainsi le dépôt ne pourrait pas être perçu comme du temps perdu puisqu'il y aurait au moins une publication à la clef. Les data papers peuvent aussi être vus comme un moyen permettant aux chercheurs de s'interroger sur les conditions de réutilisation et la richesse de leurs données. De nombreux chercheurs, en particulier parmi ceux qui pratiquent les méthodes qualitatives, apparaissent en effet assez perplexes, parfois sceptiques, quant aux possibilités que d'autres puissent réutiliser de façon pertinente leurs matériaux. Le fait que ces derniers puissent sécuriser le réusage de leurs données en produisant un data paper, peut être un atout en leur redonnant une place active dans ce processus[18]. De plus, le temps passé à réfléchir aux différents aspects de la gestion des données pendant la rédaction du DMP (obligatoire), permet, on l'a dit, d'en gagner au moment du dépôt en entrepôt et de la rédaction de data papers. Une condition reste entière toutefois : concrètement les chercheurs doivent avoir à leur disposition un vrai modèle opérationnel, qu'ils peuvent investir plus ou moins en autonomie, et qui reste limité du point de vue de l'ampleur du travail exigé, ainsi que des débouchés éditoriaux. Or, la situation apparaît un peu plus compliquée en pratique. Nous allons ici dire quelques mots des pistes que nous explorons actuellement pour dépasser deux écueils : la faible visibilité des data papers dans les communautés de recherche, en particulier dans les revues de sciences sociales, et la faible consistance de ce nouveau genre de publication.

Les data papers : quelle réalité dans les revues de sciences sociales ?

De manière générale, les chercheurs que nous rencontrons en ont globalement peu entendu parler. Les data papers, les chercheurs n'en voient pas passer dans les revues qu'ils lisent, n'en entendent pas souvent parler en séminaire, en colloque, ou dans les comités éditoriaux dont ils sont membres, et ne les voient pas mis en avant dans les critères d'évaluation de leurs établissements de tutelle. Dans les laboratoires, mis à part quelques chercheurs particulièrement sensibilisés à ce thème, ou quelques rares curieux, le sujet est principalement mis à l'ordre du jour par les ingénieurs. Plus encore, évoquer les data papers est parfois perçu comme revenant

[17] La définition usuelle des data papers prend ici tout son sens. Par exemple, selon le site du CoopIST, le data paper est un format d'article qui "informe la communauté scientifique de la disponibilité de ces jeux de données et de leur potentiel pour des utilisations futures. Contrairement à un article de recherche classique, le data paper décrit uniquement des données scientifiques et les circonstances et méthodes de leur collecte." (https://coop-ist.cirad.fr/gerer-des-donnees/rediger-un-data-paper/1-qu-est-ce-qu-un-data-paper, consulté le 28/09/2021).
[18] Basculer vers les data papers peut aussi être vu comme un gain pour les ingénieurs qui auront accès à ces informations sans avoir besoin d'interviewer les producteurs, et qui pourront consacrer leur temps à d'autres aspects liés à l'ouverture des données de la recherche, à moyens humains constants.

à ajouter une nouvelle couche de contraintes à un principe, l'ouverture des données de la recherche, qui est déjà souvent perçu comme une contrainte administrative. Tout l'enjeu ici consiste à faire en sorte que les data papers soient vus comme une solution à la problématique de l'ouverture des données de la recherche. Pour cela, encore faut-il avoir des références à suggérer, des modèles à montrer, sur lesquels les chercheurs vont pouvoir s'appuyer - ce qui est tout sauf évident.

En interne, au CDSP, nous avons certes des éléments qui pourraient s'apparenter à un modèle possible, basé sur "l'enquête sur l'enquête" développée pour les besoins de beQuali. Cependant, en l'état, il s'agit d'un exercice qui est beaucoup plus fouillé et volumineux que ne l'est un data paper[19], donc non publiable en l'état dans une revue, et difficilement reproductible à moindre coût ; de plus il manque une étape importante, qui est celle de la stabilisation du genre éditorial - les discussions sur les contours d'une enquête sur l'enquête n'ayant que rarement dépassé le périmètre de l'équipe du CDSP (Bendjaballah et al., 2017). Si on élargit la perspective pour regarder ce qui se fait autour de nous, le constat est simple, et connu : on trouve peu de spécimen de data papers en sciences sociales qui pourraient servir d'exemples, en tout cas en langue française[20] ; la littérature spécifiquement consacrée à discuter les contours de ce genre rédactionnel est encore très embryonnaire ; et il y a très peu de data journals, ce qui n'aide pas à accueillir des data papers ni à des préconisations éditoriales pour ce type de publication[21].

Face à cette situation, nous sommes partis dans plusieurs directions. La première stratégie a consisté à essayer de trouver des débouchés éditoriaux pour les chercheurs de Sciences Po qui désireraient se lancer dans la rédaction de data papers, et pour cela avoir une vision plus claire des politiques des revues en matière de données dans nos champs disciplinaires. A titre expérimental, nous avons commencé à consulter les sites des revues dans lesquelles publient les sociologues de nos unités. Nous avons cherché à voir si elles demandent les données liées aux articles, si elles encouragent leur dépôt et si elles acceptent des soumissions de data papers. Ce travail étant encore en cours, nous ne pouvons restituer ici que des proto-résultats. A ce stade, nous avons trouvé un faible nombre de revues qui acceptent officiellement les soumissions de data papers. On peut supposer qu'à l'échelle de la discipline, les tendances sont les mêmes et que ces pratiques sont très peu développées. Ce constat ne semble d'ailleurs pas propre à la sociologie et est sans doute valable dans les autres disciplines de SHS. Il est donc difficile de s'appuyer sur ce levier pour sensibiliser réellement les chercheurs à ce sujet. Néanmoins, on observe que les choses évoluent : on voit par exemple que de nouvelles revues de sciences sociales prennent position en ce domaine[22], d'autres revues encourageant le dépôt de données dans des entrepôts ou demandant des Data availability statements[23].

A terme, nous nous interrogeons sur la manière de donner à voir, aux communautés, les initiatives prises par des revues en matière de données. Si nous voulons accompagner, voire accélérer les opérations de partage, il nous semble important de valoriser des politiques volontaristes et de ne pas attendre que ces pratiques se diffusent (forcément lentement) dans ces communautés. Nous devrons également, si nous poursuivons ce travail sur les revues, analyser les différences entre les disciplines afin de pouvoir ajuster notre discours aux

[19] Les différents rapports produits dans le cadre de beQuali ont un volume conséquent, la plupart du temps compris entre 100 000 et 150 000 signes, ce qui équivaut à la taille de deux ou trois articles "classiques", dépassant ainsi très largement la taille standard des data papers en sciences sociales, souvent limitée à quelques dizaines de milliers de signes.

[20] Ou du moins rédigés par des chercheurs et chercheuses français, auxquels la communauté nationale pourrait davantage s'identifier. Quelques cas existent toutefois. Pour un exemple de data paper basé sur une base de données développée au CDSP, voir Dehousse et al., 2017. Pour un exemple très récent, voir Gay, 2021.

[21] Sur ces points voir notamment Schöpfel et al., 2020. En consultant les bases de données bibliographiques du Web of Sciences et de Scopus, nous avons nous-mêmes pu constater que très peu de data papers en SHS étaient publiés, et ce même en tenant compte des biais de ces bases bibliographiques.

[22] Par exemple récemment la Revue française de sciences de l'information et de la communication (Le Deuff, 2018).

[23] JCMS, European Law Journal, Oxford University Press (exemples tirés d'un DMP Sciences Po sur un projet de droit).

différentes sous-communautés de la recherche à Sciences Po.

Stabiliser un genre rédactionnel nouveau, trouver des débouchés éditoriaux : la pertinence d'une expérimentation collective

En parallèle de cette démarche, nous sommes en train d'essayer de monter une expérimentation collective visant à contribuer à stabiliser ce genre éditorial nouveau et à trouver des débouchés éditoriaux adaptés. Ce projet a été stimulé par l'appel à projet 2021 du Fonds national pour la science ouverte, centré sur le thème des "publications", qui nous a incité à nous mobiliser pour combler certains angles morts de la dynamique d'ouverture des données de la recherche en SHS, que nous avions pu constater au fil de nos explorations : le manque de spécimens concrets et diversifiés de data papers ; le manque de guidelines et de retours d'expériences permettant de circonscrire un genre de publication encore très nouveau ; le manque d'offre éditoriale dédiée pour publier les data papers. Nous avons commencé à nouer des liens avec des partenaires de plusieurs Universités (Grenoble, Gustave Eiffel, Lille) et à l'INRAE pour tenter de construire un dispositif destiné à accompagner des chercheurs de sciences sociales qui souhaitaient produire des data papers décrivant des jeux de données déjà déposés, ou devant être déposés dans chaque entrepôt institutionnel (tous sont basés sur la solution Dataverse). L'ambition d'un tel projet est d'aboutir à proposer des modèles et des bonnes pratiques pouvant servir de ressources futures pour une communauté élargie de sciences sociales. Des objectifs plus larges sont proprement éditoriaux : poser les bases d'un data journal francophone en sciences sociales, mais aussi dialoguer avec des revues du champ pour voir comment les aider à intégrer ce nouveau type d'articles, les data papers, à leur politique éditoriale.

En guise de conclusion, (re)penser la question du lectorat des data papers

Avant de clore cette contribution, nous voulions élargir la perspective en attirant l'attention sur une question qui semble le plus souvent absente des réflexions sur les data papers, celle de leur lectorat - et derrière elle, les formes de réutilisation qui sont anticipées et donc privilégiées lorsque les jeux de données sont partagés. Il n'est pas anodin de savoir si les data papers sont adressés par exemple en priorité à des étudiants ayant besoin de comprendre les données qu'ils manipulent pour se former aux sciences sociales, et notamment comprendre comment les chercheurs procèdent méthodologiquement dans leurs enquêtes, ou en priorité à des chercheurs désirant faire de l'analyse secondaire. Le degré de précision du data paper, sur l'état de l'art relatif au thème de la recherche, sur la réalisation du terrain, sur l'organisation des données, etc. ne sera pas le même selon qu'on pense s'adresser, en aval, à des étudiants profanes ou à des chercheurs spécialisés, étant entendu qu'entre les deux il y a évidemment tout un continuum. Clarifier le lectorat cible permettra aux chercheurs de mieux imaginer les réanalyses possibles et contribuera à façonner des modèles éditoriaux.

Références bibliographiques

Bendjaballah, Selma, Guillaume Garcia, Sarah Cadorel, Emilie Groshens, Emilie Fromont et Emeline Juillard. 2017. "Valoriser les données d'enquêtes qualitatives en sciences sociales : le cas français de la banque d'enquête beQuali". *Documentation et bibliothèques* 63 (4) : 73-85. https//doi.org/10.7202/1042306ar

Dehousse, Renaud, Selma Bendjaballah, Geneviève Michaud, Olivier Rozenberg, Florence Deloche-Gaudez, Giuseppe Ciavarini Azzi, Olivier Costa et Romain Lalande. 2017. "L'Observatory of European Institutions: Une base de données sur le processus décisionnel

dans l'Union européenne (1996-2014)". *Politique européenne* 58 : 14-42. https//doi.org/10.3917/poeu.058.0014

Duchesne, Sophie, Guillaume Garcia, Anne Both et Sarah Cadorel. 2014. "Retour vers le futur : la numérisation des enquêtes qualitatives de sciences sociales entre patrimonialisation et transformation des pratiques scientifiques", 20 février 2014. https://humanum.hypotheses.org/147.

Gay, Victor. 2021. "Mapping the Third Republic: A Geographic Information System of France (1870–1940)". *Historical Methods: A Journal of Quantitative and Interdisciplinary History*. https//doi.org/10.1080/01615440.2021.1937421

Le Deuff, Olivier. 2018. "Une nouvelle rubrique pour la RFSIC : le data paper". *Revue française des sciences de l'information et de la communication* 15. htttp://journals.openedition.org/rfsic/5275

Rovny, Jan et al. 2010. "Reliability and validity of the 2002 and 2006 Chapel Hill expert surveys on party positioning". *European Journal of Political Research* 49 : 687–703. https//doi.org/10.1111/j.1475-6765.2009.01912.x

Rovny Jan et al. 2015. "Measuring party positions in Europe: The Chapel Hill expert survey trend file, 1999–2010". *Party Politics* 21 (1) 143–152. https//doi.org/10.1177/1354068812462931

Schöpfel, Joachim, Dominic Farace, Hélène Prost et Antonella Zane. 2020. "Data papers as a new form of knowledge organization in the field of research data". Knowledge organization, ergon verlag 46 (8) : 622-638. http://dx.doi.org/10.5771/0943-7444-2019-8-622