Le data paper, de la littérature grise à l'article de données. Retour sur l'expérience du CDSP et de l'accompagnement des chercheurs à Sciences Po

Anna Egea, Guillaume Garcia

Le data paper, de la littérature grise à l'article de données
Retour sur l'expérience du CDSP et de l'accompagnement des chercheurs à Sciences Po

Anna Egea, CSO, CNRS Sciences Po
Guillaume Garcia, CDSP, Sciences Po

Journée d'étude "Un *data journal* interdisciplinaire pour les sciences humaines et sociales - Enjeux scientifiques et mise en œuvre pratique"
Université de Nancy, 10 mars 2023

Plan de l'intervention

Partie 1 : Retour sur l'expérience du CDSP avant l'émergence des data papers

Partie 2 : Retour sur la mise en place d'un service d'accompagnement aux data papers à Sciences Po

Conclusion : Quels enjeux autour des data papers ?

Partie 1

Retour sur l'expérience du CDSP avant l'émergence des data papers

Le contexte originel : la diffusion des enquêtes à Sciences Po

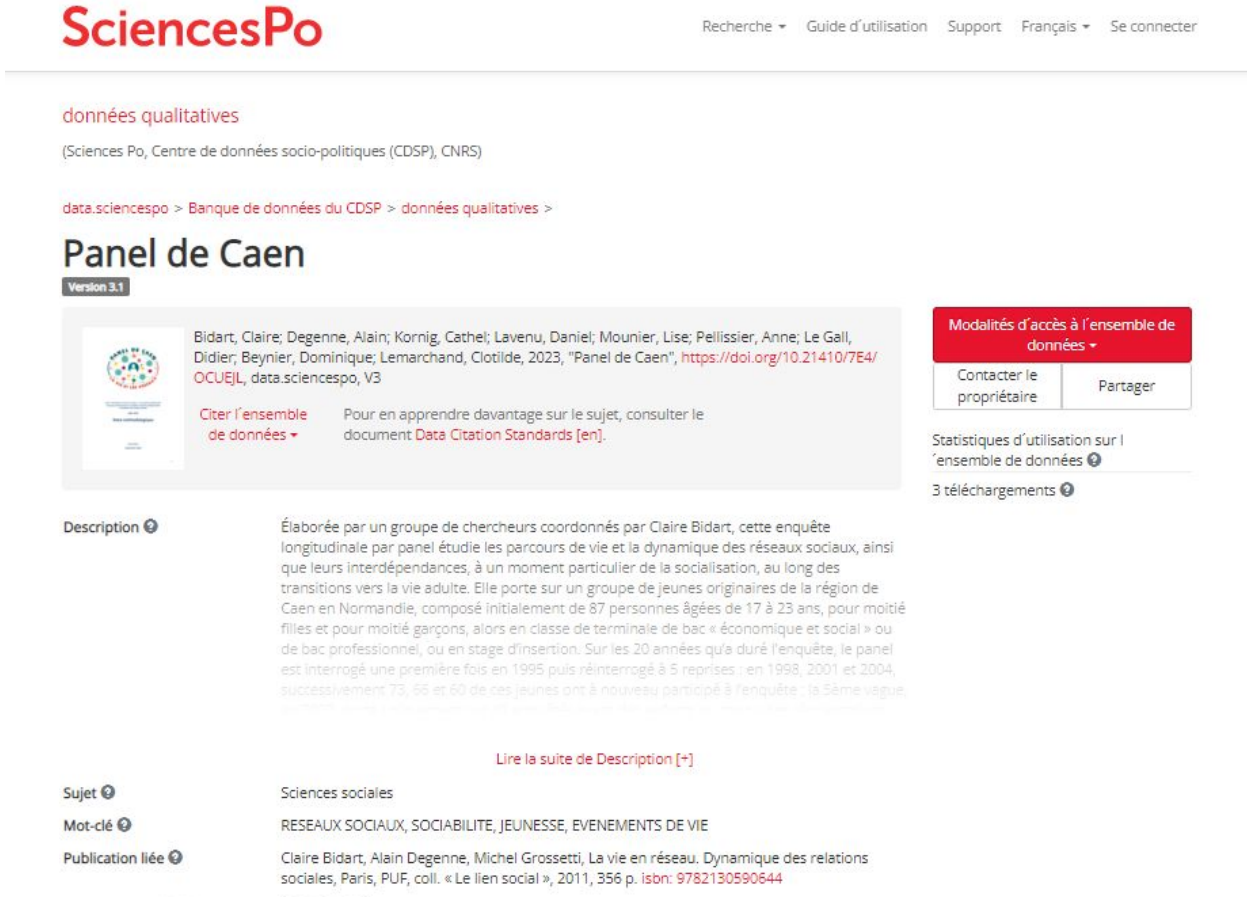
Depuis 2006 au CDSP - essentiellement via des ingénieurs BAP D en documentation

Au départ : données "quantitatives"

A partir de 2013, intégration des données "qualitatives"

Des modèles de documentation mis en oeuvre avant l'émergence des data papers

La banque de données du CDSP sur Datasciencespo



Apports et limites de la documentation "classique" des données

L'enquête, en tant que processus de recherche global, est documentée avec une notice, via des standards de métadonnées de type DDI, ELSST, etc.

Les métadonnées d'une notice d'enquête

SciencesPo

Recherche ▾ Guide d'utilisation Support Français ▾ Se connecter

Métadonnées géospatiales

Couverture géographique	France (la), Normandie, Caen
Unité géographique	Région

Métadonnées sur les sciences sociales et les sciences humaines

Unité d'analyse	Individu
Univers	Panel de 87 jeunes de la région de Caen regroupant en 1995 : des lycéens de section ES, âgés de 17 à 20 ans ; des lycéens en bac. professionnel, âgés de18 à 23 ans ; des stagiaires, âgés entre 18 et 24 ans .
Méthode temporelle	Longitudinale : cohorte / basée sur un événement
Responsable de la collecte de données	Bidart, Claire
Fréquence	Tous les trois ans pour les cinq premières vagues, puis huit ans plus tard pour la dernière.
Taille de l'échantillon cible	87
Mode de collecte	Entretien en face-à-face
Type d'instrument de recherche	Questionnaire semi-structuré
Caractéristiques de la collecte de données	Enquête réalisée sur 6 vagues, entre 1995 et 2015, à travers des questionnaires et des entretiens auprès des 87 enquêtés. L'étude se déroulant sur le temps long, l'enquête subie des attritions et le nombre d'enquêté répondant varie selon les vagues. Les élèves ont été sensibilisés au projet en classe, et devaient ensuite se porter volontaires pour participer au panel. Les volontaires étaient contactés par téléphone afin de procéder au premier entretien. Les élèves étaient recrutés au sein de trois formations différentes : bac ES, bac professionnel, et stagiaires travaillant déjà.
Mesures visant à minimiser les pertes	Des nouvelles régulières étaient échangées avec les panélistes afin d'éviter l'attrition.
Pondération	Il n'existe aucune pondération pour ces données qui n'ambitionnent pas d'être représentatives.
Remarques	Nombre de documents : 926 fichiers numériques ou numérisés. Les documents n'ayant pas été numérisés restent consultables aux archives MSH de Caen Transcription : Documents textuels. Langage documentation : Français. Anonymisation : Anonymisation réalisée par le chercheur déposant et l'équipe du CDSP.

Apports et limites de la documentation "classique" des données

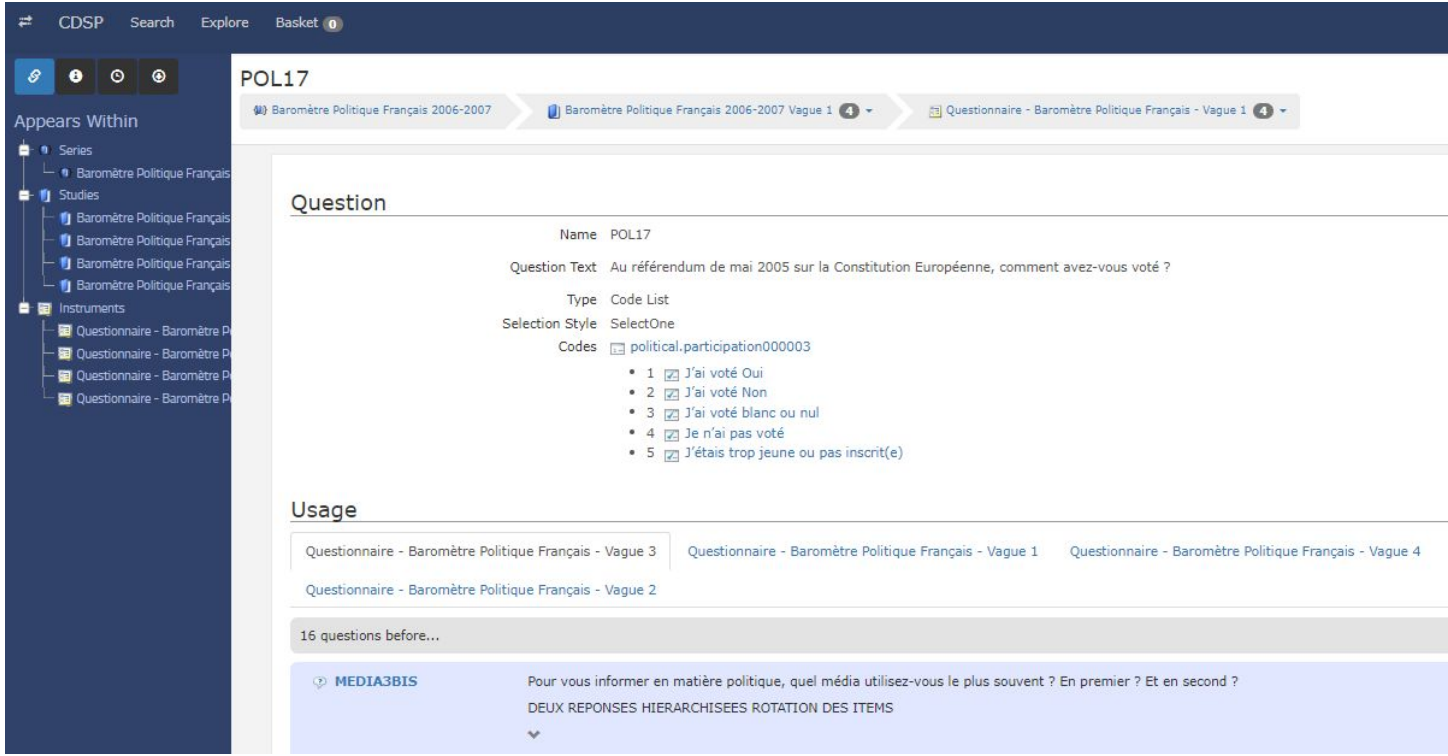
L'enquête en tant que processus de recherche global est documentée via des standards de métadonnées de type DDI, ELSST, etc.

Les descriptions détaillées des questions et des variables sont faites soit dans les fichiers eux-mêmes soit dans des documents à part.

On a ainsi accès à des informations sur la position de la question dans le questionnaire, les filtres dans le questionnaire, les techniques de recodage, les consignes données aux enquêteurs, etc.

La documentation fine des questions et variables

https://explore.cdsp.sciences-po.fr/



La description papier de l'ensemble des variables d'un questionnaire

Etats-Unis

Emmanuel LAZEGA, Fondation Nationale des Sciences Politiques (FNSP)

Le phénomène collégial

Study Documentation

February 13, 2020

Table of Contents

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.

Apports et limites de la documentation "classique" des données

L'enquête en tant que processus de recherche global est documentée via des standards de métadonnées de type DDI, ELSST, etc.

Les descriptions détaillées des questions et des variables sont faites soit dans les fichiers eux-mêmes soit dans des documents à part ; on a ainsi accès à des informations sur la position de la question dans le questionnaire, les filtres dans le questionnaire, les techniques de recodage, les consignes aux enquêteurs, etc.

Beaucoup d'informations sont présentes.

L'inconvénient est qu'elles sont disséminées entre plusieurs documents / sources.

La globalité et des détails du processus de recherche doivent être recherchés dans les publications.

Apports et limites de l'"enquête sur l'enquête" (pour les entretiens)

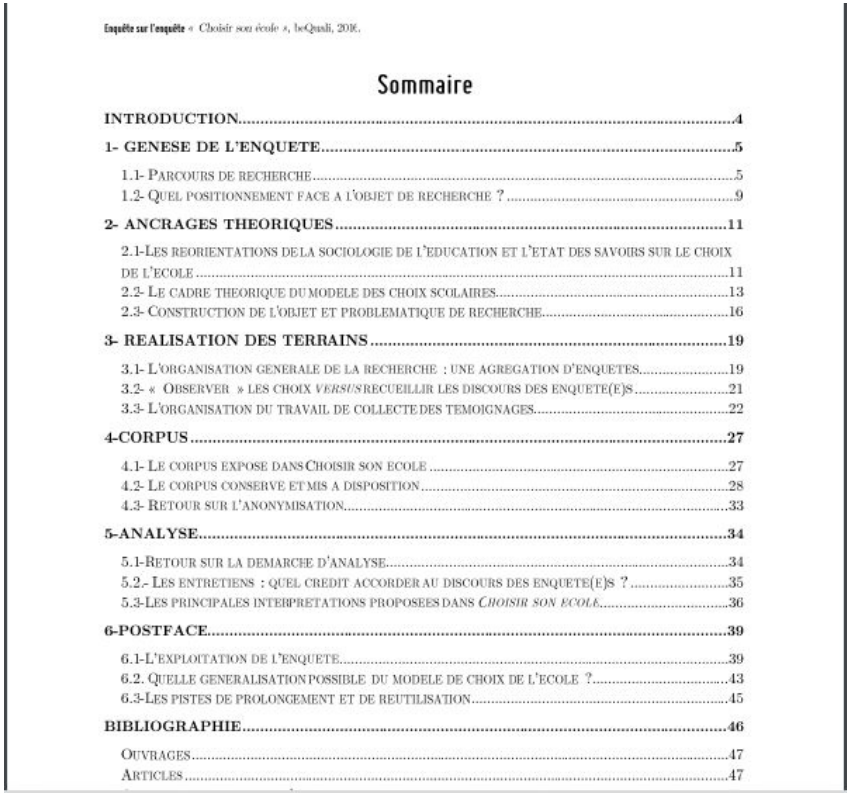
Idée = rassembler, dans un document unique, les informations nécessaires à la compréhension & manipulation du corpus de données issues d'un processus de recherche

On est relativement proche du format data paper dans la structuration des rubriques & dans la forme littéraire

Cela donne des rapports de plusieurs dizaines de pages (2 à 3 fois la taille d'articles standards)

Page d'accueil de l'enquête sur l'enquête "Le choix de l'école"

Sommaire du rapport de l'enquête sur l'enquête "Le choix de l'école"



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

Sommaire

Premiers bilans

On se situe dans le modèle du "**faire pour - et - avec**"

Un **coût d'entrée élevé** pour les ingénieurs extérieurs à l'enquête & un coût de documentation globalement élevé

Une **nécessaire collaboration** - même minimale - avec le chercheur déposant (éviter la mise à distance)

Les **limites de l'exercice**

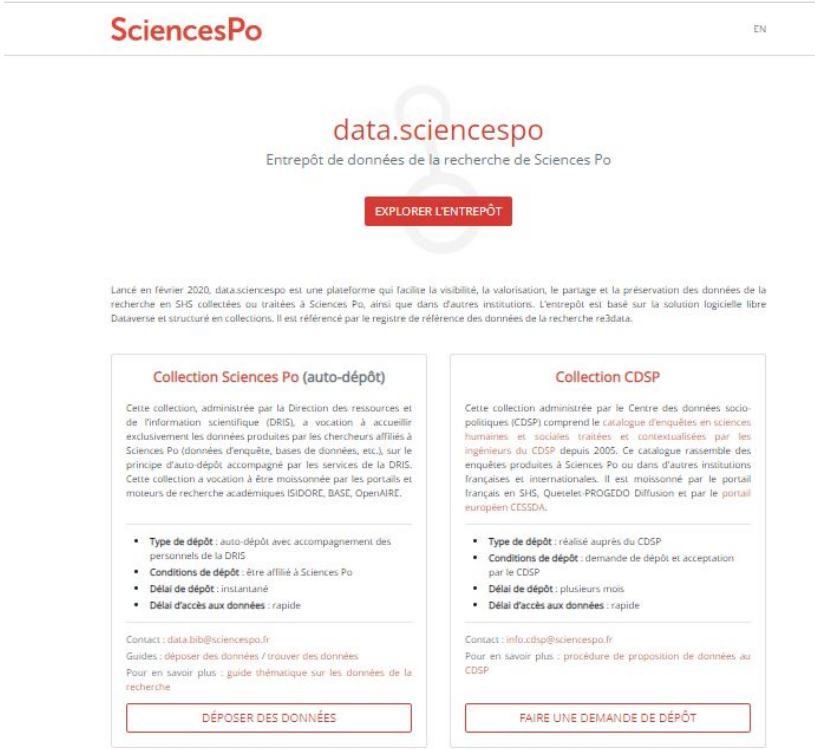
- quelle posture adopter ? posture compréhensive *versus* posture critique
- jusqu'où aller dans la recontextualisation ?

Les **freins de publicisation** liés au format "littérature grise"

Partie 2

Retour sur la mise en place d'un service d'accompagnement à Sciences Po

La collection en auto-dépôt de data.sciencespo



Création d'un groupe de travail dédié

Un **groupe multi-métiers** (data librarians, documentalistes, ingénieurs BAP D)

But : mettre en capacité des ingénieurs de laboratoires, de la bibliothèque de Sciences Po, du CDSP
d'**orienter** et de **conseiller** les chercheurs (choix de revues cibles, aide ou participation à la rédaction de data papers)
<u>après</u> ou <u>au moment</u> d'un dépôt d'enquête

Les linéaments et les contours de ce projet sont décrits dans un chapitre à paraître dans les actes du colloque #DHNord 2021 sur les data papers

Les premières actions menées

Etat des lieux des pratiques en vigueur dans l'écosystème éditorial impliquant la science politique et la sociologie principalement

A partir d'un **échantillon raisonné** de data journals anglophones ou des revues académiques plus classiques (disciplinaires ou méthodologiques) en France ou dans le monde anglophone

Démarche incrémentale : principales revues dans lesquelles les chercheurs de Sciences Po publient & des Data journals ou revues disciplinaires qui se font fait connaître sur le sujet

Principe : aider à **déterminer la cible** aisément (positionnement de la revue, coûts, particularités du processus d'évaluation, etc.)

Articulation de 2 logiques :

- une **logique réactive** : conseiller les (rares) chercheurs déjà informés et motivés
- une **logique proactive** : se mettre en capacité de promouvoir cette possibilité auprès de chercheurs plus distants

Les actions en cours

Mise en place de **tutoriels / guidelines**

- "scripts" sur les plans, informations demandées, points d'attention, check lists… permettant d'alerter sur les besoins d'élaboration & de rédaction
- mise en place de "modèles" de DP servant de cadres de référence
- sert de "maïeutique" dans un échange entre chercheur et ingénieur

Mise en place de **tests en grandeur nature** avec des "cobayes"

Les limites principales de cette démarche : **encore peu de débouchés** (gratuits) dans les revues disciplinaires ou les Data Journals, même si dans le même temps il y a encore peu de candidats
⇒ quel degré d'adéquation entre l'offre et la demande ?

Conclusion

Quels enjeux autour des data papers ?

Un flou sur les contours éditoriaux des data papers

Une difficulté à **formaliser un modèle de data paper** pour les disciplines visées, la différence et l'articulation avec la forme "article de recherche" n'étant pas encore très lisibles

Une question non résolue : le **statut de cette publication**, de l'évaluation dont elle peut faire l'objet, en lien avec le ou les articles scientifiques consécutifs de la même recherche

Question du **support de publication**, des rôles respectifs des revues classiques et des data journals, généralistes ou spécialisés, avec un risque d'effet de niche

Une question souvent éludée : quel(s) l**ectorat(s) cible(s)** des data papers, en lien avec les **pistes de réutilisation** fournies par les auteurs - recherche *versus* pédagogique

Un flou sur le périmètre méthodologique des data papers

Un **tropisme "quanti"**, lié aux différences d'ancrage de la pratique de partage / réutilisation dans les différentes cultures professionnelles

Quelle place pour les données "quali" ?

Un **point aveugle** : quid des données du Web ?

Une question centrale mais souvent évitée : les coûts humains de la rédaction des DP

Quel **bilan coûts / avantages** impliqués par la production des data papers, pour les chercheurs ou les ingénieurs en soutien ?

Quelles formes de **division du travail entre chercheurs et personnels IST** (archivistes, data librarians, documentalistes, ingénieurs Bap D) ?

Vers un changement du **positionnement des ingénieurs** face aux chercheurs en matière d'ouverture des données de la recherche ? en particulier s'agissant de la nature des services qu'ils rendent

Merci de votre attention

anna.egea@sciencespo.fr
guillaume.garcia@sciencespo.fr