

Préserver et réutiliser les données de la recherche : regards croisés d'ingénieurs données et d'experts en numérisation à Science Po

Comment garantir une réutilisation efficace et pérenne des données de la recherche ? Un éclairage concret sur l'importance des standards de métadonnées et de la numérisation des matériaux de recherche.

L'Organisation de coopération et de développement économiques (OCDE) définit les données de la recherche 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 »¹.

Après la publication des deux Plans nationaux de la Science ouverte en France², la réutilisation des données de la recherche s'impose aujourd'hui comme un enjeu central pour les Sciences humaines et sociales (SHS). Elle transforme les pratiques scientifiques, favorise la collaboration interdisciplinaire et ouvre la voie à de nouvelles formes de cumulativité des savoirs. À Sciences Po, université de recherche internationale, cette dynamique s'appuie, entre autres, sur deux piliers complémentaires : la documentation, diffusion et préservation des données

issues de méthodes quantitatives et qualitatives au sein du Centre de données socio-politiques (CDSP), et la numérisation des archives de la recherche portée par la Direction des ressources et de l'information scientifique (DRIS). Ensemble, ces expertises contribuent à rendre les données plus accessibles, interopérables et réutilisables, conformément aux principes FAIR³ (*Findable, Accessible, Interoperable, Reusable*), tout en respectant des standards internationaux de valorisation des données de la recherche, comme *Data Documentation Initiative* (DDI)⁴.

RENDRE ACCESSIBLES ET RÉUTILISABLES LES DONNÉES DE LA RECHERCHE

Réutiliser⁵ des données, c'est permettre à des chercheurs, enseignants ou étudiants d'exploiter des matériaux produits dans un autre contexte scientifique, pour répondre à de nouvelles questions, vérifier des résultats ou enrichir des corpus existants. Cette pratique suppose que les données soient non

[1] Consulter le rapport « Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics » (2007) https://www.oecd.org/fr/publications/2007/04/oecd-principles-and-guidelines-for-access-to-research-data-from-public-funding_g1gh7fe5.html

[2] Consulter le Plan national pour la Science ouverte. MESRI (2018-2021) <https://www.enseignementsup-recherche.gouv.fr/fr/bilan-du-plan-national-pour-la-science-ouverte-des-engagements-tenus-des-avancees-majeures-realisees-48296> et le Deuxième Plan national pour la Science ouverte. MESRI (2021-2024) <https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte>

[3] Wilkinson, Mark D., Michel Dumontier, et al. « The FAIR Guiding Principles for Scientific Data Management and Stewardship. » (2016)

[4] Consulter « Data Documentation Initiative : Metadata Specification for the Social, Behavioral, and Economic Sciences » DDI Alliance (202) <https://ddialliance.org>

[5] « Data reuse is use of data one which did not collect oneself » (« La réutilisation des données est une utilisation de données que le chercheur n'a pas collecté lui-même »). Fear, KM. 2013. *Measuring and anticipating the impact of data reuse*, PhD thesis, University of Michigan. <https://deepblue.lib.umich.edu/handle/2027>



Crédit photo Adobe Stock - Björn Wylezich

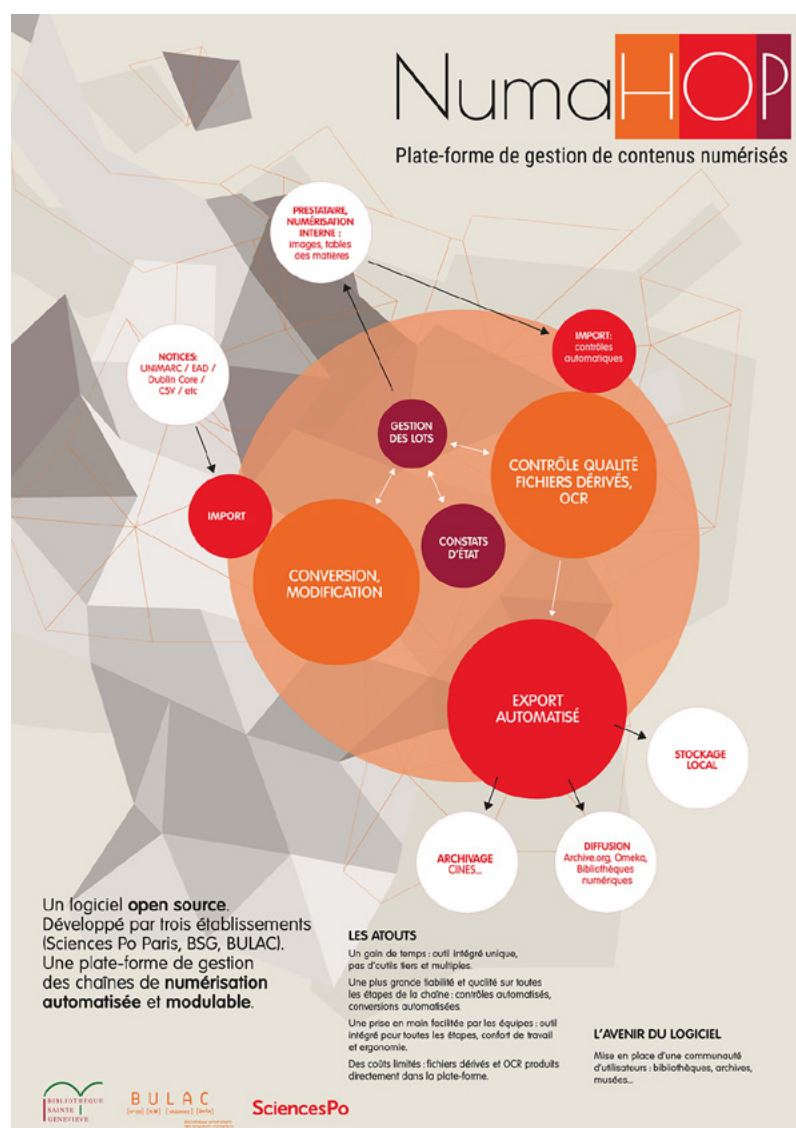
seulement disponibles, mais aussi compréhensibles, bien documentées et techniquement exploitables. Dans les SHS, où les données sont souvent hétérogènes en termes de méthodes ou de supports disponibles (enquêtes quantitatives, entretiens, archives, corpus numérisés...), la réutilisation exige un travail méthodologique et technique conséquent. La réutilisation ne peut être effective que si les données respectent un ensemble de standards. Les principes FAIR constituent aujourd'hui le cadre de référence international. Ils impliquent, entre autres :

- Une description riche et normalisée des données en respectant des standards comme DDI
- L'usage de formats ouverts et pérennes pour les données et les métadonnées
- Une structuration permettant l'interopérabilité entre plateformes et surtout la mise en place de plateformes de confiance
- Une documentation complète sur les méthodes de collecte, de traitement et de contextualisation.

À Sciences Po, un exemple de ce type de travail conjoint est la collaboration entre ingénieurs en *data management*, également responsables d'entrepôts de données⁶ et experts de la numérisation qui gèrent des plateformes de numérisation et de diffusion. L'objectif est de garantir que les données puissent circuler entre disciplines, infrastructures et communautés scientifiques. Dans la pratique, cela signifie par exemple harmoniser les métadonnées d'une enquête électorale de 1988 avec celles d'un panel contemporain, ou encore aligner les formats de numérisation de carnets de terrain provenant de laboratoires différents.

L'APPORT DU CDSP : VALORISER, DIFFUSER ET PRÉSERVER LES DONNÉES SOCIO-POLITIQUES

Le CDSP, unité d'appui et de recherche de Sciences Po, joue un rôle clé dans la mise en œuvre de ces standards. En effet, le CDSP documente et diffuse des données de la recherche depuis 2005 dans le respect des standards internationaux comme le DDI, mène des projets de recherche sur la thématique de la réutilisation des données comme le projet FAIRwDDI⁷, financé par l'Agence Nationale de Recherche. Les données quantitatives produites ou archivées au CDSP sont régulièrement utilisées dans des publications scientifiques ou des enseignements. Un exemple est la réutilisation des enquêtes électorales diffusées par le CDSP dans l'ouvrage *"Controlling the Electoral Market Place"*⁸, de Joost van Spanje. Ce livre étudie la réaction des partis politiques établis face aux partis d'extrême gauche et d'extrême droite qui ont fait un bond dans de nombreuses démocraties à travers le monde. Les réutilisations montrent que la valeur d'un jeu de données dépasse largement son objectif initial : une enquête ou jeu de données peut servir à analyser la



thématique de recherche initialement conçue par ses auteurs, mais elle peut avoir d'autres usages de réutilisation pour la recherche ou l'enseignement, dès lors qu'elle est correctement documentée et accessible. Par ailleurs, ouvrir un jeu de données a de nombreux avantages pour l'auteur et favorise la citation de ses travaux.

Malgré des moyens humains et financiers limités, le CDSP renforce sa politique de diffusion des données en formant et en accompagnant la communauté à la mise en œuvre de pratiques de valorisation des données conformes aux principes FAIR, pour favoriser la réutilisation des données. Le CDSP contribue ainsi à structurer et professionnaliser la gestion des données, en mettant l'accent sur l'importance des métadonnées structurées type DDI, encore méconnues dans le paysage français. Des efforts restent à accomplir dans ce domaine.

[6] Un exemple d'entrepôt est la Banque de données du CDSP, certifiée CoreTrustSeal, qui met à disposition de la communauté académique résultats d'enquêtes et jeux de données socio-politiques

[7] FAIRwDDI, Implémentation des principes FAIR avec DDI : <https://www.sciencespo.fr/cdsp/fr/projets/projets-en-cours/fairwithddi>

[8] Van Spanje, Joost. *Controlling the Electoral Marketplace: How Established Parties Ward Off Competition*. Springer International Publishing (2017)

L'APPORT DE LA DRIS : NUMÉRISER, PRÉSERVER ET RENDRE EXPLOITABLES LES ARCHIVES DE LA RECHERCHE

La DRIS collecte, signale, diffuse et conserve les ressources et matériaux pour l'enseignement et la recherche. Dans le cadre de la préservation des données, elle intervient par exemple au niveau de la numérisation et le traitement post-capture. En effet, la réutilisation ne concerne pas uniquement les données structurées, elle dépend aussi de la disponibilité des archives de la recherche, souvent dispersées, fragiles ou difficilement exploitables. Le service de numérisation et d'archivage numérique de la DRIS jouent ici un rôle déterminant. Ce travail s'appuie notamment sur NumaHOP⁹, la plateforme de pilotage des chaînes de numérisation utilisée pour assurer la traçabilité, la qualité et la cohérence des traitements appliqués aux archives de la recherche. Grâce à cette plateforme, chaque étape du processus de numérisation est documentée, ce qui garantit la reproductibilité des opérations et facilite l'intégration des fichiers dans des infrastructures de diffusion ou d'archivage pérenne. Cette traçabilité est essentielle pour assurer la confiance dans les données numérisées et permettre leur réutilisation dans des contextes scientifiques variés.

En numérisant les carnets de terrain, les questionnaires papier, les corpus iconographiques ou les documents administratifs liés aux projets de recherche, la DRIS contribue à préserver des matériaux essentiels et à les rendre accessibles dans des formats pérennes. Un exemple emblématique est la numérisation d'archives d'enquêtes qualitatives menées dans les années 2010 et 2020. Dans l'un des projets, l'équipe avait traité plus de 3 000 pages de carnets manuscrits provenant d'une enquête sociologique sur les mobilisations étudiantes. Ces documents, longtemps restés dans des fonds personnels, deviennent exploitables pour de nouvelles recherches dès lors qu'ils sont numérisés, décrits et intégrés dans des *workflows* de préservation. Leur mise à disposition ouvre des perspectives inédites pour l'histoire des sciences sociales, la sociologie politique ou l'analyse des politiques publiques. Autre exemple : la numérisation d'un corpus photographique produit dans le cadre d'un projet sur les transformations urbaines. Les images, initialement conservées sur des supports hétérogènes, ont été harmonisées, indexées et intégrées dans une chaîne de traitement permettant leur diffusion¹⁰ dans des formats ouverts.

Si les infrastructures existent, les besoins restent considérables. Les enquêtes qualitatives, en particulier, nécessitent des moyens humains et techniques importants : préparation matérielle, numérisation haute qualité, contrôle qualité, structuration des métadonnées, archivage pérenne. Un investisse-

ment accru permettrait de préserver des corpus encore inaccessibles et de soutenir la réutilisation de matériaux précieux pour la recherche contemporaine.

Ces réutilisations montrent que la valeur d'un jeu de données dépasse largement son objectif initial : une enquête électorale peut servir à analyser la confiance politique, les comportements générationnels ou les effets territoriaux, dès lors qu'elle est correctement documentée et accessible.

RENDRE LES DONNÉES RÉUTILISABLES, UN EFFORT COLLECTIF

La diffusion des données ne dépend pas uniquement de leur numérisation ou de leur structuration selon des standards. Elle suppose également le respect d'autres exigences : anonymisation, gestion des droits, consentement, protection des données sensibles, qui dépassent le cadre de cet article mais conditionnent la mise à disposition d'un jeu de données ou corpus. La mise en place de plateformes de confiance, impliquant une gestion transparente, est d'une importance capitale dans ce processus. L'expérience de Sciences Po montre que la réutilisation des données est un processus collectif, mobilisant ingénieurs en *data management*, experts en numérisation, bibliothécaires et chercheurs. Dans un projet récent, par exemple, la mise à disposition d'entretiens réalisés dans le cadre d'une enquête sur les trajectoires migratoires a nécessité la collaboration étroite entre juristes, ingénieurs données, archivistes et chercheurs pour garantir la conformité éthique et juridique du corpus. En articulant expertise technique, normes internationales et infrastructures pérennes, Sciences Po contribue à renforcer la qualité, la visibilité et l'impact de la recherche en sciences sociales.

OLESEA DUBOIS

responsable du Service Bibliothèque Numérique,
Sciences Po
olesea.dubois@sciencespo.fr

ALINA DANCIU

responsable de l'équipe Documentation -
Diffusion du CDSP, Sciences Po
alina.danciu@sciencespo.fr

SciencesPo

NumaHOP
Plate-forme de gestion de contenus numérisés

[9] NumaHOP, plateforme open source de gestion des chaînes de numérisation : <https://www.numahop.fr>

[10] Consulter : <https://bibnum.sciencespo.fr/s/catalogue/page/accueil?u=http://dossierspresse.sciences-po.fr/consult>