



Pratiques de gestion et de partage des données de recherche. Retour d'expérience de quatre projets SHS lauréats du Prix Science Ouverte 2025

Joachim Schöpfel

► To cite this version:

Joachim Schöpfel. Pratiques de gestion et de partage des données de recherche. Retour d'expérience de quatre projets SHS lauréats du Prix Science Ouverte 2025. 2026. <hal-05634994v2>

HAL Id: hal-05634994

<https://lilloa.hal.science/hal-05634994v2>

Preprint submitted on 29 May 2026

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.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

Pratiques de gestion et de partage des données de recherche

Retour d'expérience de quatre projets SHS lauréats du Prix Science Ouverte 2025

SCHÖPFEL Joachim

MCF émérite

Univ. Lille, ULR 4073 - GERiiCO - Groupe d'Études et de Recherche Interdisciplinaire en Information et Communication, F-59000 Lille, France

ORCID : <https://orcid.org/0000-0002-4000-807X>

Mail : joachim.schopfel@univ-lille.fr

Sciences de l'information et de la communication

Données de la recherche, science ouverte, plan de gestion de données, gestion de données de la recherche, partage de données, valorisation, réutilisation, principes FAIR, article de données

Research data, open science, data management plan, research data management, data sharing, outreach, reusage, FAIR principles, data paper

Cet article propose une étude comparative de quatre projets de recherche lauréats du Prix Science Ouverte 2025 issus des sciences humaines et sociales (humanités numériques, linguistique, sociologie quantitative et longitudinale), en se concentrant sur leurs pratiques de gestion des données de recherche. À partir d'entretiens qualitatifs, l'étude examine les types de données produites, les modalités de gestion, les responsabilités, ainsi que les stratégies de partage et de valorisation. L'analyse est structurée selon les principes FAIR (Findable, Accessible, Interoperable, Reusable). Les résultats mettent en évidence une forte hétérogénéité des pratiques, mais aussi une convergence vers une professionnalisation accrue de la gestion des données et une intégration progressive des exigences de la science ouverte. Une attention particulière est accordée à l'impact du Prix Science Ouverte des Données de la Recherche, dont les effets se manifestent principalement en termes de reconnaissance, de visibilité et de consolidation des pratiques.

This article presents a comparative study on four research projects awarded the Open Science Prize 2025 from social sciences and humanities (digital humanities, linguistics, quantitative and longitudinal sociology), focusing on their research data management practices. Based on qualitative interviews, the study examines the types of data produced, management approaches, responsibilities, as well as data sharing and outreach strategies. The analysis is structured around the FAIR principles (Findable, Accessible, Interoperable, Reusable). The results highlight a strong heterogeneity of practices, but also a convergence towards increased professionalization of data management and a gradual integration of open science requirements. Particular attention is paid to the impact of the Open Science Award for Research Data, whose effects appear mainly in terms of recognition, visibility, and consolidation of practices.

Introduction

La gestion des données de recherche constitue aujourd'hui un enjeu central des politiques de science ouverte. En France, la structuration et l'ouverture, selon des niveaux variables (ouvert, sur

demande, restreint), des données sont au cœur du deuxième axe des deux plans nationaux pour la science ouverte qui préconise, entre autres, des mesures pour adopter des politiques de données en conformité avec les principes FAIR, pour généraliser les plans de gestion des données dans les appels à projets de recherche, pour développer une culture de la donnée par des dispositifs d'(auto)formation, pour développer les articles de données (*data papers*) et pour reconnaître et amplifier la réutilisation des données partagées¹. Inscrit dans le deuxième Plan national pour la science ouverte 2021-2024, le Prix Science Ouverte des données de la recherche fait partie des mesures pour reconnaître et amplifier la réutilisation des données de la recherche, par la récompense des équipes et projets exemplaires en termes de préparation à la réutilisation ou de réutilisation des données de recherche. Son objectif : « mettre en lumière aussi bien des projets, des individus que des collectifs engagés dans des pratiques exemplaires de gestion et de partage des données de recherche, que des projets de recherche s'appuyant sur des données produites par d'autres »².

Cet article s'intéresse aux quatre projets issus des SHS récompensés en 2025. Son objectif est de présenter, sous forme d'un retour d'expérience, les pratiques de gestion et de partage des données de recherche par les équipes de ces projets. Il s'agit de rendre visible le travail des personnels scientifiques et de valoriser leurs résultats. Comment ont-ils organisé la collecte, le traitement et la publication de leurs données ? Comment ont-ils fait face aux contraintes techniques, juridiques et éthiques ? Quelles ressources ont-ils mobilisé ? Qu'ont-ils fait pour faciliter la réutilisation de leurs données ? Et puis, nous avons également voulu savoir ce que cette récompense leur a apporté. Mais pour commencer, quelques remarques sur l'environnement de cet article.

L'environnement

Dans le cadre de l'avènement de la science fondée sur les données, considérée comme quatrième paradigme (Hey et al., 2009) et « la mère de la science »³ (Igarashi et al., 2016), la gestion des données de la recherche est devenue une pratique globale et indispensable pour le progrès de la science, soutenue (et exigée) surtout par les agences de financement et les éditeurs de revues, mais confrontée au manque de ressources, de temps, d'infrastructures, d'interopérabilité et de reconnaissance (Chawinga & Zinn, 2019). Faire entrer la gestion des données dans les projets dès le montage est difficile, demande du temps, des moyens humains, techniques (équipements) et financiers, et implique la contribution et la coordination de plusieurs métiers (Cavallo & Noûs, 2019), y compris les professionnels de l'information, dans les laboratoires, les services de la documentation et les bibliothèques universitaires (Koltay, 2019).

Par ailleurs, les concepts même de la donnée et du partage de données varient selon les disciplines et les communautés scientifiques (Schöpfel et al., 2017 ; Thoegersen & Borlund, 2021). Cavallo & Noûs (2019) évoquent même des « disparités persistantes entre pays » (p.112). La diversité d'approches et de pratiques est la règle, pas l'exception.

De leur côté, les principes FAIR, formulés en 2016 (Wilkinson et al., 2016), proposent un cadre normatif visant à améliorer la découvrabilité, l'accessibilité, l'interopérabilité et la réutilisabilité des données. Ils sont ainsi en passe de devenir « des standards de facto, avec une cascade de normes et de standards mobilisés » (Schöpfel, 2018). Toutefois, le constat est le même : malgré leur caractère normatif et convergent, la mise en œuvre des principes FAIR varie selon les disciplines, les types de données et les contextes institutionnels (Wang et al., 2024 ; cf. aussi Wilkinson et al., 2018)⁴.

Très justement, Rebouillat (2019) fait remarquer que la donnée est dépendante du contexte dans lequel elle est utilisée, rejoignant Borgman (2015) qui quelques années auparavant a fait le constat

¹ Cf. Ouvrir la Science <https://www.ouvrirlascience.fr/plan-national-pour-la-science-ouverte/>

² Cf. <https://www.ouvrirlascience.fr/les-candidatures-pour-le-prix-science-ouverte-des-donnees-de-la-recherche-2025-sont-ouvertes/>

³ En anglais: data-driven science is *mother of science*, namely, a scientific framework that drives many fields of science.

⁴ Cf. aussi les informations sur le site de l'initiative GO FAIR <https://www.GO-FAIR.org>

qu'une donnée ne naît pas donnée mais le devient quand on l'utilise (collecte, analyse, interprétation...) comme preuve dans un but de recherche. Sur le terrain, la diversité des contextes de recherche et de leurs incidences sur la gestion des données impose l'élaboration d'une politique publique intégrant les spécificités disciplinaires, techniques et socio-économiques ainsi que les besoins propres aux différents champs scientifiques. Dans les sciences expérimentales et appliquées, caractérisées par des coûts élevés de collecte et d'analyse, elle requiert en outre des financements supplémentaires dédiés à la gestion des données (Rebouillat, 2019).

Un nombre toujours croissant d'études et d'enquêtes⁵ mettent en évidence un paysage dynamique, caractérisé par des pratiques, méthodes et approches extrêmement variées, en fonction des disciplines et thématiques, des équipements et outils, des environnements institutionnels et des communautés, des partenariats industriels, des ressources humaines, techniques et financières disponibles mais aussi, tout simplement, en fonction des données elles-mêmes. Les SHS ne font pas exception. Plusieurs enquêtes et études ont révélé une « extraordinaire hétérogénéité » des pratiques et des besoins (Malingre et al., 2019 ; cf. aussi Schöpfel & Prost, 2016), y compris par rapport à l'environnement et aux conditions de travail sur le terrain des laboratoires (Bendjaballah & Garcia, 2024), mais également des différences importantes avec les domaines en sciences, technologie et médecine, dans les pratiques aussi bien que dans les représentations (Amiel et al., 2020). Comparé à ces études antérieures, ce retour d'expérience, loin de vouloir être exhaustif ou représentatif, constitue une sorte de zoom sur les pratiques sur le terrain d'un projet de recherche au quotidien.

A propos de la méthode de ce retour d'expérience

Pour réaliser ce retour d'expérience, nous avons contacté par mail les porteurs des quatre projets SHS lauréats du Prix Science ouverte des données de la recherche 2025⁶ (tableau 1) pour leur expliquer notre démarche et pour les solliciter pour un entretien à distance d'environ 30 minutes.

Projet	Domaine	Responsable	Catégorie du prix
<i>e-NdP</i>	Histoire	Julie Claustre	Créer les conditions de la réutilisation
<i>Panel de Caen</i>	Sociologie	Claire Bidart	Mention spéciale du jury
<i>Subwork</i>	Urbanisme	Nicolas Rimbault	Créer un jeu de données manquantes
<i>Les Vocaux</i>	Linguistique	Julie Glikman	Créer un jeu de données manquantes

Tableau 1. Les quatre projets de l'étude

⁵ Cf. notamment le recensement exemplaire des enquêtes en France depuis 2015 par le groupe de travail science ouverte du consortium COUPERIN <https://gtso.couperin.org/gtdonnees/recensement-enquetes-sur-les-donnees-de-la-recherche/>

⁶ En 2025, la quatrième édition du Prix de la Science ouverte a récompensé 24 lauréats, dont huit dans le domaine des données de la recherche. De ces huit lauréats, les quatre qui sont issus des SHS constituent notre échantillon.

Toutes les personnes contactées ont confirmé leur intérêt et leur disponibilité. Les entretiens dont la trame a été communiquée au préalable par mail aux interlocuteurs ont eu lieu entre le 16 février et le 24 mars 2026, avec les porteurs et, dans un cas, aussi avec deux autres membres de l'équipe projet. Les entretiens ont été conduits en mode semi-directif en visioconférence et enregistrés sur Zoom. Ils ont duré entre 30 et 45 minutes.

Les enregistrements ont fait l'objet d'une transcription assistée par l'IA. Les transcriptions ont été vérifiées, corrigées, structurées et synthétisées par l'auteur. Les transcriptions synthétiques – de trois à quatre pages – ont été communiquées aux interlocuteurs pour relecture et validation. Les enregistrements ont été supprimés après validation des transcriptions par les interlocuteurs.

L'analyse du contenu de ces entretiens a été réalisée manuellement, par l'auteur. L'objectif était triple : essayer de dégager les aspects caractéristiques de chaque projet par rapport aux thématiques de l'entretien (type de données, organisation de la gestion etc.) ; identifier des cas particuliers ; et tenter une comparaison entre les quatre projets, notamment par rapport aux principes FAIR. Certains résultats ont été présentés sous forme de tableau, pour une meilleure compréhension. La discussion met l'accent sur plusieurs thèmes essentiels, comme la tension entre l'ouverture et la protection des données, l'adhésion aux valeurs de la science ouverte, l'importance des moyens mis à disposition, et le rôle central du travail en équipe.

Dans le texte, les verbatims issus des transcriptions validées sont cités en italique.

Diversité des données et des chaînes de production

Les quatre projets étudiés illustrent une grande diversité de types de données : corpus textuels numérisés, données audio, bases statistiques massives et données longitudinales mixtes. Tous reposent sur des processus de transformation importants, produisant des données dérivées à forte valeur ajoutée.

Au cœur du projet e-NpD se trouve un corpus exceptionnel mais jusqu'alors difficilement exploitable : les procès-verbaux des réunions des chanoines de Notre-Dame de Paris, conservés de 1326 jusqu'à la Révolution française [...] Malgré leur importance pour l'histoire institutionnelle, religieuse et sociale de la cathédrale - et plus largement pour l'histoire de Paris et du royaume - ces documents n'étaient ni indexés ni transcrits de manière systématique, ce qui limitait fortement leur accessibilité.

Sous l'égide des Archives Nationales, 26 registres manuscrits de 1326 à 1504, issus des archives de Notre-Dame (données primaires), ont été numérisés puis transcrits automatiquement pour produire des fichiers texte en format XML ALTO 4.3 (données produites) et construire un corpus interrogeable d'environ 4 600 pages et de cinq millions de mots. La transcription par reconnaissance d'écriture manuscrite (Handwritten Text Recognition, HTR) a impliqué la mise en place d'une chaîne de transformation forte : *Nous avons formé un groupe d'une quinzaine de chercheurs et de conservateurs à l'entraînement des algorithmes de transcription automatique. Tout cela a été fait en une année, et c'était assez efficace.*

La qualité de cette transformation automatique du patrimoine en corpus numérique est imparfaite et varie selon les mains et les périodes, avec un taux de fiabilité compris entre 88 % et 94 %. *Consciente de ces limites, l'équipe a fait le choix d'une diffusion intégrale et transparente : les images originales sont systématiquement présentées en regard des transcriptions, et les marges d'erreur sont explicitement signalées.* L'équipe projet assume le caractère perfectible du résultat, privilégiant une logique d'ouverture rapide des données *[qui] s'inscrit pleinement dans les principes de la science ouverte : accès libre en ligne, documentation des méthodes employées, partage des données et des outils, et possibilité pour la communauté scientifique de réutiliser, corriger et enrichir progressivement le corpus.* En conséquence, le projet e-NdP a également développé des outils de recherche adaptés, *capables de prendre en compte les variantes graphiques et les erreurs probables, et facilitent ainsi l'exploration thématique, prosopographique ou lexicale du corpus.*

Le *Panel de Caen*, quant à lui, a utilisé des données mixtes, quantitatives et qualitatives (calendriers de vie, entretiens, réseaux sociaux) dans une approche longitudinale en six vagues sur vingt ans,

afin de produire un corpus complexe, riche et diversifié à plusieurs niveaux avec des bases nettoyées, des codages et des indicateurs (données dérivées). L'enjeu majeur de ce projet d'envergure était la combinaison de *plusieurs types d'informations complémentaires* [...]

- *Des données quantitatives structurées, recueillies à chaque vague d'enquête, telles que des calendriers de vie retraçant les parcours scolaires, professionnels, résidentiels et familiaux des enquêtés, ainsi que des listes de relations permettant d'identifier les personnes importantes de leur entourage [...]*
- *Des matériaux qualitatifs approfondis, issus d'entretiens longs dans lesquels les participants développent des récits détaillés de leur trajectoire et donnent du sens à leurs expériences et à leurs relations [et]*
- *Des données relationnelles, au cœur de la démarche, qui permettent de cartographier les réseaux personnels des individus et d'en suivre les transformations au fil du temps.*

Ancré en sciences sociales et dans le domaine de l'urbanisme, le projet *Subwork* a exploité comme données primaires des bases administratives et statistiques, notamment de l'INSEE, afin de constituer (reconstruire) une base de données nationale *couvrant l'ensemble des actifs en France, soit environ 26 millions de personnes*, des jeux de données anonymisés et diffusables (95 fichiers) et des indicateurs spatialisés à une échelle infra-communale. L'intégration de ces données de sources différentes, leur harmonisation et l'anonymisation a nécessité une forte ingénierie de données : *Les données les plus détaillées, notamment celles du recensement de la population et de la base Tous Salariés⁷, ont été exploitées dans l'environnement sécurisé du Centre d'accès sécurisé aux données (CASD)⁸, afin de reconstruire des nomenclatures adaptées aux besoins du projet et de localiser précisément les emplois et les établissements. Ces matériaux ont été complétés par des données pseudonymisées, telles que celles de l'Enquête Emploi⁹, ainsi que par des ressources en open data, incluant les référentiels géographiques et administratifs nécessaires au traitement spatial¹⁰ [...]* Le traitement des données ne s'est pas limité à une exploitation descriptive ; il a impliqué une véritable construction d'outil statistique, combinant ingénierie de données, requalification conceptuelle des catégories et production d'indicateurs spatialisés adaptés à l'analyse de la ville productive.

Le projet de recherche en linguistique *Les Vocaux* a exploité des fichiers audio natifs de SMS dans de formats différents selon les plateformes (OGG, WebM...) et, parfois, des fichiers audio (piste sonore) extraits de vidéos : *Dans certains cas particuliers, lorsque les fichiers ne pouvaient être exportés directement, des participants ont réalisé des captures vidéo de la lecture du message vocal, dont la piste sonore a ensuite été extraite.* Le projet a également collecté *un nombre limité de métadonnées [...]* dans le respect d'un principe de minimisation des données personnelles. A partir de ces données primaires, l'équipe projet a produit des transcriptions en format adapté pour l'analyse (TXM, XML). Un enjeu majeur a été la collecte et le traitement des données personnelles sensibles (voix), avec des fortes contraintes éthiques.

Tous les projets produisent des données dérivées à forte valeur ajoutée, avec une diversité croissante des formats et des structures. La transformation et l'enrichissement des données est au cœur de chaque projet. Cette dimension transformationnelle apparaît comme un trait commun majeur : les données ne sont pas simplement collectées, mais reconstruites, enrichies et structurées pour répondre à des problématiques scientifiques spécifiques.

⁷ Base Tous salariés de l'INSEE <https://www.insee.fr/fr/metadonnees/source/serie/s1998>

⁸ Source : Centre d'accès sécurisé aux données (CASD) <https://www.casd.eu/>

⁹ La série des enquêtes Emploi sur PROGEDO <https://www.progedo.fr/la-serie-des-enquetes-emploi/>

¹⁰ IRIS-INSEE <https://www.insee.fr/fr/information/7708995>

Formalisation et organisation de la gestion des données

La présence d'un plan de gestion des données (PGD) varie fortement. Seul le projet *e-NdP* financé par l'ANR¹¹ dispose d'un plan formel : *Comme c'est un projet ANR, il a fallu rédiger un plan de gestion en trois versions, avant le début du financement, à mi-chemin, et à la fin. Le plan a été rédigé par la responsable du projet avec un ingénieur CNRS en informatique d'un des laboratoires de recherche (LaMOP). Le plan a été déposé sur HAL (Brochard & Claustre, 2021). Les autres projets reposent sur des dispositifs implicites ou élaborés au préalable.*

Ceci peut être lié au caractère même des données et de leur traitement. Pour le projet *Subwork*, par exemple, *si aucun document formel intitulé « plan de gestion des données » n'a été rédigé, il est évident que ses composantes aient existé de facto, notamment à travers les dossiers d'accès aux données sécurisées, les conventions signées avec les producteurs de données et les protocoles d'anonymisation mis en place pour la diffusion de la base finale. Autrement dit, un ensemble de pratiques et de procédures équivalentes ont été déployées, constituant un plan de gestion des données implicite.*

Projet	Plan de gestion	Responsabilité
<i>e-NdP</i>	Formel (trois versions)	Distribuée
<i>Panel de Caen</i>	Non	Centralisée
<i>Subwork</i>	Implicite (de facto)	Distribuée
<i>Les Vocaux</i>	Non	Centralisée (données primaires) et distribuée (données anonymisées)

Tableau 2. Plan de gestion et responsabilité

Dans deux projets, la responsabilité de la gestion des données a été centralisée, auprès de la responsable du projet pour des données non anonymisées ou auprès d'un ingénieur pour l'anonymisation. Cependant, la tendance est plutôt l'inverse, c'est-à-dire, la responsabilité de la gestion des données évolue vers des modèles distribués, avec par exemple un(e) responsable par partenaire institutionnel (*e-NdP*) ou par traitement (*Subwork*), impliquant une pluralité d'acteurs : chercheurs, ingénieurs, institutions et infrastructures.

Les deux modèles – une responsabilité centralisée et distribuée – peuvent cohabiter au sein du même projet ; ainsi, dans le projet *Les Vocaux*, *pour les données non anonymisées, la directrice du projet était l'unique responsable pour la réception, la modération, l'anonymisation et la conversion des données ; tandis que pour les fichiers anonymisés, la responsabilité des données a été répartie entre les membres du projet, en fonction du traitement et de l'analyse (transcription collaborative etc.).*

Dans le projet *Subwork*, *la responsabilité des données n'a pas reposée sur une seule personne, mais sur un dispositif articulant responsabilité scientifique, responsabilité opérationnelle des utilisateurs habilités et responsabilité institutionnelle de l'établissement porteur, le tout encadré par les règles fixées par les producteurs de données et les infrastructures sécurisées d'accès.*

Autrement dit, la responsabilité de la gestion est intimement liée à l'expertise des acteurs et leur apport spécifique au projet. Dans le projet *Subwork*, par exemple, *une ingénieure de recherche de PROGEDO-Loire était en charge de la FAIRisation de l'ensemble des données, tandis qu'un enseignant-chercheur de l'Université de Montréal était responsable de la coordination autour des outils, en particulier ceux proposés par l'infrastructure numérique de recherche Huma-Num (GitLab, Serveur R, application Shiny...).* Le rôle des ingénieurs apparaît central dans tous les projets, en particulier pour le traitement, l'anonymisation et la FAIRisation des données, ce qui implique parfois aussi le développement d'outils, marquant une professionnalisation croissante de la gestion

¹¹ Depuis 20189, l'ANR exige un PGD pour tous les projets sélectionnés comme préalable du financement.

des données. Citation (e-NdP) : *ce qui a été indispensable, ça a été les ingénieurs ou assistants ingénieurs en informatique des divers laboratoires.*

Les équipes ont rarement fait appel à un soutien ou accompagnement extérieur au projet pour la gestion des données ; le projet e-NdP est l'exception qui confirme la règle : ici, il n'y a pas eu de formation particulière à la gestion des données ou de soutien de la part d'un service commun de documentation (SCD) ou d'un autre service mais *comme les données étaient à plusieurs endroits, à un moment, il a fallu qu'on fasse faire par un consultant externe (un archiviste-paléographe expert en valorisation numérique patrimoniale) une sorte de géographie, de carte de nos données numériques. Parce qu'on n'y voyait plus clair, nous-mêmes quoi, et ça a été très efficace.* Mais la règle était plutôt d'internaliser l'expertise et les compétences nécessaires, *notamment via l'intégration d'une ingénieure de données dans l'équipe projet. En revanche, la question des ressources et moyens disponibles – outils infrastructures dans le cloud, acquisition des données, coût des traitements etc. – était cruciale tout au long du projet, comme dans le projet Subwork.*

En résumé, un PGD formel existe dans l'unique projet récent financé par l'ANR, tandis que les autres reposent sur des pratiques implicites. La gestion évolue vers une responsabilité collective et structurée, avec une forte professionnalisation. Et les contraintes juridiques peuvent structurer la gestion même sans PGD ; on y reviendra tout de suite.

Contraintes juridiques et éthiques

Les contraintes juridiques, notamment liées au Règlement Général sur la Protection des Données (RGPD), jouent un rôle structurant dans les projets impliquant des données personnelles. Elles influencent non seulement les modalités de collecte, mais aussi les conditions de stockage, de traitement et de diffusion.

La protection des données à caractère personnel, avec la question du consentement explicite, est cruciale dans le projet *Les Vocaux* qui a mobilisé un accompagnement institutionnel important : *La déléguée à la protection des données (DPO) a joué un rôle central dans la rédaction des formulaires d'information et de consentement, ainsi que dans la définition des modalités de stockage et de circulation des données. Une attention particulière a été portée à la question du droit à la voix, dès lors que le corpus conserve des enregistrements vocaux identifiables [...] la voix constitue en elle-même une donnée personnelle. Le protocole de recherche a ainsi été élaboré en parallèle d'une réflexion approfondie sur la protection des participants, et non a posteriori.* D'après la responsable du projet, l'anticipation a eu des conséquences directes sur le calendrier du projet puisque *la phase de collecte [a] dû être retardée afin de finaliser les documents juridiques nécessaires.* Pour elle, l'expérience souligne qu'il faut intégrer cet aspect très en amont d'un projet, *notamment pour les jeunes chercheurs soumis à des contraintes temporelles fortes.*

La protection des données personnelles joue un rôle important aussi dans le projet *Subwork* ; ce projet a mobilisé des données pseudonymisées en complément, pour affiner certaines analyses, notamment en matière de conditions d'emploi et de caractéristiques individuelles *dans une logique de triangulation, afin de consolider l'interprétation des résultats obtenus à partir des bases exhaustives.* Pour produire une base de données diffusable, il a fallu supprimer des informations potentiellement identifiables et mettre en place des règles garantissant la confidentialité statistique : *L'enjeu principal consistait à rendre diffusable un matériau initialement construit à partir de données confidentielles. Les traitements réalisés en environnement sécurisé ont permis de produire des fichiers dérivés, agrégés et anonymisés, ne comportant plus d'informations directement identifiantes. Ce travail a impliqué la suppression ou la transformation des variables sensibles, l'application de seuils de diffusion pour éviter les faibles effectifs et la vérification du risque de ré-identification. L'objectif était de constituer une base autonome, exploitable hors des dispositifs sécurisés, tout en respectant strictement les règles fixées par les producteurs de données.*

Le projet plus ancien du *Panel de Caen* montre une adaptation a posteriori à ces exigences, révélant une évolution rapide du cadre normatif de la recherche. Ainsi, ce projet (et plus particulièrement un ingénieur de recherche du CNRS avec son équipe) a dû réaliser *un important travail d'anonymisation [...] afin de rendre les données partageables tout en garantissant la protection des personnes [...] Cette tâche était particulièrement complexe, car les entretiens contiennent de*

nombreux noms de personnes et de lieux, ce qui est accentué par la nature même de l'étude portant sur les réseaux personnels. Les données mises à disposition ont donc été soigneusement anonymisées, ce qui constitue aujourd'hui la principale garantie de protection.

Ouverture et partage des données

Tous les projets s'inscrivent dans une démarche de science ouverte, mais selon des modalités différenciées. Si les données patrimoniales du projet e-NdP peuvent être diffusées librement, les données sensibles des autres projets nécessitent des dispositifs d'accès contrôlé. L'ouverture des données repose sur plusieurs leviers (ou bonnes pratiques) - dépôts dans des entrepôts, documentation détaillée, multiplicité des formats et développement d'interfaces facilitant leur (ré)utilisation.

Les images et textes du projet e-NdP sont librement accessibles sur le site de l'École des Chartes (et) les données produites ont été déposées sur Zenodo.

Les jeux de données de *Subwork* sont accessibles sur NAKALA, sous forme de 95 fichiers, dont une documentation des fichiers de données déposés de 44 pages. Par ailleurs, la structuration des fichiers et la production d'une documentation associée - dictionnaires de variables, description des traitements, explicitation des périmètres analytiques - ont constitué un levier essentiel de valorisation. La documentation est essentielle : En rendant explicites les opérations de construction des données, le projet a renforcé leur réutilisabilité par d'autres équipes de recherche ou par des acteurs institutionnels, comme des agences d'urbanisme.

Le corpus du projet *Les Vocaux* a été diffusé en libre accès sur la plateforme disciplinaire ORTO-LANG, accompagné d'une documentation détaillée. Il a également été mis en ligne sur un site de consultation hébergé par l'université de Liège pour faciliter l'appropriation par des utilisateurs ne souhaitant pas manipuler les fichiers sources localement. Une particularité de cette diffusion est que le corpus est mis en ligne dans plusieurs formats afin de faciliter la réutilisation : fichiers audio bruts, transcriptions, versions exploitables par des logiciels d'analyse textométrique tels que TXM, et, à terme, versions enrichies comportant des annotations et un alignement au signal au format XML.

Pour ces projets, l'ouverture est une norme, mais modulée par la sensibilité des données et les contraintes juridiques. A titre d'exemple, les données du *Panel de Caen* ont été déposées dans l'entrepôt de données data.sciencespo de Sciences Po Paris où elles sont diffusées avec une licence Creative Commons mais en accès restreint. Il faut donc demander explicitement l'accès et s'engager à respecter les conditions générales d'utilisation de cet entrepôt : une manière de faciliter la réutilisation des données tout en protégeant les données sensibles du projet.

Analyse selon les principes FAIR

L'analyse FAIR met en évidence des niveaux globalement élevés de conformité, bien que différenciés selon les dimensions. Si la trouvabilité ou découvrabilité (findability) des données est bien assurée dans les projets récents grâce aux entrepôts de données, l'accessibilité est fortement contrainte par la sensibilité des données ; l'interopérabilité des données dépend du degré de structuration et de standardisation, et leur réutilisabilité repose avant tout sur la qualité de la documentation. Voici une tentative de synthèse (tableau 3).

Projet	F	A	I	R
e-NdP	Fort	Très ouvert	Fort	Fort
Panel de Caen	Moyen	Restreint	Moyen	Très fort
Subwork	Fort	Contrôlé	Très fort	Très fort

Les Vocaux	Fort	Ouvert mais encadré	Fort	Fort
------------	------	---------------------	------	------

Tableau 3. La conformité des projets avec les principes FAIR

L'appréciation de la conformité FAIR présentée dans ce tableau (très fort, fort, moyen...) reflète l'interprétation par l'auteur de l'article des différences entre les projets. Il ne s'agit en aucun cas d'une évaluation objective ou standard à partir d'une échelle ou d'un test. Quelques commentaires :

- F : Tous les projets récents assurent une bonne découvrabilité, par une diffusion sur une ou plusieurs plateformes ; le *Panel de Caen* est en transition d'un partage plus informel vers une infrastructure récente mais plus visible.
- A : L'accessibilité des données dépend directement de la sensibilité des données, avec une forte tension entre ouverture et protection ; l'ouverture des données personnelles nécessite un travail important en amont et, bien sûr, depuis le RGPD, le consentement des personnes concernées.
- I : Le projet *Subwork* peut servir de modèle pour une très forte interopérabilité de données, à cause d'une structuration avancée et des référentiels INSEE ; l'interopérabilité est forte aussi dans les projets *e-NdP* et *Les Vocaux*, avec des formats standards et des outils de recherche appropriés tandis que la complexité structurelle des données du *Panel de Caen* limite la standardisation et l'interopérabilité de ses résultats.
- R : Sans doute le principe FAIR le plus caractéristique de ces projets et le haut degré de réutilisabilité de leurs données, avec des métadonnées de qualité, une documentation détaillée, un travail important d'anonymisation des données sensibles, des données structurées (*Subwork*) et des formats multiples (*Les Vocaux*). La réutilisation réelle (déjà nombreuse pour certains projets) témoigne de cette conformité avec le principe R.

Ajoutons qu'au vu de ces constats et par rapport à la gestion des données, les quatre projets apparaissent globalement conformes aux principes FAIR, mais selon des logiques distinctes (tableau 4).

Projet	Logique	Focus	Commentaire
<i>e-NdP</i>	S'inscrivant dans une logique patrimoniale	<i>heritage-driven</i>	Très bon équilibre, orienté open science
<i>Panel de Caen</i>	S'inscrivant dans une logique d'héritage historique	<i>legacy-driven</i>	FAIR élevé mais construit a posteriori
<i>Subwork</i>	Reposant sur une approche fondée sur les données	<i>data-driven</i>	Modèle FAIR avancé (ingénierie des données)
<i>Les Vocaux</i>	Relevant d'une démarche à dominante éthique	<i>ethics-driven</i>	FAIR élevé mais contraint par l'éthique

Tableau 4. Logique de la mise en conformité avec les principes FAIR

Si le projet *Subwork* reposant sur une forte ingénierie des données apparaît comme le plus avancé sur le plan FAIR, les autres illustrent des formes adaptées à leurs contraintes spécifiques.

Réutilisation et valorisation

Les données produites font l'objet de réutilisations variables, allant d'usages intensifs et interdisciplinaires à des appropriations émergentes, telles que la détection de thématiques dans un corpus linguistique. La documentation joue un rôle déterminant dans ces processus.

La valorisation des données issues de la recherche ne suit pas nécessairement une stratégie précise. Si pour *Subwork*, la valorisation s'est appuyée sur plusieurs canaux complémentaires [...] des publications scientifiques, des communications en colloques et des rapports de recherche, ailleurs, comme dans le *Panel de Caen*, les échanges de données reposaient (au départ) sur des accords informels entre collègues.

Les articles de données ou *data papers*, encore peu développés, apparaissent comme un outil prometteur mais contraint par des enjeux de temps et de reconnaissance académique. Pour l'équipe du *Panel de Caen*, l'idée de la rédaction d'un *data paper* est envisagée depuis longtemps, avec les ingénieurs qui ont travaillé à la mise à disposition du projet [...] mais elle n'a pas encore été concrétisée faute de temps. Seul l'équipe du projet *Subwork* a rédigé un tel document : La base de données elle-même, en tant qu'infrastructure méthodologique, a pu être présentée comme un apport original, notamment par la description détaillée des choix de nomenclature et des méthodes de spatialisation [et] elle a fait l'objet d'un *data paper* en cours de soumission [...] complétée d'un site web d'exploration des données, sorte de "vitrine" de la base¹² (tableau 5).

Projet	Data paper	Réutilisation
e-NdP	Pas de <i>data paper</i>	Transdisciplinaire
Panel de Caen	Envisagé	Très forte, multidisciplinaire
Subwork	En cours	Potentielle
Les Vocaux	En projet	En développement

Tableau 5. Rédaction d'un *data paper* et réutilisation des données

Pourtant, dans tous les projets la documentation des données est présente et parfois très développée, ce qui faciliterait la rédaction d'un *data paper* ; selon les mots de la responsable du projet *Les Vocaux* : *Un document de documentation détaillé accompagne le dépôt et constitue une première étape vers la rédaction d'un article de données plus formalisé.* Mais c'est peut-être aussi l'une des raisons pourquoi un *data paper* n'est pas nécessairement une priorité – puisqu'il y a la documentation, et puisque les données sont déjà (ré)utilisées. Est-ce que les chercheurs considèrent les *data papers* comme un doublon de la documentation ou comme un objet de reconnaissance encore un peu flou ? La question reste ouverte.

Une autre raison est la priorité donnée à la rédaction d'un article de recherche et/ou de méthodologie. Ainsi, l'équipe du *Panel de Caen* a déjà et d'abord publié un article qui traite de ce processus de collecte et de mise à disposition des données mais sans entrer dans les détails de leur documentation (Bidart et al., 2025). Pourtant, un *data paper* pourrait avoir un sens en complémentarité dans la mesure où il offrirait davantage de précisions pour l'accompagnement des futurs utilisateurs, car les données produites sont complexes, riches et organisées à différents niveaux (données brutes, résultats intermédiaires, analyses). Une situation similaire dans le projet e-NdP où, en parallèle avec d'autres articles de recherche le coordinateur du programme HTR et le responsable de la mission numérique de l'Ecole des Chartes ont publié plusieurs articles sur la méthodologie et le processus de la production des données, sans qu'on puisse parler de *data papers* au sens strict (cf. par exemple Torres Aguilar & Jolivet, 2023).

¹² Les données Subwork https://tlecorre.gitpages.huma-num.fr/Subwork_website/

Les équipes suivent l'usage de leurs données avec intérêt. Ainsi, l'équipe du *Panel de Caen* fait remarquer que les données du Panel ont été déjà réutilisées et exploitées plusieurs dizaines de fois, majoritairement en sociologie (pour des travaux sur des thématiques variées comme la formation du couple, les mobilités sociales ou le rapport à l'école). Néanmoins, mesurer la réutilisation réelle des données n'est pas toujours facile, faute d'indicateurs et/ou d'information sur les utilisateurs et les utilisations. A titre d'exemple, l'observation de l'équipe du projet *Subwork* : *À partir des éléments disponibles, on ne dispose pas d'indications explicites et documentées sur des réutilisations effectives de ces données, identifiées et tracées, en dehors du cadre du projet.* Pareil pour *Les Vocaux* - il est difficile de mesurer l'ampleur exacte de la réutilisation dans un contexte de diffusion libre. Trois constats s'imposent :

1. Tous les projets sont conscients de l'intérêt potentiel de leurs jeux de données. Un exemple : *Plusieurs caractéristiques du projet Subwork permettent d'inférer un potentiel élevé de réutilisation scientifique et institutionnelle. Tout d'abord, la production d'une base anonymisée diffusable, structurée en fichiers cohérents et accompagnée d'une documentation méthodologique, constitue en soi une condition essentielle de réutilisabilité. Ensuite, le caractère national et exhaustif de la base - couvrant l'ensemble des actifs et reposant sur une maille infra-communale - en fait un outil susceptible d'être mobilisé par différents types d'acteurs [...] La compatibilité avec les référentiels géographiques officiels renforce également cette possibilité d'appropriation.*
2. La réutilisation se développe dans un écosystème dynamique, constitué de plateformes, d'outils d'analyse, d'événements et de publications. Deux exemples : *Les données du Panel de Caen ont déjà été réutilisées à de nombreuses reprises [...] La mise à disposition récente dans des infrastructures dédiées permet désormais ce partage dans un cadre plus sécurisé. De son côté, la réutilisation du corpus du projet Les Vocaux commence à se développer [...] Certains collègues ont indiqué avoir testé le corpus avec leurs propres outils d'analyse de la parole. Un numéro de revue est en préparation à la suite d'une journée d'étude consacrée à l'exploitation scientifique du corpus, et des étudiants de master et doctorants ont également travaillé sur ces données. D'autres plateformes, comme des environnements de textométrie¹³, ont intégré le corpus dans leurs interfaces, parfois sans information systématique auprès de l'équipe initiale, ce qui témoigne néanmoins de son appropriation par la communauté.*
3. La réutilisation réelle ou potentielle dépasse largement les limites disciplinaires. Les données du projet d'urbanisme *Subwork* pourraient intéresser des chercheurs en sociologie, en géographie ou en économie, agences d'urbanisme [et] services d'études des collectivités territoriales. Les données du *Panel de Caen* ont suscité un intérêt dans d'autres disciplines [...] la linguistique (par exemple pour étudier l'accent normand), la démographie (effets des événements de vie sur les réseaux), la géographie (spatialisation des relations), la physique et l'informatique (dynamique des réseaux), l'anthropologie ou encore les sciences de l'éducation. A titre d'anecdote, la responsable du projet e-NdP raconte que récemment, une jeune chercheuse en linguistique a déjà demandé cet ensemble de données afin de faire des analyses informatiques de détection de thèmes dans un corpus latin.

Ajoutons ici aussi l'intérêt des données pour développer de nouveaux partenariats et monter de nouveaux projets. Deux exemples peuvent être proposés : Des chercheurs étrangers ont reproduit le protocole d'enquête du *Panel de Caen* au Québec et en Argentine, donnant lieu à des comparaisons internationales. Et la responsable du projet e-NdP envisage sur la base de cette première expérience de construction de corpus [...] de monter un projet européen (ERC) afin de comparer comment on gérait la cathédrale parisienne à la manière dont on gérait d'autres cathédrales, à partir de la documentation dans les cathédrales qui est absolument énorme – mettre à disposition des corpus similaires serait extrêmement intéressant pour les historiens.

La qualité de la documentation, l'accessibilité des données et leur standardisation semblent être les facteurs clés de la réutilisation. Mais il ne faut pas penser la situation figée ou statique, au contraire ; il faut penser la réutilisation comme un processus dans un environnement scientifique,

¹³ Dont le Lexicoscope de l'Université Grenoble Alpes http://phraseotext.univ-grenoble-alpes.fr/lexicoscope_2.0/

technologique et socioéconomique dynamique, comme une sorte d'investissement dans le temps, soutenue et favorisée par la politique de la science ouverte. Selon les mots de la responsable du *Panel de Caen* : *Malgré ces usages déjà nombreux, le potentiel des données semble encore largement sous-exploité, ce qui renforce l'importance de leur diffusion. Ces données sont maintenant déposées dans l'entrepôt de données de Sciences Po Paris, et sont donc à la fois plus accessibles et mieux protégées.*

Impact du Prix Science ouverte des données de la recherche

Les quatre projets étudiés ont été distingués par le Prix Science ouverte des données de la recherche, une récompense nationale visant à valoriser les pratiques exemplaires de gestion et de partage des données. L'analyse des entretiens permet d'identifier plusieurs effets convergents de cette distinction.

Tout d'abord, le prix joue un rôle majeur en termes de reconnaissance scientifique et institutionnelle. Il agit comme un label de qualité, validant des pratiques souvent exigeantes et encore inégalement valorisées (car invisibles) dans les carrières académiques, en particulier le travail de structuration, de documentation et d'archivage des données. La responsable du *Panel de Caen* confirme ainsi que *cette reconnaissance lui a apporté une forme de réconfort, en validant le travail considérable accompli, notamment les tâches longues et peu visibles de préparation, de nettoyage et d'archivage des données.* Pour elle, ce prix ne représente pas une forme de gloire personnelle, mais plutôt la reconnaissance d'un effort collectif par la valorisation du travail d'équipe accompli.

Ensuite, il constitue un levier important de visibilité. Dans plusieurs cas (*e-NdP*, *Subwork*, *Les Vocaux*), l'obtention du prix a permis d'accroître la diffusion des projets au sein de la communauté scientifique et au-delà, favorisant ainsi leur appropriation et leur potentiel de réutilisation.

Le prix contribue également à la consolidation et à la poursuite des projets, un effet mis en avant par l'équipe *Subwork* qui souligne l'intérêt du prix *en particulier au regard de son investissement méthodologique dans la production et la diffusion de données.* Le prix peut soutenir le développement de nouvelles fonctionnalités techniques (comme la bibliothèque numérique de l'École des Chartes dans le cadre du projet *e-NdP*), relancer certaines activités (comme la collecte de données du projet *Les Vocaux*) ou renforcer les dispositifs juridiques et organisationnels existants, en collaboration avec les services compétents (*Les Vocaux*). Il agit ainsi comme un accélérateur de dynamiques déjà engagées plutôt que comme un déclencheur.

Quatrième effet du prix, et pas le moindre : *Sur le plan symbolique, un tel prix confère une légitimité institutionnelle forte à la démarche adoptée [...] la distinction a renforcé la crédibilité de l'équipe auprès des financeurs et des partenaires publics* (projet *Subwork*).

Bien que le Prix Science ouverte puisse être associé à un soutien financier, cet aspect apparaît secondaire au regard de ses effets en termes de reconnaissance, de valorisation et de structuration. Dans l'ensemble, le prix agit moins comme un instrument de transformation radicale que comme un levier de légitimation, de visibilité et d'amplification de pratiques déjà alignées avec les principes de la science ouverte.

Discussion

Vu le nombre limité d'entretiens, notre analyse n'a pas d'ambition d'être exhaustive ou représentative. Néanmoins, illustrée de nombreux verbatims, elle permet de faire plusieurs observations pour mieux comprendre les pratiques de la gestion des données de la recherche sur le terrain des SHS.

D'une manière générale, l'analyse met en évidence une tension structurante entre ouverture et protection, particulièrement marquée pour les données personnelles. Elle souligne également une évolution vers une intégration croissante des principes FAIR, même en l'absence de formalisation explicite. Les pratiques observées montrent que la mise en œuvre des principes FAIR est moins une question de conformité formelle qu'un processus situé, dépendant des contextes disciplinaires,

des ressources disponibles, de la politique de données des établissements et des contraintes réglementaires.

Toutes les équipes de projet se positionnent clairement en faveur de la science ouverte. La volonté de partager les résultats du travail scientifique fait partie des principes du métier de chercheur. A titre d'exemple et parlant pour d'autres, la responsable du *Panel de Caen* se dit convaincue de *l'intérêt du partage des données, qu'elle pratiquait déjà de manière informelle dès le début de sa carrière. Aujourd'hui, elle apprécie que cette pratique soit institutionnalisée et intégrée dès la conception des projets, ce qui permet de gagner du temps et d'améliorer la qualité des données.* Bien sûr, cette volonté d'ouvrir les données a des limites (données personnelles et/ou confidentielles...) et nécessite dans certains cas du travail supplémentaire. Mais il n'y a pas d'opposition entre le travail scientifique et la gestion des données. La diffusion des résultats et en particulier, des données, est un enjeu central du projet. Dans le cadre de *Subwork*, la production d'une *base de données anonymisée et diffusable, organisée en un ensemble cohérent de fichiers [...] représente un apport méthodologique important, dans la mesure où il permet de concilier l'exigence d'une analyse fine des emplois et des territoires avec les contraintes liées à la confidentialité des données.* Le projet ne se limite donc pas à une enquête sur la ville productive ; il fournit également une *infrastructure de recherche susceptible d'être mobilisée par d'autres.* Avec les mots de Cavallo & Noûs (2019) : « La gestion des données implique une éthique de la donnée qui est aussi et avant tout une éthique de la recherche » (p.113).

Une autre observation est directement liée au planning et à la temporalité des projets. D'une part, comme le fait remarquer la responsable du projet *Les Vocaux*, il faut intégrer la dimension donnée dès le montage du projet : *L'expérience acquise conduit à une conclusion forte : toute démarche d'ouverture et de partage des données suppose une réflexion préalable rigoureuse sur les droits des personnes et sur les modalités de consentement. L'intégration précoce de ces dimensions apparaît comme une condition essentielle à la fois de la conformité réglementaire et de la qualité scientifique du projet.* Là où cela n'a pas été fait, il faut trouver des solutions a posteriori, comme pour le *Panel de Caen* : *Concernant le consentement des participants, il faut rappeler que les normes actuelles n'existaient pas au moment du lancement de l'enquête (1995). Les dispositifs de consentement étaient donc très limités, souvent réduits à des fiches d'information et, pour les mineurs, à des autorisations parentales. De ce point de vue, l'enquête apparaît aujourd'hui en décalage avec les exigences contemporaines.*

D'autre part, il faut rester flexible et adapter les pratiques de gestion au développement des outils, des méthodes et de l'environnement. *La chaîne de traitement des données a évolué au cours des trois années du projet (Les Vocaux). Le traitement des données dans le cadre du projet Subwork s'est organisé en plusieurs étapes successives, combinant un important travail de reconstruction statistique, d'harmonisation des nomenclatures et de spatialisation fine des emplois (Subwork).*

Une autre observation faite par les équipes de projet est l'importance des moyens informatiques, financiers et humains, souvent manquants ou insuffisants dans les projets et structures SHS. A titre d'exemple, la responsable du projet *Les Vocaux* regrette que *certaines opérations initialement envisagées, notamment un alignement très fin du signal jusqu'au niveau phonémique pour permettre des analyses phonétiques approfondies, se sont révélées trop coûteuses en temps et en ressources humaines [...] Faute de moyens suffisants, cette étape a été partiellement abandonnée, même si certains collègues ont poursuivi ce type d'analyse de manière autonome à partir des données mises à disposition.* Une manière de faire face au manque de moyens est la mutualisation des infrastructures (cf. Huma-Num et NAKALA, ORTOLANG), des méthodes (codes...) voire même des personnels, notamment des ingénieurs d'études et de recherche.

Concernant les personnels, dernière observation : Tous les interlocuteurs soulignent le rôle central du travail en équipe. La gestion des données et plus particulièrement leur ouverture et mise à disposition n'est jamais une affaire individuelle ou personnelle. Aussi, la responsable du *Panel de Caen* insiste sur l'importance du travail collectif et sur le rôle essentiel des collaborateurs et des ingénieurs, aussi pour la formation des autres membres de l'équipe, tandis que la responsable du projet *e-NdP* le formule très simplement ainsi : *Ce qui a été indispensable, ça a été les ingénieurs ou assistants ingénieurs en informatique des divers laboratoires.*

Conclusion

En définitive, l'analyse des quatre projets met en évidence une dynamique convergente vers une amélioration de la gestion, de la structuration et de l'ouverture des données de recherche. Cette convergence ne doit toutefois pas masquer la diversité des trajectoires observées, étroitement dépendantes des contextes disciplinaires, des objets étudiés et des contraintes juridiques ou techniques propres à chaque projet. À cet égard, le projet *e-NdP* se distingue par son caractère exemplaire : en articulant numérisation patrimoniale, intelligence artificielle, ingénierie documentaire et réflexion méthodologique sur la qualité des données, il incarne une transformation aboutie des pratiques en humanités numériques. En offrant un accès inédit à des corpus manuscrits de longue durée, il renouvelle les conditions de la recherche et propose un modèle reproductible pour d'autres initiatives comparables.

Plus largement, ces études de cas révèlent des niveaux de maturité différenciés, allant de pratiques encore largement informelles à des dispositifs fortement structurés et outillés. Elles soulignent également la centralité des processus de transformation et d'enrichissement des données, qui constituent un dénominateur commun à l'ensemble des projets. Cette évolution s'accompagne d'une professionnalisation croissante, marquée par le rôle déterminant des ingénieurs, des infrastructures et des dispositifs d'accompagnement, désormais essentiels à la production, à la documentation et à la diffusion des données.

Dans ce contexte, les principes FAIR offrent un cadre de référence particulièrement pertinent pour interpréter ces évolutions. Les projets analysés attestent d'avancées significatives en matière de mise à disposition et de structuration des données (accessibilité et trouvabilité), ainsi que d'une intensification des efforts en matière de documentation et d'enrichissement, contribuant à en renforcer la réutilisabilité. Toutefois, leur mise en œuvre demeure inégale, notamment en ce qui concerne l'interopérabilité, encore limitée par la diversité des formats, des standards et des pratiques disciplinaires. Par ailleurs, les contraintes liées à la protection des données personnelles et aux cadres réglementaires introduisent des tensions qui viennent nuancer l'application pleine et entière de ces principes.

Il apparaît ainsi que les pratiques effectives précèdent souvent les formalismes : même en l'absence de plans de gestion de données formalisés, des dynamiques de mise en conformité progressive avec les principes FAIR sont à l'œuvre, bien que de manière hétérogène ou informelle. La question de la documentation s'impose comme un levier central pour renforcer la réutilisabilité et l'interopérabilité des données.

Dès lors, l'enjeu principal réside dans la capacité à articuler des processus de standardisation - nécessaires à l'opérationnalisation des principes FAIR - avec la prise en compte de la diversité des données, des méthodes et des contextes de recherche. Dans cette perspective, les dispositifs de reconnaissance institutionnelle, tels que le Prix Science ouverte, apparaissent comme des leviers importants pour encourager, structurer et valoriser ces transformations, contribuant ainsi à l'émergence progressive de nouvelles normes au sein des communautés scientifiques.

Remerciements

L'auteur remercie vivement les interlocuteurs pour leur motivation, engagement et disponibilité : Claire Bidart, Julie Claustre, Julie Glikman, Thibault Le Corre, Nicolas Raimbault et Aliette Roux. Il remercie également Florence Thiault pour ses conseils lors de la préparation des entretiens et de la rédaction de l'article.

Déclaration d'intérêts conflictuels

L'auteur certifie qu'aucun conflit d'intérêt n'a influencé la conduite, l'analyse ou la rédaction de cet article.

Références bibliographiques

- Amiel, P., Frontini, F., Lacour, P.-Y., & Robin, A. (2020). Pratiques de gestion des données de la recherche : une nécessaire acculturation des chercheurs aux enjeux de la science ouverte ? *Cahiers Droit, Sciences & Technologies*, (10), 147–168. <https://doi.org/10.4000/cdst.2061>
- Bendjaballah, S., & Garcia, G. (2024). Les sciences sociales à l'épreuve de l'ouverture des données de la recherche. *Terrains & Travaux*, N° 43(2), 211–216. <https://doi.org/10.3917/tt.043.0211>
- Bidart, C., Bonneville, A., & Garcia, G. (2025). Ouvrir les coulisses d'une recherche longitudinale sur les parcours de vie et les reseaux personnels : le Panel de Caen (1995-2015). *L'année sociologique*, 75(1), 259-299. <https://shs.cairn.info/revue-lannee-sociologique-2025-1-page-253?lang=fr>
- Borgman, C. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge MA, The MIT Press. <https://mitpress.mit.edu/books/big-data-little-data-no-data>
- Brochard, P. & Claustre, J. (2021). *Plan de gestion de données pour le projet e-NdP*. (hal-03419065)
- Cavallo, D., & Noûs, C. (2019). Données à penser. Enjeux pratiques et éthiques autour des données dans le montage de projets de recherche européens. *Tracés : Revue de Sciences Humaines*, 19 (hors-série), 101-113. <https://doi.org/10.4000/traces.10793>
- Chawinga, W. D., & Zinn, S. (2019). Global perspectives of research data sharing: A systematic literature review. *Library & Information Science Research*, 41(2), 109–122. <https://doi.org/10.1016/j.lisr.2019.04.004>
- Cocaud, S., L'Hostis, D., & Aventurier, P. (2018). Les plans de gestion de données (PGD). *Cahier des Techniques de l'INRA*, N° Spécial : Données de la recherche. <https://hal.inrae.fr/hal-04693775v1>
- Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). *The Fourth Paradigm. Data-Intensive Scientific Discovery*. Redmond WA, Microsoft Corp. <http://research.microsoft.com/en-us/collaboration/fourthparadigm/>
- Igarashi, Y., Nagata, K., Kuwatani, T., Omori, T., Nakanishi-Ohno, Y., & Okada, M. (2016). Three Levels of Data-Driven Science. *Journal of Physics: Conference Series*, 699, 012001. <https://doi.org/10.1088/1742-6596/699/1/012001>
- Koltay, T. (2019). Accepted and Emerging Roles of Academic Libraries in Supporting Research 2.0. *The Journal of Academic Librarianship*, 45(2), 75–80. <https://doi.org/10.1016/j.acalib.2019.01.001>
- Malingre, M.-L., Mignon, M., Pierre, C., & Serres, A. (2019). Construction(s) et contradictions des données de recherche en SHS. *Recherche d'information, Document et Web Sémantique*, 2(1), 1–21. <https://doi.org/10.21494/ISTE.OP.2019.0336>
- Rebouillat, V. (2019). Les données scientifiques face aux enjeux de la recherche en Sciences, Technologie et Médecine : enquête exploratoire à l'Université de Strasbourg. *Études de communication - Langages, information, médiations*, 52, 99-116. <https://doi.org/10.4000/edc.8681>
- Schöpfel, J. (2018). Hors norme ? Une approche normative des données de la recherche. *Revue COSSI*, (5). <https://revue-cossi.info/numeros/n-5-2018-processus-normalisation-durabilite-information/730-5-2018-schopfel>
- Schöpfel, J., & Prost, H. (2016). Research data management in social sciences and humanities: A survey at the University of Lille 3 (France). *LIBREAS. Library Ideas*, 29, 98–112. <https://doi.org/https://doi.org/10.18452/9097>
- Schöpfel, J., Kergosien, E., & Prost, H. (2017). « Pour commencer, pourriez-vous définir 'données de la recherche' ? » Une tentative de réponse. *Atelier VADOR : Valorisation et Analyse Des Données de La Recherche*, INFORSID 2017, 31 Mai 2017 Toulouse (France). <https://lilloa.hal.science/hal-01530937/>

Thoegersen, J. L., & Borlund, P. (2022). Researcher Attitudes Toward Data Sharing in Public Data Repositories: A Meta-Evaluation of Studies on Researcher Data Sharing. *Journal of Documentation*, 78(7), 1–17. <https://doi.org/10.1108/JD-01-2021-0015>

Torres Aguilar, S., & Jolivet, V. (2023). Handwritten Text Recognition for Documentary Medieval Manuscripts. *Journal of Data Mining and Digital Humanities*, jdmdh:10484, 22 décembre 2023, Documents historiques et reconnaissance automatique de texte. <https://doi.org/10.46298/jdmdh.10484>

Wang, J., Li, Y., Wang, S., Shi, L., & Duan, B. (2024). Global Impact Analysis and Implementation Strategy Recommendations of the FAIR Principles. *Chinese Science Bulletin*, 69(9), 1183–1191. <https://doi.org/10.1360/TB-2023-0585>

Wilkinson, M. D., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*, 3(1), 160018. <https://doi.org/10.1038/sdata.2016.18>

Wilkinson, M. D., Sansone, S.-A., Schultes, E., Doorn, P., Bonino da Silva Santos, L. O., & Dumontier, M. (2018). A Design Framework and Exemplar Metrics for FAIRness. *Scientific Data*, 5(1), 180118. <https://doi.org/10.1038/sdata.2018.118>

Liste des projets

La description des projets correspond aux verbatims des transcriptions validées. Les liens ont été ajoutés.

E-NPD

Le projet récompensé par le Prix de la science ouverte s'inscrit dans un programme ANR mené entre 2021 et 2025, consacré à Notre-Dame de Paris et son cloître, dans le contexte du vaste chantier scientifique ouvert après l'incendie de 2019. Alors que de nombreuses recherches se concentraient sur la matérialité de l'édifice - pierre, charpente, vitraux, techniques de construction - l'équipe a choisi d'explorer un autre versant du patrimoine de la cathédrale : sa « cathédrale de parchemin et de papier », c'est-à-dire l'ensemble des archives et de la bibliothèque historique du chapitre cathédral. Ce déplacement du regard permettait d'appréhender non seulement le monument comme objet architectural, mais aussi comme institution vivante, insérée dans des réseaux religieux, politiques, économiques et sociaux sur la longue durée.

Domaine : histoire / humanités numériques

Site du projet

<https://endp.chartes.psl.eu/endp/>

Les données

Sur un serveur de l'École nationale des chartes : <https://endp.chartes.psl.eu/endp/ressources>

Sur Zenodo : Claustre, J., Smith, D., Torres Aguilar, S., Bretthauer, I., Brochard, P., Canteaut, O., Cottureau, E., Delivré, F., Denglos, M., Jolivet, V., Julerot, V., Kouamé, T., Lusset, E., Massoni, A., Nadiras, S., Perreaux, N., Regazzi, H., & Treglia, M. (2023). The e-NDP project : collaborative digital edition of the Chapter registers of Notre-Dame of Paris (1326-1504). Ground-truth for handwriting text recognition (HTR) on late medieval manuscripts. (1.0, p. <https://zenodo.org/record/7575693>) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.7575693>

Sur Zenodo : Torres Aguilar, S., & Jolivet, V. (2024). TRIDIS: HTR model for Multilingual Medieval and Early Modern Documentary Manuscripts (11th-16th) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.10800223>

Le prix

Prix Science ouverte des données de la recherche 2025, catégorie « créer les conditions de la réutilisation » - cf. Site Ouvrir la science du MESR du 1^{er} décembre 2025

PANEL DE CAEN

Le projet de recherche porte sur un panel, c'est-à-dire une enquête longitudinale consistant à interroger les mêmes personnes à plusieurs reprises dans le temps. L'objectif de cette recherche est de comprendre comment les relations interpersonnelles - famille, amis, collègues, voisins - influencent les parcours de vie. Les individus ne se construisent jamais seuls, et leur trajectoire est constamment façonnée par leur entourage, lui-même évolutif. L'équipe de recherche a donc souhaité étudier conjointement l'évolution des réseaux relationnels et celle des trajectoires individuelles, afin de saisir leur co-construction. L'originalité de l'enquête réside dans sa durée exceptionnelle : elle s'étend sur vingt ans (1995-2015), avec six vagues d'interrogation d'une même cohorte de jeunes initialement situés à Caen, d'où le nom de « panel de Caen ». Le dispositif reposait à la fois sur des outils très structurés (calendriers, listes de relations) et sur de longs entretiens approfondis, pouvant durer entre cinq et onze heures.

Domaine : sociologie

Site du projet

<https://panelcaen.hypotheses.org/>

Les données

Dans l'entrepôt des données de Sciences Po (Banque de données du Centre de données sociopolitiques CDSP) : Bidart, Claire; Degenne, Alain; Kornig, Cathel; Lavenue, Daniel; Mounier, Lise; Pellissier, Anne; Le Gall, Didier; Beynier, Dominique; Lemarchand, Clotilde, 2023, "Panel de Caen", <https://doi.org/10.21410/7E4/OCUEJL>, data.sciencespo, V4

Le prix : Prix Science ouverte des données de la recherche 2025, catégorie « mention spéciale du jury » - cf. Site Ouvrir la science du MESR du 1^{er} décembre 2025 <https://www.ouvrirelascience.fr/remise-des-prix-science-ouverte-des-donnees-de-la-recherche-edition-2025/>

SUBWORK

Le projet Subwork s'inscrit dans le cadre d'un programme de recherche soutenu par le Plan Urbanisme Construction Architecture et consacré aux transformations contemporaines de la « ville productive ». Son ambition principale est d'interroger empiriquement cette notion en l'ancrant dans l'analyse fine des activités économiques, des catégories d'emploi, des groupes sociaux associés, et de leurs localisations dans la ville. À rebours d'une approche strictement sectorielle ou morphologique, le projet propose de partir des emplois eux-mêmes afin de mieux comprendre la place qu'occupent les activités productives dans les dynamiques urbaines contemporaines et, en particulier, dans la structuration sociale des territoires.

L'hypothèse centrale est que l'étude des emplois relevant des sphères productives permet de renouveler l'analyse du travail des classes populaires en ville. En croisant une définition rigoureuse des activités productives avec les catégories socioprofessionnelles correspondantes, le projet vise à identifier précisément les groupes sociaux concernés et à mesurer leur poids relatif. Cette démarche s'accompagne d'un effort de spatialisation particulièrement poussé, reposant sur une localisation des emplois à une échelle infra-communale, notamment celle des IRIS. Proposer une géographie des emplois au lieu de travail à l'échelle infracommunale est inédit au regard des données aujourd'hui diffusées par l'Insee. Il devient ainsi possible de cartographier avec finesse les espaces urbains où se concentrent ces activités et d'analyser leurs logiques d'implantation.

Domaine : sciences sociales / urbanisme

Site du projet

<https://www.urbanisme-puca.gouv.fr/subwork-les-espaces-suburbains-de-production-quels-a2207.html>

Les données

Sur NAKALA : Heitz, A., Le Corre, T., Raimbault, N., ROUX, A., Subwork, E., & Tranchant, L. (2023). Subwork : pour de nouvelles géographies du travail (Version 1) [Data set]. NAKALA - <https://nakala.fr> (Huma-Num - CNRS). <https://doi.org/10.34847/NKL.C8CALJC9>

Cf. aussi le site d'exploration des données sur Huma-Num, vitrine de la base de données https://tlecorre.gitpages.huma-num.fr/Subwork_website/

Le prix

Prix Science ouverte des données de la recherche 2025, catégorie « créer un jeu de données manquantes » - cf. Site Ouvrir la science du MESR du 1^{er} décembre 2025 <https://www.ouvrirlascience.fr/remise-des-prix-science-ouverte-des-donnees-de-la-recherche-edition-2025/>

LES VOCaux

L'entretien porte sur le projet LesVocaux dont le corpus est constitué de SMS vocaux authentiques, effectivement échangés entre usagers sur leurs téléphones mobiles sur des applications de discussion (comme WhatsApp, etc.). C'est le premier corpus rassemblant des données de ce nouveau mode de communication. Pour l'étude de la langue, c'est un accès privilégié à des données spontanées produites en contexte non surveillé et relevant de l'intime. Ce projet nous permet ainsi d'étudier la parole naturelle sans les biais des recueils en laboratoire, mais aussi en contexte de communication médiée. Il a été conçu dans une perspective de linguistique de la parole et inscrit dans une démarche affirmée de science ouverte.

Domaine : linguistique / communication

Site du projet

<https://www.atilf.fr/ressources/corpus-les-vocaux/>

Les données

Sur ORTOLANG : Julie Glikman, Nicolas Mazziotta, Christophe Benzitoun, et al. (2025). LesVocaux [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, v0.0.2, <https://doi.org/10.82270/lesvocaux/v0.0.2>.

Le prix

Prix Science ouverte des données de la recherche 2025, catégorie « créer un jeu de données manquantes » - cf. Site Ouvrir la science du MESR du 1^{er} décembre 2025 <https://www.ouvrirlascience.fr/remise-des-prix-science-ouverte-des-donnees-de-la-recherche-edition-2025/>

Trame d'entretien

1. Présentation du projet (5 min)

Pouvez-vous présenter brièvement le projet pour lequel vous avez reçu le Prix de la Science ouverte ?

Quels étaient ses objectifs principaux ?

2. Données produites et organisation (10 min)

Quels types de données avez-vous produits ou collectés ?

Aviez-vous mis en place un plan de gestion des données ?

Comment la responsabilité de la gestion des données était-elle organisée au sein de l'équipe ?

Quels outils ou plateformes (GitLab, entrepôts disciplinaires, cahiers de labo électroniques) ont été cruciaux pour la traçabilité de votre travail ?

3. Compétences, accompagnement et pratiques (5 min)

Avez-vous bénéficié de formations ou d'un accompagnement sur la gestion des données ?

Quels appuis ont été les plus déterminants pour vous ?

4. Ouverture et valorisation des données (5 min)

Les données ont-elles été publiées ou rendues accessibles ?

Avez-vous rédigé ou envisagé un article de données ?

Quels bénéfices ou difficultés reprenez-vous de cette démarche ?

Avez-vous rencontré des verrous juridiques ou éthiques (RGPD, propriété intellectuelle) ? Si oui, comment les avez-vous levés pour permettre l'ouverture ?

5. Apport du prix et perspectives (5 min)

Qu'a apporté l'obtention du Prix de la Science ouverte à votre parcours ou à votre projet ?

Ces pratiques ont-elles été poursuivies ou étendues à d'autres projets ?

Avez-vous déjà eu des retours de chercheurs ayant réutilisé vos données ? Ou imaginez-vous des usages détournés/inattendus de votre jeu de données dans d'autres disciplines ?