Sous la direction de Claire Scopsi, Clothilde Roullier, Martine Sin Blima-Barru et Édouard Vasseur

Les nouveaux paradigmes de l'archive

La recherche ouverte et les données en Lettres, Sciences humaines et sociales (LSHS)

Le cas des GLAM

Gérald Kembellec et Claire Scopsi

Éditeur : Publications des Archives nationales
Lieu d'édition : Pierrefitte-sur-Seine
Publication sur OpenEdition Books : 9 février 2024
Collection : Actes
ISBN numérique : 978-2-86000-390-2

OpenEdition Books
https://books.openedition.org

Référence numérique
Kembellec, Gérald, et Claire Scopsi. « La recherche ouverte et les données en Lettres, Sciences humaines et sociales (LSHS) ». *Les nouveaux paradigmes de l'archive*, édité par Claire Scopsi et al., Publications des Archives nationales, 2024, https://doi.org/10.4000/books.pan.7298.

La recherche ouverte et les données en Lettres, Sciences humaines et sociales (LSHS)

Le cas des GLAM

Gérald Kembellec et Claire Scopsi

Introduction

1 Depuis 2018, la « recherche ouverte » est une des priorités du ministère de l'Enseignement supérieur et de la Recherche français [MESR]. Il s'agit d'un ensemble de mesures progressives impactant les acteurs de la recherche française : chercheurs, personnels des universités et des institutions scientifiques, éditeurs scientifiques. L'objectif est de doter le pays d'une politique cohérente et systémique conduisant à apporter plus de transparence à la recherche financée grâce à des fonds publics et à en restituer largement les résultats. Deux plans triennaux se sont succédé, le second s'achèvera en 2024[1]. De 2018 à 2021, l'accent a été porté sur l'ouverture des publications et notamment l'accès ouvert aux revues scientifiques[2], ainsi que l'élaboration de plans de gestion de données dès la conception des projets de recherche. De 2021 à 2024, les mesures visent l'ouverture des codes source produits dans le cadre de la recherche publique et le partage des données *via* des plateformes dédiées et ouvertes.

2 Nous revenons dans cet article sur ces mesures en montrant de quelle façon elles impliquent les chercheurs, influencent la composition des équipes projet et renouvellent les modalités de publication. Dans une première partie, nous présentons les directives, les instances et les ressources qui constituent l'écosystème de la politique publique de l'ouverture des données de la recherche. Nous regardons également la manière dont les fondamentaux de ces politiques, la conservation et le partage, rencontrent les conditions actuelles d'exercice du métier de chercheur en sciences humaines et sociales, invité à se rapprocher des méthodes des sciences « dures ». Nous constatons qu'elles impliquent inévitablement d'intégrer la dimension

pluridisciplinaire dans les projets. Dans la partie suivante, nous nous appuyons sur l'exemple des GLAM[3], pour montrer que les experts de la discipline, qui travaillent sur des corpus numériques, numérisés ou transformés en données, n'abandonnent pas les méthodes et les bonnes pratiques traditionnelles, mais doivent les prolonger dans l'univers numérique. Cependant, la modélisation des objets de recherche, qui est au centre des humanités numériques, les conduit à s'appuyer sur l'expertise des informaticiens et les archivistes sont des alliés précieux lorsqu'il s'agit d'anticiper le cycle de vie des données par la rédaction d'un plan de gestion des données [PGD]. La troisième partie présente les solutions disponibles pour la publication et le partage des données et de leur documentation.

L'ouverture des données de la recherche

3 Les mesures exposées dans le second plan national pour la science ouverte ont pour objectif d'adopter les principes structurants de la recherche ouverte pour toutes les productions de la recherche publique, c'est-à-dire les mémoires, publications, données, logiciels, etc. Toutes les productions émanant « d'une activité de recherche financée au moins pour moitié par des dotations de l'État, des collectivités territoriales ou des établissements publics, par des subventions d'agences de financement nationales ou par des fonds de l'Union européenne[4] » doivent respecter cette politique et les règles de partages qui y sont associées. Rappelons que ces principes, désignés par l'acronyme FAIR, sont « faciles à trouver, accessibles, interopérables et réutilisables ».

Une vision politique

4 En application de la loi pour une république numérique[5] promulguée en 2016, le deuxième axe, « Structurer, partager et ouvrir les données de la recherche » (MESRI, 2021, p. 12-15), pose trois mesures concrètes : la mise en œuvre de l'obligation de diffusion des données de recherche financées sur fonds publics (mesure 4), la création de la plateforme nationale fédérée *Recherche Data Gouv* (mesure 5) et la promotion de l'adoption d'une politique de données sur l'ensemble du cycle des données de la recherche (mesure 6).

5 Ces orientations se concrétisent par la mise en place d'un « écosystème » constitué de réseaux de correspondants, plateformes et de financement de projet. Sur le thème des données, l'animation de la politique de gestion des données de la recherche repose sur de multiples initiatives :

- le Comité pour la science ouverte [CoSo] a pour mission de définir la politique de sciences ouvertes, d'en coordonner la mise en œuvre et de la développer au niveau national et international. Présidé par le directeur général de la recherche et de l'innovation [DGRI][6], il s'appuie sur un comité de pilotage, composé de représentants des établissements de l'enseignement supérieur et de la recherche et des acteurs de la science ouverte, qui se réunit deux fois par an ;
- la mise en place d'un réseau d'administrateurs au sein des établissements afin d'assurer la coordination. Des ateliers de la donnée, recrutés par appels à projets, sont mis en place dans une quinzaine d'établissements supérieurs[7]. Labellisés par le CoSO et structurés en réseau autour d'un bureau national, ils appuient en local les établissements et les chercheurs pour propager la culture de l'*open science,*

mettre en place une offre de formation sur la gestion des données de la recherche ou aider à la rédaction des plans de gestion de données désormais obligatoire dans la constitution de dossiers de réponse aux appels à projets nationaux et européens ;

- des plateformes de partage de connaissances et de fédération des ressources sur l'*open science* émanant de réseaux de partenaires, d'établissements ou d'associations prolongent les actions et assurent l'information et l'initiation aux techniques d'ouverture des données ;
- la visibilité des données produites est assurée par un entrepôt pluridisciplinaire, *Recherche Data Gouv*[8], confié à l'INRAE et ouvert en 2022. Sa mission est d'héberger des jeux de données et de référencer par moissonnage ceux qui sont hébergés dans d'autres entrepôts institutionnels, disciplinaires ou internationaux de façon à fournir un catalogue complet ;
- le choix de licences libres pour les données et code sources. En matière informatique, la reproductibilité des résultats de la recherche demande non seulement la mise à disposition des données, mais également celle des codes sources des logiciels qui en assurent le traitement, qu'il s'agisse de scripts, de macros Excel ou de logiciels plus élaborés (MESRI, 2022). L'ouverture des codes implique que les détenteurs des droits patrimoniaux sur le code soient identifiés et qu'ils optent pour une licence libre. Le code peut être archivé et décrit dans la plateforme *Software Héritage*[9] et référencé dans Hal ;
- le Fonds national pour la science ouverte [FNSO] est un GIS [groupement d'intérêt scientifique] créé en 2019, constitué de financements publics et privés. Il soutient les projets sélectionnés dans le cadre d'appels thématiques. Il a permis le financement de la plateforme et des ateliers de la donnée de recherche.data.gouv.fr. Un troisième appel à projet a été lancé en 2023.

6 Comme l'indique la mesure 6 du second plan, la politique de données concerne l'ensemble du cycle des données de la recherche. L'Inist[10] s'inspire du modèle de *Research Data Life* de la *UK Data Archive* pour définir ce cycle en six étapes *:*

- traitement des données (*processing data*),
- analyse des données (*analysing data*),
- conservation des données (*preserving data*),
- accès aux données (*giving access to data / data discovery*),
- réutilisation des données (*reusing data*).

7 Toutefois les appels à projet du FNSO sont encore très orientés vers la publication. On peut cependant relever quelques soutiens à des plateforme d'archives ouvertes[11].

8 Nous assistons donc à la mise en place d'un ensemble hétérogène d'initiatives, d'acteurs, d'outils méthodologiques et techniques et d'actions de support qui visent à modifier la posture et les pratiques de tous les chercheurs à l'égard des données qu'ils manipulent. À l'autre bout de la chaîne, sur le terrain même de la recherche, comment les chercheurs abordent-ils cette injonction à l'ouverture et à la conservation des données qu'ils recueillent ou produisent ? En effet, en termes pratiques, penser les données de la recherche peut s'avérer une démarche complexe.

L'impact pour les professionnels de la recherche : pourquoi apprendre à penser les données ?

9 Il est évident que les politiques publiques évoquées précédemment vont avoir un impact sur les métiers des acteurs de la recherche. Si l'on se réfère à une vision socio-anthropologique du chercheur en tant qu'individu socialisé en milieu professionnel, il est un certain nombre de besoins et d'objectifs auxquels un acteur de la science est confronté.

10 Latour et Woolgar (1986), Bourdieu (1972, 1977) entre autres, tous repris dans une synthèse par Vigni *et al.* (2023) ont théorisé successivement – sur le temps long – la question du « cycle de crédibilité » qui présente les résultats scientifiques comme une « monnaie d'échange » sociale, mobilisée pour obtenir du crédit (social et financier) qui permet au chercheur de poursuivre son travail de recherche. D'autres auteurs, issus des milieux internationaux des *library and information sciences*, ont également participé à la réflexion. En examinant les besoins des chercheurs, ils montrent que dans le cycle du travail scientifique, collecter, analyser, partager et discuter des résultats[12] sont des étapes importantes qui doivent être prises en compte dans l'élaboration des outils et services documentaires (Palmer *et al.*, 2009, Unsworth, 2000, Nakakoji *et al.*, 2015).

11 D'un point de vue pratique, afin de faciliter la circulation des travaux de recherche, des théoriciens du Web comme Berners-Lee[13] (2004, 2010), Shotton (2009), Peroni (2017) ou Kuhn (2017) proposent des principes et des outils d'édition en ligne qui rendent les corpus facilement découvrables et compréhensibles par les machines : des dispositifs de lecture équipée, des moteurs de recherche ou des outils d'extraction et de filtrage automatique (*scraping*). Ces principes sont mis en œuvre, par exemple, dans des systèmes d'édition et d'écriture scientifique comme « Stylo »[14] de l'éditeur canadien Érudit. Ils sont également proposés dans des dispositifs comme Omeka S, inclus dans l'offre de l'infrastructure de recherche française Huma-Num. Ces outils techniques, fondés sur des technologies, des identifiants et des modèles de données standardisés, proposent de relier entre eux les éléments associés à une production scientifique comme les articles, les données, les logiciels éventuels ou les profils des chercheurs. Cette bonne pratique passe par l'observation des règles du concept de *semantic publishing* (Peroni *et al.*, 2017, Shotton, 2009) qui garantit la bonne accessibilité FAIR et donc la bonne visibilité, la diffusion et la ré-exploitation des travaux, des données et des documentations incluses. Toutefois, cette exploitation du Web de données, socle conceptuel et technique du *semantic publishing*, est complexe et requiert, autour du chercheur lui-même, la collaboration d'informaticiens (pour la bonne implémentation technique) et des archivistes[15] (pour garantir les bonnes pratiques de classement, de nommage et de préservation des corpus).

Les humanités numériques : des méthodes de plus en plus « dures »

12 Dans le domaine des humanités numériques, Pierre Mounier a clairement démontré que les évolutions techniques, la mise en données de la société, la compétition, la course aux financements transforment le métier des chercheurs et de leurs équipes dans leurs pratiques quotidiennes. Cette transformation a commencé dans les sciences dites dures, habituées à manipuler des données avec des protocoles stricts. Elle affecte désormais également les sciences humaines et sociales avec des méthodes similaires

(Mounier, 2018, p. 9-19). Cette analyse était d'ailleurs partagée par Bruno Latour depuis bien longtemps :

> Vous pouvez bien établir toutes les nuances que vous voulez entre sciences *soft* et sciences *hard*, vous êtes obligé de reconnaître que les sciences sociales se constituent elles aussi par l'intermédiaire de questionnaires, de collections, de banques de données [...]. (Latour, 2001, p. 70)

13 On assiste notamment à une injonction à appliquer des méthodes issues des sciences dites dures aux disciplines des lettres, sciences humaines et sociales [LSHS] pour crédibiliser leurs méthodologies auprès des organismes financeurs des projets, car la transparence et la reproductibilité des recherches sont des gages supposés de crédibilité et de véracité (Mounier, 2018, p. 9-19). Penser la création ou la numérisation de corpus pour en faire des objets numériques manipulables fait partie de ces nouvelles méthodes, transformer des textes issus d'archives ou des œuvres d'art en bases de données également.

Modéliser un corpus : une approche multidisciplinaire

14 Cette transformation s'accompagne d'une réflexion pointue, car les informations contenues dans les corpus sont loin d'être vues de manière unanime au sein d'une même discipline, d'un courant à un autre et même d'un chercheur à un autre. Il y a donc une part importante de subjectivation dans l'appréciation de l'objet étudié. Dans le cadre d'un projet interdisciplinaire, la problématique se complexifie encore. Cela va donc avoir une importance sur la méthode de description du corpus et de ses « documents » – et, à l'instar de Suzanne Briet, nous partons du principe générique que « tout » est document, c'est-à-dire objet descriptible[16] –, c'est ce qu'on appelle la modélisation. L'enjeu de ces questions est à la fois info-documentaire (comment l'on décrit), informatique (la manière d'inscrire, de stocker, de propager et de consulter), linguistique au sens large du terme (les idées derrière les mots et les individus), tout en relevant, bien sûr, de la discipline originale liée au projet de recherche. De prime abord, cette tâche peut sembler rebutante voire insurmontable aux chercheurs des LSHS[17] qui, s'ils sont spécialistes de leur domaine, n'ont pas forcément les connaissances nécessaires pour en réaliser seuls la modélisation. Ils n'ont pas toujours non plus le temps, ni l'envie, de s'investir dans un processus d'apprentissage coûteux. Heureusement, nous verrons que cette étape peut être grandement facilitée par des collaborations interdisciplinaires et/ou par l'aide des nombreux services d'appui à la recherche, qu'ils soient nationaux ou locaux.

Implications interdisciplinaires dans les projets sur corpus numériques : l'exemple des données des GLAM

15 Les recherches liées aux disciplines historiques, à l'histoire de l'art, à la muséologie, à l'archivistique ou encore à la gestion info-documentaire sont un secteur particulièrement intéressant à considérer pour traiter du processus de mise en données et de partage des matériaux de la recherche. Les acteurs de ces disciplines se sont très tôt dotés d'outils adaptés au travail sur de corpus numériques. Rappelons que c'est le centre historique *Roy Rosenzweig*[18] qui a développé le logiciel d'édition bibliographique Zotero et Omeka, le logiciel de publication et de valorisation en ligne de collections

numériques ou numérisées. Si, au premier abord, les documents d'archives qui constituent les corpus d'études des historiens semblent bien éloignés de la problématique des *data*, des projets innovants, comme Biblissima[19] (autour des manuscrits numérisés) ou Gloss-e[20] (étude de la Glose), ont su nous convaincre qu'une relation harmonieuse et fructueuse pouvait exister entre le numérique et les études historiques.

16 Quelles sont les conditions d'un dialogue interdisciplinaire autour des données en sciences humaines ? Nous proposons ici d'analyser comment chercheurs, informaticiens et archivistes peuvent associer leurs expertises au sein d'un même projet.

Questions de discipline : valider les données et leurs sources

17 Pour les historiens, comme pour les archivistes, la méthode diplomatique[21] est fondamentale. Ces disciplines se doivent de vérifier l'authenticité et la fiabilité d'un document original par une analyse rigoureuse de sa forme : c'est un élément de la crédibilité du fait historique. Mais les humanités numériques ajoutent des étapes de transformation de cet original (par exemple numérisation, océrisation, extractions de contenus) afin d'y appliquer des outils d'analyse automatique. Il devient nécessaire de s'assurer que ces opérations n'ont pas altéré la structure et le sens de la source originale. C'est la raison pour laquelle les chercheurs des disciplines historiques ne devraient s'appuyer sur des bases de données qu'à condition d'en connaître les étapes de constitution et en d'en vérifier la bonne exécution.

18 Dans l'idéal, le chercheur en humanités ne devrait pas se fier aux données stockées dans une base sans avoir consulté un fac-similé de qualité de la source originale ou même, idéalement, d'avoir compulsé la source originale. Malheureusement cela peut être difficile si ces sources sont conservées dans un lieu éloigné et peut remettre en question l'un des attraits du partage des données numériques de la recherche : donner facilement et rapidement accès à des corpus déjà assemblés.

19 C'est pourquoi différents outils et méthodes doivent être identifiés pour expliquer et garantir la qualité des opérations de constitution des bases :

1. connaître l'emplacement de la source originale, si elle existe encore, afin de pouvoir s'y reporter en toute extrémité et la compulser physiquement ;
2. disposer d'un fac-similé de qualité et pas seulement des données extraites de la source, afin de disposer des précieuses informations de contexte : ratures, disposition originelle du texte, éléments graphiques ou graphologiques, etc. N'oublions pas aussi qu'une base de données peut faire oublier les caractéristiques de l'époque et décorréler une information de son contexte ;
3. une base de données ne dispense pas de représenter visuellement les éléments dans le contexte de leur époque, lorsque l'on élabore un dispositif de filtrage et de visualisation.

20 Ainsi, dans un projet sur l'implantation des immigrés allemands recensés à Paris au milieu du XIX^e^ siècle, il était tout à fait inutile de représenter lesdits immigrés sur une carte issue de « *Google map* » : cela n'a pas de pertinence, il a donc fallu transposer les adresses d'époque en adresses modernes puis en coordonnées cartographiques en intégrant les découpages administratifs et infrastructures d'époque, sur un fond de

carte de l'époque. Cela n'aurait pas été possible sans l'apport des Archives nationales, des archives de la Ville de Paris, des fonds de cartes de la BnF et des données du projet Alpage[22] (König *et al.*, 2023). Ces divers matériaux de recherche et archives historiques, numérisés patiemment par d'autres projets de recherche et d'archivistique, ont permis de faire avancer un nouveau projet. C'est là que les données numérisées forment un nouveau paradigme archivistique dans le cadre de la recherche en GLAM.

21 Transposer dans le monde numérique les pratiques et les exigences d'une discipline demande donc de la créativité et parfois une collaboration interdisciplinaire. D'autres outils méthodologiques sont à mobiliser pour faire dialoguer les disciplines autour de l'objet commun.

Élaborer un modèle

22 Comme nous l'avons évoqué, ce genre de projet se trouve obligatoirement à l'intersection de plusieurs disciplines, toutes aussi pointues et exigeantes les unes que les autres. Dans le projet « Bibliographies de critiques d'art francophones[23] », des chercheuses et chercheurs en histoire de l'art, en infocom et en informatique se sont retrouvés pour définir la notion naissante du critique d'art entre 1870 et 1950 en lien avec l'avènement de la presse écrite et des salons artistiques. La méthode incluait la prosopographie : la sociologie des acteurs de la critique, les lieux de production et d'édition des textes ainsi que leurs supports de publication, sans oublier l'impact du contexte historique (Gispert et Méneux, 2020 ; Kembellec, 2020). Ces diverses questions ont été l'objet d'intenses réflexions et d'un dialogue interactif entre historiens, historiens de l'art, archivistes et chercheurs en infocom. Six mois ont été nécessaires pour élaborer, corriger et reconstruire plusieurs fois le modèle conceptuel représentant les acteurs et les objets intervenant dans l'activité de critique ainsi que leurs interactions. Il a fallu confronter les points de vue, tant sur le fond (le fait) que sur la forme (le modèle) et, enfin, la manière de stocker les données, visualiser l'information et proposer des représentations de connaissances[24]. Cette étape de modélisation, que nous nommons « maïeutique de recherche » en hommage au dialogue socratique, pousse les différents interlocuteurs à interroger les autres pour amener chacun à expliciter le plus simplement et le plus clairement possible ses besoins et contraintes afin de « négocier » un dispositif d'accès aux connaissances compilées, produites et vérifiées par les historiens de l'art. Avant d'être mise en œuvre, cette étape de modélisation s'appuie sur des méthodes issues de l'informatique de gestion comme UML ou Merise[25], mais aussi sur la production de métadonnées descriptives et d'une connaissance approfondie de l'objet étudié. Il faut donc impérativement savoir s'entourer pour avancer dans un programme GLAM, sous peine d'avoir au final un dispositif inutilisable car techniquement mal pensé ou scientifiquement approximatif. Dans le cadre de ce dispositif, l'interface éditait dynamiquement des notices exportables vers Omeka, référençables par Zotero, téléchargeables en plusieurs formats et identifiables par les moteurs de recherche pour alimenter le Web de données[26].

Penser le cycle de vie de ses données

23 Une fois le projet de recherche terminé et les données exploitées, il convient (légalement) de rendre accessibles, non seulement les articles, livres et rapports associés, mais aussi les données brutes capitalisées comme nous l'avons expliqué en début de chapitre. Nous avons vu, dans le cadre du projet sur l'immigration allemande à Paris au XIX^e^ siècle, à quel point les données d'autres projets avaient été capitales pour sa bonne mise en œuvre. Le dépôt des nouvelles données peut être vu comme une contrainte, mais ce n'est pas du temps perdu, car il est aussi, à son tour, un don aux futurs chercheurs. C'est aussi l'occasion de faire connaître le projet grâce à la publication de ses données qui conduit l'usager vers les travaux associés. Du point de vue de l'archiviste, il convient, autant que possible, de documenter et de transférer la qualité de preuve historique de la source à sa version numérisée, tant par la qualité de la numérisation que par la possibilité de localiser la source originale (si elle existe encore) pour vérifier la fidélité de la copie numérique. Enfin, dans le cas d'un dispositif de consultation en ligne, la granularité des métadonnées issues du modèle peut être exposée pour être comprise à la fois par les humains (visuellement) que par les machines (moteur de recherche ou Zotero). Il s'agit là de l'une des règles de base du FAIR : l'interopérabilité. En plus d'être libres d'accès et de droit, dans des formats libres, accessibles avec des outils si possible gratuits, les données doivent être décrites avec des métadonnées explicatives. Elles peuvent aussi bien s'appliquer à l'annotation textuelle avec des liens conceptuels ou bibliographiques qu'à la description de faits historiques, de modèles artistiques au moyen d'annotations d'images. Les pionniers du domaine depuis David Shotton (Shotton, 2009) ont nommé cette méthode de valorisation des archives de données ou de textes scientifiques en ligne « *semantic publishing* ». Ce concept est repris par Tobias Kuhn (2017) et le *genuine semantic publishing*, comme vecteur du FAIR, mais aussi de sérendipité et rend possible le raisonnement conceptuel automatisé, selon des méthodes et des formats techniques également consensuels élaborés dans le cadre du Web de données.

24 Rob Sanderson, directeur du département *Cultural Heritage Metadata* de l'université de Yale, œuvre beaucoup dans la documentation des archives numériques et particulièrement les images numérisées et partagées en contexte GLAM. Il insiste sur l'importance d'assurer la pérennité et de donner accès sur Internet non seulement aux archives elles-mêmes, mais aussi aux annotations qui leurs sont appliquées. Par exemple, grâce à l'*International Image Interoperability Framework* [IIIF], son institution offre de charger à la demande tout ou partie d'une image en diverses tailles et qualités dans un article culturel ou scientifique en ligne *via* une URL paramétrable de manière standardisée. Ce protocole permet également de fournir de manière standardisée toutes les métadonnées de contextualisation disponibles dans le catalogue de la base[27].

25 On l'a compris, la documentation du projet est fondamentale à la réutilisation des fonds d'archives, car il s'agit d'un point d'entrée dans les données du projet pour les générations futures. C'est la raison pour laquelle la politique d'ouverture des données impose, comme condition au financement de la recherche publique, une démarche de plan de gestion des données ou PGD qui ne se limite pas à la conservation des données.

26 Le plan de gestion de données, PGD ou *Data Management Plan* [DMP], est un outil phare de la politique d'ouverture des données. Le décret n° 2021-1572[28] stipule à l'article 6 que

« les établissements publics et fondations reconnues d'utilité publique [...] veillent à la mise en œuvre par leur personnel de plans de gestion de données[29] [...] ».

27 Il s'agit d'un document rédigé au commencement d'un projet de recherche, qui doit être mis à jour tout au long du projet. Il est par exemple requis dans les six premiers mois d'un projet financé par l'Agence nationale de la recherche[30]. Puis une version mise à jour doit être transmise à mi-projet et la remise de la version définitive conditionne le dernier versement du financement. Dans les faits, la réflexion sur la gestion des données démarre en amont du projet, puisque la gestion des données est mentionnée dans la réponse à l'appel d'offre.

28 Un PGD type comporte (INRAe, 2021, p. 5) :

- une présentation du projet et/ou de la structure,
- les caractéristiques des données (nature, volume) et leurs modalités de production et de traitement,
- les métadonnées et les documentations qui les accompagnent,
- les modalités de stockage et de sécurisation,
- les informations légales les concernant : propriété des données, respect de l'éthique et du RGPD,
- les conditions d'accès et de partage,
- le plan d'archivage à long terme,
- les responsabilités et les budgets affectés à la gestion de ces données.

29 L'enjeu de la formation est d'importance : il s'agit de sensibiliser et de former tous les chercheurs à la rédaction des PGD et les initiatives de supports de formation à la rédaction des PGD fleurissent sur les sites des établissements de recherche. Nous ne citons que deux initiatives :

- la plateforme DoraNum[31], réalisée par l'Inist-CNRS et le GIS Réseau Urfist, propose depuis 2015 une centaine de ressources et d'outils de formation. Elle est associée depuis juillet 2022 à l'écosystème national *Recherche Data Gouv* en tant que centre de ressources supports pédagogiques et e-formation, afin de mutualiser les ressources pédagogiques issues des initiatives locales, par exemple des Ateliers de la donnée ;
- DMP OPIDor[32] est une plateforme d'aide à la rédaction de PGD. Issue de codes sources ouverts du *Digital Curation Centre* (Royaume-Uni) et de l'*University of California Curation Center* (États-Unis), personnalisés par des chercheurs français. Des dizaines d'exemples et de modèles de PGD élaborés par des établissements français y sont disponibles en téléchargement. Mais surtout, il est possible de créer son propre PGD à partir du modèle de son choix et de le compléter en ligne de façon collaborative.

30 Ces outils qui illustrent et appliquent les principes du FAIR, tout en contribuant à leur propagation, montrent la rapidité des mises en œuvre des PGD dans la recherche française[33]. Ce dynamisme de cette dernière est vertueux, mais, comme nous allons le montrer avec la problématique de la publication, il conduit à un foisonnement d'initiatives complémentaires, parfois difficilement lisibles par les chercheurs. Publier et partager ses données demande de faire des choix, d'identifier les partenaires et les modalités de publication les plus adaptées parmi les offres disponibles.

Les stratégies de publication et de partage en ligne des données de la recherche

31 Concrètement, l'ouverture des données est un acte de publication et de partage dont les modalités ne sont pas imposées, mais sont à choisir au cas par cas en fonction de l'environnement du projet, des types de données et des contraintes réglementaires ou techniques. La plateforme Datapartage[34] de l'Institut national de recherche pour l'agriculture, l'alimentation et l'environnement identifie quatre stratégies de partage de données.

Publier ses données dans un entrepôt

32 Un entrepôt de données de la recherche est une infrastructure publique ou privée qui offre, pour les jeux de données de la recherche, un service de publication, de référencement, de documentation et d'accès pérenne. Le partage des données de la recherche est une opération de communication qui répond aussi à des injonctions réglementaires et institutionnelles, il faut donc bien choisir son entrepôt[35]. Souhaite-t-on être associé à une société privée ou à un partenaire institutionnel ? Dispose-t-on d'un budget pour le faire ? Quel est le volume de données à partager ? Quelle est la pérennité de l'infrastructure ? Ces questions méritent d'être débattues dans le PGD. Or le choix d'une plateforme peut être difficile, car la mise en place des entrepôts répond à des logiques multiples. Certaines sont mises en place par des éditeurs, et les données seront parfois obligatoirement associées à un article ou un livre porté au catalogue de cet éditeur. Mais ce n'est pas toujours le cas. *Data Mendeley*, entrepôt de l'éditeur scientifique néerlandais Elsevier, par exemple, accepte des données associées ou non à ses publications. D'autres plateformes sont liées à un établissement ou une institution et réservées aux chercheurs qui y sont rattachés. Il existe aussi des entrepôts par discipline comme Nakala, plateforme d'Huma-num, qui accueille des données en SHS, tandis que le CNRS a ouvert *CNRS Research Data*, entrepôt destiné aux communautés scientifiques qui ne disposent pas encore d'entrepôt thématique.

33 L'accroissement de l'écosystème de partage de données s'accompagne de développements de logiciels de valorisation *ad hoc*. Nous donnons ici l'exemple de la plateforme POUNT[36] de l'université de Strasbourg (Witz *et al.*, 2019) dont le code est librement utilisable par d'autres universités. POUNT est conçu pour sauvegarder et versionner les jeux de données, les structurer selon le standard de la discipline, configurer les droits d'accès, visualiser les fichiers de tous types (3D, vidéos, etc.) et les enrichir par des liens ou des informations. La qualité des fonctionnalités offertes par ces logiciels peut aussi peser sur le choix.

Fournir ses données sous la forme de matériel supplémentaire à la publication

34 Les revues scientifiques en ligne en sciences, techniques et médecine [STM] ont pour habitude de publier en annexes les jeux de données qui étayent l'article, en plus de leur méthode d'exploitation (protocoles de calcul et algorithmes associés). Les publications qui respectent le plan traditionnel *Introduction, Methods, Results, and Discussion* [IMRaD], particulièrement usité en science dures, restent en relation étroite avec leurs données

qui garantissent la reproductibilité de la démonstration. Pour aller encore plus loin, certaines revues publient directement des articles dits « exécutables » sous la forme de carnets, par exemple les carnets Jupyter (Dombrowski, Gniady et Kloster, 2020), qui incluent de la programmation (des statistiques ou du calcul scientifique). Le résultat est visuellement un article traditionnel mais dont les résultats et les schémas sont calculés dynamiquement depuis les données. Cela permet aux évaluateurs et aux autres chercheurs de valider en toute transparence les résultats et de faire des tests avec d'autres paramètres (choix d'algorithme ou de variables exploitées, sous-ensembles de données étudiés) sans forcément être spécialistes, car ils ont l'accès aux données et au raisonnement.

35 Ces méthodes commencent à être intégrées par les LSHS, puisqu'il existe déjà une revue, le *Journal of Digital History*[37], qui ne publie que des carnets exécutables. Frédéric Clavert, l'un des coordinateurs de cette revue, présente deux échelles de lecture des sources historiennes : d'une part, le *close reading* au plus près du texte/*distant reading* globalisant[38] et, d'autre part, une échelle lecture humaine/lecture computationnelle (Clavert, 2014). La clé de la lecture et de l'interprétation des sources de l'histoire à l'ère numérique réside dans les allers-retours constants entre *close reading* et *distant reading* et entre appréhension humaine et appréhension computationnelle des sources primaires. Cette approche dite distante permet de prétraiter un plus grand volume de données, plus rapidement et, ensuite, de mettre les résultats intermédiaires à disposition des chercheurs pour une analyse plus qualitative. Les travaux issus de ces nouvelles méthodes de travail en LSHS se doivent de fournir les données étudiées comme matériau complémentaire à l'écrit scientifique.

Publier ses données dans un *Data Paper* (article de données)

36 Les données se doivent d'être publiées, mais il est évident que les données seules, même avec leurs métadonnées, forment un ensemble trop aride pour être exploitable en l'état. Pour utiliser une métaphore triviale, c'est un peu comme proposer un meuble suédois en kit sans y adjoindre de notice d'utilisation : c'est peu utilisable. C'est là qu'intervient le concept de *data paper*. Un *data paper* est un article scientifique, généralement assez court, qui présente un ou plusieurs jeux de données et explique brièvement les objectifs du projet de recherche associé, explicite les méthodes de collecte, de numérisation – voire de calcul –, qui ont amené à la production du jeu de données publié dans l'entrepôt. Il faut bien le distinguer des articles de résultats scientifiques : le *data paper* n'est là que pour documenter les données du corpus. Ce document est indispensable à la bonne compréhension des données, des règles de nommage et apporte tous les éléments de contextualisation utiles. Comme le *data paper* est un article scientifique, il peut être cité et être pris en compte dans l'évaluation de la production scientifique de son auteur[39]. Bien que ce type d'article vienne initialement des sciences dures, il pénètre les LSHS et particulièrement les humanités numériques[40].

37 Un *data paper* en sciences humaines et sociales peut présenter le plan suivant :

- contexte et résumé : succincte description de données produites, leur contexte scientifique ainsi que leurs utilisations potentielles ;
- méthodes : description précise du processus de production des données afin que celui-ci soit reproductible ;

- fichiers de données : description de chaque jeu de données associé avec le *data paper* (variables, noms de fichiers, localisation, formats et taille) ;
- validité des données décrites : analyses ou procédures ayant permis de confirmer la validité des données décrites (confrontation avec différentes sources ou avec des données comparables) ;
- notes d'usage : procédures de réutilisation des données, licence ;
- disponibilité du code (éventuellement) : reproductibilité, un éventuel accès au code de reproduction du jeu de données.

Publier dans le Web des données

38 Outre la publication des données, il est courant en humanités numériques, et plus spécifiquement dans les GLAM, que les équipes de recherche souhaitent partager des fac-similés numériques des documents patrimoniaux (ouvrages, documents administratifs, œuvres d'art, cartes) participant au corpus. En effet, ces documents numérisés et mis à disposition sur des dispositifs de consultation en ligne sont des sources précieuses pour les futurs projets de recherche.

39 Le Web de données repose sur le fait que des moteurs de recherche, des systèmes d'affichage ou de requête du web s'appuient sur des fichiers contenant des métadonnées structurées selon une norme de modélisation (RDF) et souvent non visibles à l'écran, car intégrées au code source du document. Cela permet de filtrer, regrouper, relier des fichiers du Web selon le sens de leurs contenus. Par exemple, l'accès à une page d'ouvrage rare numérisée pourra être réalisé au moyen de filtres de recherche tels que les lieux ou les entités nommées désambiguïsées qui y sont évoqués, le type de police d'écriture, ou même sur la présence de *marginalia* ou d'enluminures. Ces informations, issues du travail d'enrichissement effectué par les chercheurs, sont les métadonnées figurant dans les fichiers de données structurées que les techniques de documentation du Web des données vont ainsi rendre accessibles aux outils de collecte et de filtrage.

40 Ces quatre procédés n'aboutissent pas au même degré d'ouverture et ne visent pas les mêmes cibles. La « démocratisation de l'accès aux savoirs, utile à l'enseignement, à la formation, à l'économie, aux politiques publiques, aux citoyens et à la société dans son ensemble[41] », préconisée par le gouvernement, passera peut-être davantage par le Web de données que par l'accès aux *data papers* ou aux entrepôts, plus adaptés aux partages entre chercheurs. Ils requièrent aussi des degrés de technicité différents. Un chercheur peut publier un *data paper* en autonomie, les entrepôts proposent des interfaces appropriables par les chercheurs ou les archivistes, mais l'accès au Web de données demande l'aide d'un ingénieur de recherche ou d'un *data librarian*. L'interdisciplinarité reste un atout dans cette étape vers l'ouverture.

Conclusion

41 Nous avons livré dans ce texte un état des lieux, à la fin de 2023, de la politique de FAIRification des données de la recherche publique française qui concerne autant les sciences dures que le domaine des LSHS et plus particulièrement des humanités numériques dont les méthodes s'appuient sur des corpus et des traitements numériques. Pour atteindre l'objectif de partage, de conservation pérenne et de

réemploi, les modèles de données doivent être clairement documentés et doivent, le plus possible, respecter des normes ouvertes de structuration. Cela requiert la collaboration de plusieurs profils d'intervenants : informaticiens, professionnels de l'information ou des archives, et spécialistes de la discipline. Même si l'époque actuelle est davantage préoccupée par la mise en place des instances, outils et formations, les enjeux de la collaboration ne doivent pas être négligés, car des collègues dont les statuts, missions et cultures professionnelles diffèrent doivent parvenir à harmoniser leurs langages, leurs méthodes et leurs objectifs au sein de l'équipe du projet. Si ce processus peut sembler rébarbatif ou chronophage, il peut néanmoins être gratifiant, car la publication des jeux de données sur des plateformes ouvertes, liée aux profils des contributeurs, permet aux acteurs des supports à la recherche de valoriser leur travail et de faire reconnaître leur expertise. Le rôle de la recherche ouverte ne se limite pas à la valorisation des données, elle promeut aussi les savoir-faire scientifiques.

BIBLIOGRAPHIE

Baromètre français de la Science Ouverte (décembre 2022), consulté le 3 août 2023, à l'adresse : https://barometredelascienceouverte.esr.gouv.fr/

Zoé ANCION, Francis ANDRE, Francis CADOREL, Romain FERET, Odile HOLOGNE, Kenneth MAUSSANG, Marine MOGUEN-TOURSEL et Véronique STOLL, « *Plan de gestion de données – Recommandations à l'ANR*, 2019, Ministère de l'enseignement supérieur et de la recherche », https://doi.org/10.52949/7

Pierre BOURDIEU, *Esquisse d'une théorie de la pratique*, Genève, Droz, 1972, 10.3917/droz.bourd.1972.01

Pierre BOURDIEU, « La production de la croyance », *Actes de la recherche en sciences sociales*, 1977, 13 (1), p. 3-43.

Suzanne BRIET, *Qu'est-ce que la documentation ?,* Paris, EDIT, 1951.

Frédéric CLAVERT, « Vers de nouveaux modes de lecture des sources », dans Olivier LE DEUFF (dir.), *Le temps des humanités digitales*, Limoges, Fyp Éditions, 2014.

Quinn DOMBROWSKI, Tassie GNIADY et David KLOSTER, « Introduction aux carnets Jupyter », traduction par François Dominic Laramée, *Programming Historian en français* 2, 2020, https://doi.org/10.46430/phfr0014

Marie GISPERT et Catherine MÉNEUX, « Bibliographies de critiques d'art francophones », *Cahiers Octave Mirbeau*, 2020, 27, p. 315-320.

James HENDLER et Tim BERNERS-LEE, « From the Semantic Web to social machines: A research challenge for AI on the World Wide Web », *Artificial intelligence*, 2010, 174 (2), p. 156-161.

INIST, « Le plan de gestion des données », dans *Une introduction à la gestion et au partage des données de la recherche,* (s.d.), consulté le 5 août 2023, à l'adresse https://www.inist.fr/wp-content/uploads/donnees/co/module_Donnees_recherche_26.html

INRAe, « Rédiger un plan de gestion de données (PGD) », OpenClass, 2021, https://ist.inrae.fr/wp-content/uploads/sites/21/2021/11/OpenClass-PGD-Octobre2021.pdf, consulté le 06/08/2023.
Gérald KEMBELLEC et Thomas BOTTINI, « Réflexions sur le fragment dans les pratiques scientifiques en ligne : entre matérialité documentaire et péricope », *20e Colloque International sur le Document Numérique : CiDE.20*, novembre 2017, Villeurbanne, France.
Gérald KEMBELLEC, « Dialogie disciplinaire en Humanités Numériques : vers une percolation épistémique et méthodologique négociée. Le cas de l'analyse des acteurs de la critique d'art (1850-1950) », *Sens public*, 2020, p. 1-31. https://doi.org/10.7202/1079443ar
Gérald KEMBELLEC et Olivier LE DEUFF, « Poétique et ingénierie des data papers », *Revue française des sciences de l'information et de la communication*, 2022, 24, ⟨10.4000/rfsic.12938⟩ ou ⟨hal-03850522⟩
Mareike KÖNIG, Gérald KEMBELLEC et Evan VIREVIALLE, « Data paper en humanités numériques : Adressbuch 1854 » 2023, preprint, à paraître dans *Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux*, inPress. https://hal.science/hal-03947294/
G. KUCK, « Tim Berners-Lee's Semantic Web », *South African Journal of information management*, 2004, 6(1).
Tobias KUHN et Michel DUMONTIER, « Genuine semantic publishing », *Data Science*, 1(1-2), 2017, p. 139-154.
Bruno LATOUR et Steve WOOLGAR, « *La vie de laboratoire : la production des faits scientifiques* », Paris, La Découverte, 1986.
Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, « *Deuxième Plan national pour la science ouverte 2021-2024* », consulté le 12 novembre 2023, à l'adresse : https://www.enseignementsup-recherche.gouv.fr/sites/default/files/content_migration/document/2e-plan-national-pour-la-science-ouverte-2021-2024-7794.pdf
Ministère de l'Enseignement supérieur et de la Recherche et Université de Lille, « *Science ouverte. Codes et logiciels* », 2022, https://www.ouvrirlascience.fr/wp-content/uploads/2022/10/Passeport_Codes-et-logiciels_WEB.pdf
Kumiyo NAKAKOJI, Yasuhiro YAMAMOTO, Mina AKAISHI et Koichi HORI, « Interaction design for scholarly writing: hypertext representations as a means for creative knowledge work », *The New Review of Hypermedia and Multimedia*, Special issue: Scholarly hypermedia, 2015, vol. 11, nº 1, Taylor et Francis.
Carole L. PALMER, Lauren C. TEFFEAU et Carrie M. PIRMANN, « Scholarly Information Practices in the Online Environment – Themes from the Literature and Implications for Library Service Development », rapport, OCLC Research, 2009.
Silvio PERONI, Francesco OSBORNE, Angelo DI IORIO, Andrea Giovanni NUZZOLESE, Francesco POGGI, Fabio VITALI et Enrico MOTTA, « Research Articles in Simplified HTML: a Web-first format for HTML-based scholarly articles », *PeerJ Computer Science*, 2017, https://doi.org/10.7717/peerj-cs.132
David SHOTTON, « Semantic publishing: the coming revolution in scientific journal publishing », *Learned Publishing*, 2009, vol. 22, nº 2, p. 85-94.
John UNSWORTH, « Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this? », *Symposium on Humanities Computing: formal methods, experimental practice*, sponsored by King's College, London, 2000 May 1st.

Régis WITZ, Julia SESÉ, Ana SCHWARTZ, Stéphanie CHEVIRON et Vincent LUCAS, « Science Ouverte : sauvegarder, visualiser et partager vos données », *Jres, Journées réseaux de l'enseignement et de la recherche*, Dijon, 16 décembre 2019, https://conf-ng.jres.org/2019/document_revision_5259.html?download, consulté le 8 août 2023.

NOTES DE BAS DE PAGE

1. https://www.enseignementsup-recherche.gouv.fr/fr/le-plan-national-pour-la-science-ouverte-2021-2024-vers-une-generalisation-de-la-science-ouverte-en-48525

2. 67 % des publications scientifiques françaises parues en 2021 étaient en accès ouvert en décembre 2022 (source : *Baromètre français de la Science Ouverte* (décembre 2022), consulté le 3 août 2023, à l'adresse : https://barometredelascienceouverte.esr.gouv.fr/).

3. Acronyme anglais pour *Galleries, Libraries, Archives and Museums* (en français galeries, bibliothèques, archives et musées).

4. Définition issue du Code de la recherche, article L533-4.

5. Loi n ° 2016-1321 du 7 octobre 2016 pour une république numérique, https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746

6. En 2023, il s'agit de Claire Giry.

7. https://recherche.data.gouv.fr/fr/page/ateliers-de-la-donnee-des-services-generalistes-sur-tout-le-territoire

8. https://recherche.data.gouv.fr/fr

9. https://www.softwareheritage.org/

10. Une introduction à la gestion et au partage des données de la recherche, https://www.inist.fr/wp-content/uploads/donnees/co/module_Donnees_recherche_7.html (consulté le 5 août 2023).

11. Citons par exemple le projet arXiv qui concerne une archive ouverte de prépublications dans les domaines des mathématiques, physique, informatique, économie, et le projet *Software Heritag*e, archive ouverte qui collecte, préserve et partage les codes sources de tous les logiciels publiquement disponibles.

12. Voir l'article de synthèse sur le sujet par Kembellec et Bottini (2017).

13. Le « père » du Web.

14. Stylo est un éditeur de textes scientifiques, conçu par l'équipe de la chaire de recherche du Canada sur les écritures numériques avec le soutien d'Érudit et Huma-Num. Les articles produits dans Stylo sont structurés, peuvent être annotés collaborativement et exportés en différents formats selon les plateformes ou processus de publication visés. Pour en savoir plus : https://apropos.erudit.org/stylo-un-outil-dedition-numerique-innovant-adapte-a-la-publication-savante/

15. Les structures de classement et l'indexation des corpus en ligne sont des transpositions conceptuelles de l'archivistique « physique » : plans de classement, règles de nommage des documents, normes d'encodage, accès pérennes, organisation des stockages (ici de fichiers), contextualisation et description à l'aide de référentiels.

16. En 1951, Suzanne Briet pose le principe que tout objet ou même être vivant peut devenir document s'il est intégré à un système de connaissance matérialisé par des outils d'inventaire ou de description : « Une étoile est-elle un document ? Un galet roulé par un torrent est-il un document ? Un animal vivant est-il un document ? Non. Mais sont des documents les photographies et les catalogues d'étoiles, les pierres d'un musée de minéralogie, les animaux catalogués et exposés dans un zoo » (Briet, 1951, p. 7).

17. Lettres et sciences humaines et sociales.
18. Centre Roy Rosenzweig pour l'Histoire et les Nouveaux Médias, voir https://rrchnm.org/our-story
19. Voir le projet portail éponyme qui inventorie une partie des textes et livres écrits, traduits, enluminés, collectionnés ou inventoriés de l'Antiquité au XVIII^e siècle : https://portail.biblissima.fr
20. Voir le projet *Glossae Scripturae Sacrae-electronicae* qui édite plus de 320 000 sentences exégétiques associées au texte de la Bible, encodées au format XML/TEI pour en permettre le filtrage par différents critères : https://gloss-e.irht.cnrs.fr/
21. Selon Marie-Anne Chabin : « La diplomatique est l'étude de l'authenticité et de la fiabilité des actes écrits au travers de leur processus d'élaboration, de leur forme (support et format, mais aussi structure et mise en page), de leur diffusion » (Marie-Anne Chabin « Diplomatique », blog *Esprit critique et grain de sel*, [s.d.], https://www.marieannechabin.fr/diplomatique/).
22. Voir projet ALPAGE : AnaLyse diachronique de l'espace urbain Parisien : approche GEomatique – Alpage (huma-num.fr), https://alpage.huma-num.fr/
23. Voir le dispositif de consultation lié au projet : https://critiquesdart.univ-paris1.fr/
24. Les éléments méthodologiques présentés dans cette partie sont résumés de manière pragmatique dans une ressource tutorielle interactive du projet Données de la recherche apprentissage numérique (Doranum) : 10.13143/7a03-1j03.
25. *Unified Modeling Language* et Merise sont deux systèmes conventionnels graphiques, utilisés pour représenter les processus et les objets que l'on intégrera à un développement informatique.
26. Le lecteur curieux pourra retrouver en bibliographie les deux exemples de projets présentés dans cette partie, « *Adressbuch der Deutschen in Paris von 1854* » (König *et al.*, 2023) et « Bibliographies de critiques d'art francophones » (Kembellec, 2020).
27. Voir le site du consortium du standard IIIF (https://iiif.io/) et les collections numérisées de la *Yale University Library* (https://collections.library.yale.edu/).
28. Décret n° 2021-1572 du 3 décembre 2021 relatif au respect des exigences de l'intégrité scientifique par les établissements publics contribuant au service public de la recherche et les fondations reconnues d'utilité publique ayant pour activité principale la recherche publique.
29. Texte complet de l'article 6 : « Les établissements publics et fondations reconnues d'utilité publique mentionnés au troisième alinéa de l'article L. 211-2 du code de la recherche définissent une politique de conservation, de communication et de réutilisation des résultats bruts des travaux scientifiques menés en son sein. À cet effet, ils veillent à la mise en œuvre par leur personnel de plans de gestion de données et contribuent aux infrastructures qui permettent la conservation, la communication et la réutilisation des données et des codes sources. » Décret n° 2021-1572 du 3 décembre 2021 relatif au respect des exigences de l'intégrité scientifique par les établissements publics contribuant au service public de la recherche et les fondations reconnues d'utilité publique ayant pour activité principale la recherche publique.
30. https://anr.fr/fr/lanr/engagements/faq-pgd/ consulté le 6 août 2023.
31. https://doranum.fr/
32. https://dmp.opidor.fr/
33. On peut consulter dans cet ouvrage le texte d'Annick Boissel et Véronique Ginouvès qui retrace les étapes d'élaboration d'un PGD autour du fonds de l'anthropologue Jean-Pierre Olivier de Sardan à la Maison méditerranéenne des sciences de l'homme.
34. https://datapartage.inrae.fr/Partager-Publier
35. Voici quelques exemples d'entrepôts de recherche gratuits ou *freemium* disponibles au moment de l'écriture de ce texte (consultés le 16 novembre 2023) :
– Mendeley Data (Elsevier), https://data.mendeley.com/, 10 GB par dataset (tous les fichiers) ;
– Harvard Dataverse, https://dataverse.harvard.edu/, 2 GB par fichier ;
– Open Science Framework, https://osf.io/, 5 GB par fichier ;

– Zenodo (CERN, OpenAIRE), https://zenodo.org/, 50 GB par fichier ;
– Science Data Bank, https://www.scidb.cn/, 8 GB par fichier.

36. Plateforme OUverte Numérique Transdisciplinaire [POUT] : https://pount.unistra.fr/

37. https://journalofdigitalhistory.org

38. Les notions de *close* et *distant reading* sont empruntées à Franco Moretti pour distinguer, sans les opposer, la lecture attentive de l'humain et l'analyse computationnelle qui peut faire émerger de nouvelles hypothèses que le spécialiste humain sera à même d'interpréter. Ces notions sont donc complémentaires.

39. Voici quelques exemples de revues internationales publiant des *data papers : Data in Brief,* revue en libre accès, coéditée par ScienceDirect et Elsevier, spécialisée dans les *data papers* dans toutes les disciplines ; *Scientific data,* revue en libre accès éditée par Nature Publishing Group, publie des *data papers* dans toutes les disciplines ; *Harvard Data Science Review,* multidisciplinaire, favorise le dialogue entre les chercheurs, les formateurs et les praticiens des données.

40. La *Revue française des sciences de l'information et de la communication* a consacré en 2022 un dossier spécial à la méthode des *data paper* (Kembellec et Le Deuff, 2022).

41. Présentation du second Plan national pour la science ouverte : https://www.enseignementsup-recherche.gouv.fr/fr/le-plan-national-pour-la-science-ouverte-2021-2024-vers-une-generalisation-de-la-science-ouverte-en-48525 (consulté le 20 novembre 2023).

AUTEURS

Gérald Kembellec
Maître de conférences en sciences de l'information, chercheur au Dicen-IDF Cnam/ Paris (EA 7339)

Claire Scopsi
Maître de conférences en sciences de l'information, chercheuse au Dicen-IDF, Cnam/ Paris (EA 7339)