# Les enjeux de la curation des jeux de données aujourd'hui : l'exemple de la BnF

Aude Eychenne

ÉCOLE NATIONALE DES CHARTES

# Les enjeux de la curation des jeux de données aujourd'hui : l'exemple de la BnF

Aude Eychenne
Diplômée de Maîtrise de Lettres Modernes

Mémoire pour le diplôme de master
« Technologies numériques appliquées à l'histoire »
2023

# Résumé

Ce mémoire, au cœur des préoccupations de la science ouverte qui incitent toujours plus à l'ouverture des données de la recherche, s'intéresse à la publication des jeux de données. Depuis les années 2000, le libre accès aux données occupe une place stratégique dans les politiques de recherche. Ces enjeux ont été relayés par des professions supports, notamment par les bibliothécaires, via des services dédiés destinés à accompagner les chercheurs dans l'application des recommandations de gestion de leurs données.

Dans ce contexte, la BnF, forte d'une transformation numérique de ses services, de ses métiers et de son organisation à l'œuvre depuis vingt ans, se réinterroge aujourd'hui, avec une production de données désormais évaluée à 6 petaoctets, sur la stratégie de publication des jeux de données qu'elle a mise en place cinq ans auparavant en les diffusant sur le site api.bnf. Le BnF DataLab depuis 2021, porte, avec sa mission d'appui aux activités scientifiques liées aux collections de la bibliothèque, cette réflexion, en lien avec d'autres départements producteurs de jeux de données.

Les dernières avancées des politiques publiques en matière de science ouverte, le nouveau paysage des entrepôts de données, les techniques de curation qu'ils font émerger, ont fortement évolué ces dernières années, en particulier récemment avec la création de l'initiative Recherche Data Gouv (RDG).

Une réflexion fondée sur l'observation de ces changements fait l'objet de ce présent mémoire, en faisant état des limites de l'exposition actuelle des jeux de données de la BnF. Afin de proposer une solution méthodologique et technique de référencement consolidée et élargie, un processus de préparation et de publication d'un jeu de données a été mis en place. Cette expérimentation permet d'appréhender les moyens de curation requis pour rendre les données diffusées réutilisables. Elle met également en évidence la réponse aux besoins d'exposition actuels des données qu'offre, sur de nombreux points, la publication dans un entrepôt de données de la recherche fédérateur.

**Mots-clés**

Science ouverte, open data, données de la recherche, entrepôt de données, jeu de données, métadonnées, curation, identifiant pérenne, moissonnage.

**Informations bibliographiques** : Aude Eychenne, *La curation des jeux de données aujourd'hui : l'exemple de la BnF,* mémoire de master « Technologies numériques appliquées à l'histoire », dir. Emmanuelle Bermès, École nationale des chartes, 2023.

# Remerciements

Je tiens à remercier tout particulièrement Emmanuelle Bermès, et lui témoigner ma reconnaissance pour l'énergie si positive avec laquelle elle m'a accompagnée à découvrir, dans le monde des données patrimoniales et de la recherche, leurs couches et sous-couches numériques. La vision claire et très documentée qu'elle a de ces objets et des enjeux qui les portent m'a ouvert la voie. Ses conseils et relectures, sa disponibilité généreuse, ont constitué un précieux soutien. Nos discussions, qu'elles aient porté sur le cadrage du sujet, les méthodes de restitution de mes investigations, ou sur certains épisodes de l'aventure des bibliothèques, ont été stimulantes, et nos rencontres chaque fois m'ont donné de l'élan pour poursuivre.

Mes plus sincères remerciements vont à Marie Carlin et Jean-Philippe Moreux, qui se sont rendus disponibles pour encadrer mon stage, me guider à chaque étape de mes missions, m'ouvrir les bonnes portes à la BnF, et finalement valoriser mon travail dans un moment de restitution dédié qu'ils ont organisé. Ils vont également aux nombreux collègues des différents services de la BnF rencontrés, aussi bien qu'aux personnes des équipes de recherche ou des entrepôts de données qui ont bien voulu me consacrer du temps pour partager leur expérience. Je salue ici en particulier, parmi tous mes voisins et voisines de bureau si accueillants du Service de l'accompagnement à la recherche, Géraldine Camile et Slimane Tounsi pour la gentillesse, l'accompagnement, les encouragements avec lesquels ils ont soutenu mon travail au quotidien.

Merci à mes camarades du M2 TNAH, en particulier Thomas, Marion, Giorgia. Ils ont été une très bonne compagnie pour les heures d'école.

Enfin, je dois beaucoup à mes proches qui m'ont chaque fois rappelé, quand je le perdais, le sens que je mettais dans cette démarche d'année d'études, et qui m'ont parfois écouté sans toujours bien comprendre, quand je m'enthousiasmais. Merci enfin à toi, Emmanuel, pour l'humour que tu as opposé à mon désarroi, et pour tes coups de main salutaires.

# Bibliographie

## Ouverture des données de la recherche


BERKOWITZ (Héloïse) et DELACOUR (Hélène), « Ouvrir les données de la recherche : Quelles implications pour les sciences sociales ? », *M@n@gement*, vol. 25 n° 4 (2022), p. 1-31. En ligne : https://www.cairn.info/revue-management-2022-4-page-1.htm

BERMÈS (Emmanuelle), « Interopérabilité des données culturelles et patrimoniales: le FAIR sans en avoir l'air », *Culture et recherche*, vol. 144 (no printemps-été 2023). En ligne : https://u-picardie.hal.science/public/Culture_et_Recherche_144_La_science_ouverte.pdf

Maurel (Lionel), « Quel statut pour les données de la recherche après la loi numérique ? », *S.I.Lex*, 3 novembre 2016. En ligne : https://scinfolex.com/2016/11/03/quel-statut-pour-les-donnees-de-la-recherche-apres-la-loi-numerique/. Consulté le 15 mai 2023.

« Comprendre les enjeux de l'ouverture des données publiques culturelles », Ministère de la Culture, Datactivist. Novembre 2023. En ligne : https://datactivist.coop/ministere-culture/jour1.html#1

DUPRAT (Julie), *Les données de la recherche à l'Université Bordeaux Montaigne*, rapport, Université Bordeaux Montaigne, 2019. En ligne : https://hal.science/hal-02020141. Consulté le 12 juin 2023.

HADROSSEK (Christine), JANIK (Joanna), LIBES (Maurice), LOUVET (Violaine), QUIDOZ (Marie-Claude), RIVET (Alain) et ROMIER (Geneviève), « *Guide de bonnes pratiques sur la gestion des données de la Recherche* ». En ligne : https://hal.science/hal-03152732. Consulté le 6 août 2023.

JACQUEMIN (Bernard), SCHÖPFEL (Joachim) et FABRE (Renaud), « Libre accès et données de recherche. De l'utopie à l'idéal réaliste », *Études de communication. langages, information, médiations*, n° 52 (1 juin 2019), p. 11-26.

REBOUILLAT (Violaine), *Ouverture des données de la recherche : de la vision politique aux pratiques des chercheurs*, thèse de doct., Conservatoire national

des arts et métiers - CNAM, 2019. En ligne : https://theses.hal.science/tel-02447653.

WILKINSON (Mark D.) et al, « The FAIR Guiding Principles for scientific data management and stewardship », *Scientific Data*, vol. 3 nº 1 (15 mars 2016).

## Science ouverte

CARACO (Alain), « Open access et bibliothèques », *Arabesques*, nº 93 (1 avril 2019), p. 6-7. En ligne : http://publications-prairial.fr/arabesques/index.php?id=543. Consulté le 5 août 2023.

DELHAYE (Marlene), *Science ouverte : qu'est-ce qui change avec Horizon Europe ?*, 5 février 2021. En ligne : https://oaamu.hypotheses.org/2722. Consulté le 27 juillet 2023.

MOALIC (Anthony), LEHOUX (Élise), PION (Christophe) et LASNE (Christophe), « La science ouverte à l'épreuve de la sobriété », *Arabesques*, nº 109 (1 avril 2023), p. 12-14. En ligne : http://publications-prairial.fr/arabesques/index.php?id=3418

*Guide d'application de la Loi pour une République numérique pour les données de la recherche, Ministère de l'enseignement supérieur et de la recherche*, 2022. En ligne : https://hal-lara.archives-ouvertes.fr/hal-03968218. Consulté le 16 mai 2023.

## Services à la recherche

BARTHELEMY (Antoine), BAUDRY (Julien), BRAUD (Aurélia), CHARAZAC (Christelle) et GALOT (Delphine), « Open access en bibliothèque universitaire : de nouveaux enjeux de médiations », *Revue française des sciences de l'information et de la communication*, nº 8 (1 janvier 2016). En ligne : https://journals.openedition.org/rfsic/1854. Consulté le 20 juillet 2023.

CARLIN (Marie) et LABORDERIE (Arnaud), « La BnF et les services à la recherche à l'heure des humanités numériques », *Arabesques*, nº 105 (1 avril 2022), p. 8-9.

CARLIN (Marie) et LABORDERIE (Arnaud), « Le BnF DataLab, un service aux chercheurs en humanités numériques », *Humanités numériques*, vol. 4 (1 décembre 2021). En ligne : https://hal-bnf.archives-ouvertes.fr/hal-03285816.

CORMIER (Paul), *Le positionnement des bibliothèques universitaires et de recherche françaises dans les politiques publiques des données de la recherche*, (2022), mémoire de master en Sciences de l'information et de la communication. En ligne : https://memsic.ccsd.cnrs.fr/mem_03940727v1. Consulté le 9 août 2023.

DARDENNE (Nadine), « DARIAH : une infrastructure numérique au service des sciences humaines et sociales », *Blog Hypothèses*, 6 février 2014. En ligne : https://humanum.hypotheses.org/155. Consulté le 12 avril 2023.

GEROUDET (Madeleine), « La science ouverte, nouvelles pratiques numériques, nouvelles compétences », *Le numérique universitaire des BU*, n° 20 (avril 2022), p. 8-9.

LARROUSSE (Nicolas) et JOUGUET (Hélène), « Huma-Num : une infrastructure au service des sciences humaines et sociales », *Arabesques*, n° 105 (1 avril 2022), p. 6-7. En ligne :
https://publications-prairial.fr/arabesques/index.php?id=2876

LETROUIT (Carole), *La place des bibliothèques universitaires dans le développement de la science ouverte*, rapport, 2021. En ligne : https://www.enssib.fr/bibliotheque-numerique/notices/69936-la-place-des-bibliotheques-universitaires-dans-le-developpement-de-la-science-ouverte. Consulté le 6 août 2023.

MOIRAGHI (Eleonora), *Le projet Corpus et ses publics potentiels.*, rapport, Bibliothèque nationale de France, 2018. En ligne : https://hal-bnf.archives-ouvertes.fr/hal-01739730. Consulté le 11 avril 2023.

## Entrepôts de données

BENKHALID (Mohammed), « La plate-forme Nakala refait peau neuve », *Blog Arkeogis*, 13 avril 2021.
En ligne : https://arkeogis.org/la-plate-forme-nakala-refait-peau-neuve/. Consulté le 12 avril 2023

« Recherche Data Gouv, la plateforme nationale fédérée des données de recherche – Le nouveau portail de management des données de la recherche », *Blog*

*Hypothèses*, (2022). En ligne : https://dlis.hypotheses.org/6101. Consulté le 17 mai 2023.

GUIRLET, (Marielle), « Guide décisionnel et vade-mecum pour la mise à disposition d'un dépôt de données de recherche ouvertes en Suisse », (18 décembre 2020). En ligne : 10.5281/zenodo.4357133

PAIN (Marilou), *Les données de la recherche et leurs entrepôts, de la documentation à la réutilisation : étude de cas pour l'archive HAL*, Enssib, 2016. En ligne : https://memsic.ccsd.cnrs.fr/mem_01374509.

PROST (Hélène) et SCHÖPFEL (Joachim), « Les entrepôts de données en sciences de l'information et de la communication (SIC). Une étude empirique », *Études de communication. langages, information, médiations*, n° 52 (1 juin 2019), p. 71-98.

## Curation de données

*An Introduction to Humanities Data Curation*. En ligne : https://guide.dhcuration.org/contents/intro/. Consulté le 6 août 2023.

« Traduction et adaptation du modèle de description des données « *Datasheet for Datasets* », 13 mars 2019. *Blog #teamopendata,* En ligne : https://teamopendata.org/t/traduction-et-adaptation-du-modele-de-description-des-donnees-datasheet-for-datasets/1400. Consulté le 16 mai 2023.

CANDELA (Gustavo) et al, *A Checklist to Publish Collections as Data in GLAM Institutions*. En ligne : https://doi.org/10.48550/arXiv.2304.02603

GEBRU (Timnit), MORGENSTERN (Jamie), VECCHIONE (Briana), et al., *Datasheets for Datasets*. En ligne : http://arxiv.org/abs/1803.09010.

JOHNSTON (Lisa R), CARLSON (Jacob), HUDSON-VITALE (Cynthia), et al., « How Important is Data Curation? Gaps and Opportunities for Academic Libraries », *Journal of Librarianship and Scholarly Communication*, vol. 6 n° 1 (26 avril 2018). En ligne : https://www.iastatedigitalpress.com/jlsc/article/id/12803/. Consulté le 23 juin 2023.

RICHARD (Vincent), *Métadonnées pour la science ouverte : rôle et action des bibliothèques et des professionnels de l'information scientifique et technique, Zenodo*, mémoire d'étude, 2021. En ligne : https://zenodo.org/record/4662580.

*Gestion d'identifiants pérennes » — Documentation RéférentielSAEM 1.0.0*. En ligne : https://cubicweb-saem-ref.readthedocs.io/fr/latest/ark.html. Consulté le 20 juillet 2023.

## Les données patrimoniales de la BnF et la Recherche

*La recherche à la BnF : document de synthèse – 2023*, 2022. En ligne : https://www.bnf.fr/fr/actualites/la-recherche-la-bnf-document-de-synthese-2023. Consulté le 30 mai 2023.

LABORDERIE (Arnaud) et BASTARD (Irène), « La découvrabilité des collections numériques patrimoniales sous l'angle des usages de Gallica », *Bulletin des Bibliothèques de France*, (juin 2023). En ligne : https://hal.science/hal-04143824. Consulté le 9 août 2023.

BERMÈS (Emmanuelle), *Le numérique en bibliothèque : naissance d'un patrimoine : l'exemple de la Bibliothèque nationale de France (1997-2019)*, thèse de doct., Paris, Ecole nationale des chartes, 2020. En ligne : https://theses.hal.science/tel-02475991.

BERMÈS (Emmanuelle), « BnF : des métadonnées au service de projets de recherche innovants », *Arabesques*, n° 95 (1 octobre 2019), p. 8-9. En ligne : http://publications-prairial.fr/arabesques/index.php?id=1302

ILLIEN, (Gildas), « Le web sémantique, nouveau levier de la valeur pour les services d'information ? » I2D - Information, données & documents, 52, (2015). 59-60. En ligne : https://doi.org/10.3917/i2d.154.0059

ILLIEN, (Gildas) dans *Des catalogues de bibliothèque dans le Linked Open Data : servez-vous, la BnF ouvre ses métadonnées*. Le ministère de la Culture et de la Communication en partenariat avec Inria, 27 janvier 2014, 26 mn.
En ligne : https://www.dailymotion.com/video/x1bqoku

JACQUOT (Olivier), « Api et jeux de données de la BnF : découvrez et utilisez les données de la BnF », 23 novembre 2017. En ligne : https://bnf.hypotheses.org/2205. Consulté le 5 mai 2023.

LABORDERIE (Arnaud) et TFIBEL (Florence), « Ouvrir les données de la Bibliothèque nationale de France pour la recherche », *Culture et recherche*, n° 144 (1 juin 2023). En ligne : https://bnf.hal.science/hal-04074665.

# Introduction

Le numérique transforme nos sociétés mais aussi la production de la connaissance scientifique. Par la multitude de données qu'elle génère, la transition numérique contribue à élargir les connaissances, ouvrir les données et valoriser les résultats de la recherche scientifique. En parallèle, elle offre des instruments d'investigation favorisant de nouveaux modes de production des savoirs à travers la modélisation, la visualisation et l'exploration interactive des corpus. À l'heure de la science ouverte, chercheurs d'un côté et professionnels de l'information scientifique et technique (IST)[1] de l'autre, modifient leurs pratiques en profondeur. Les chercheurs sont désormais tenus de partager leurs données et de les rendre accessibles à la communauté scientifique en les publiant sur les plateformes numériques spécialisées que sont les entrepôts de données. Les professionnels de l'IST, eux, font face à de nombreux défis dans la gestion de données des productions académiques.

Aussi, la Bibliothèque nationale de France (BnF), bibliothèque patrimoniale de premier plan engagée dans la recherche, investigue-t-elle sur la réutilisation par les acteurs de la recherche de ses collections numériques. Cela se décline notamment dans l'objectif de faire connaître les jeux de données primaires ou dérivés produits à partir de ses collections. Aujourd'hui, la réponse à cet enjeu est essentiellement portée par le site api.bnf.fr[2], qui agrège de la documentation technique sur les collections numériques, les API et les jeux de données de la BnF. En structurant l'ensemble de son offre de diffusion de données, le portail a constitué un premier pas dans la simplification de l'accès aux données afin de susciter de nouveaux usages comme l'alimentation de catalogues, la création d'applications innovantes, et la fouille de données, auprès de ses publics. Aujourd'hui, après cinq ans d'existence, la question émerge : le site api.bnf répond-il aux besoins de réutilisation qu'il vise ?

Le BnF DataLab[3], au sein du département de la découverte des collections et de l'accompagnement à la recherche, doté d'une mission d'accompagnement des

[1] Liste des sigles. Annexe 7, p. 103.
[2] Portail BnF api et jeux de données
[3] Site de la BnF. https://www.bnf.fr/fr/bnf-datalab

chercheurs sur les collections de la bibliothèque, est né entre-temps, en 2021, et porte aujourd'hui cette réflexion, en lien avec d'autres départements producteurs de jeux de données. Ces interrogations ont donné lieu à l'élaboration d'une mission de stage vouée à expérimenter d'autres modes de diffusion des jeux de données et encadrée conjointement par le DataLab et le Département de la coopération numérique (DCP). En tant qu'opérateur de la coopération numérique impulsée auprès des bibliothèques françaises par la BnF à partir de 2009, le DCP est chargé du pilotage de la bibliothèque numérique Gallica[4], de l'amélioration de l'accès aux ressources numériques, d'innovation et d'accompagnement des projets recherche et développement.

La dynamique de la science ouverte prend un nouveau tournant depuis 2018. Les dernières avancées des politiques publiques redéfinissent les enjeux juridiques, techniques et environnementaux liés aux données. Les bibliothèques, fidèles à leurs missions, - l'accès à l'information en représentant le cœur même -, sont partie prenante du mouvement de diffusion des données, à travers les services à la recherche qu'elles développent. La BnF, avec son statut singulier de bibliothèque patrimoniale sous tutelle du Ministère de la Culture, élargit elle aussi sa mobilisation ; le jeune DataLab lui offre de nouvelles opportunités d'impulser une stratégie d'exposition des données plus ouverte.

Aujourd'hui, les données issues de la recherche sont au centre des préoccupations de la communauté scientifique, impliquant leur gestion active selon les principes FAIR (Facile à trouver, accessible, interoperable, réutilisable). Pour répondre à ces enjeux de gestion, d'usages et de volumes croissants des données, les infrastructures techniques que sont les entrepôts de données dédiés à leur publication, développent de nouveaux services. La nature des données de la BnF, leurs modalités d'accès et de réutilisation, leurs usages en contexte de recherche sont autant d'éléments qui permettent de réévaluer les besoins d'exposition au plus près des chercheurs. Dans ce nouveau paysage, la BnF s'interroge sur les limites et les perspectives d'évolution du portail api.bnf. Quels choix peut-elle faire, au regard de sa politique d'établissement, du profil de ses collections et de leurs usages, pour améliorer le référencement et le dépôt de ses jeux de données ?

[4] Site de Gallica, la bibliothèque numérique de la BnF. https://gallica.bnf.fr/accueil/fr/content/accueil-fr?mode=desktop

Le dépôt dans les entrepôts de données implique une gestion active des données de recherche à mesure qu'on les crée, les met à jour, les utilise, les archive, les diffuse et les réutilise. Cette technique qu'on appelle la curation de données soulève des problématiques spécifiques et appelle le développement de nouvelles compétences afin de fournir des métadonnées de qualité aux données en accès ouvert, et de rendre les données non plus seulement disponibles mais réutilisables.

Une étude des enjeux de l'ouverture des données, de l'évolution des politiques publiques incitatives en la matière, des infrastructures techniques mises en place ces dernières années, et des nouvelles activités scientifiques et documentaires qui en découlent, a fait l'objet de ce présent mémoire. Sa réflexion se poursuit ensuite à travers la présentation d'un processus de préparation et de publication de données mis en place durant le stage, afin de proposer une solution méthodologique et technique de publication consolidée et élargie. Autant d'éléments qui tentent de répondre à la question de savoir de quelle façon rendre visible des jeux de données à l'heure actuelle.

Dans une première partie, ce mémoire fera état de l'actualité des politiques publiques d'ouverture de la science aux niveaux national et européen et dont découle la dynamique de service à la recherche qui a cours dans les bibliothèques, puis présentera la place prise par le DataLab dans l'activité de recherche de la BnF.

Une deuxième étape de réflexion abordera l'importance des données de la recherche et de leur FAIRisation, ainsi que les enjeux de développement des entrepôts de données, et les critères qui président à leur choix. Elle fera également état de la nature des données patrimoniales de la BnF et de leur usage par les communautés de recherche, du portail api.bnf où elles sont déposées, afin de présenter ensuite une sélection d'entrepôts de données appropriés pour leur publication.

La troisième partie développera les enjeux de la curation des données et des activités qu'elle recouvre, et restituera une expérience de préparation et de publication d'un jeu de données en vue de sa diffusion dans un entrepôt de données. L'expérience s'assortira de propositions de référencements complémentaires, permettant ainsi de proposer des nouvelles stratégies de publication des jeux de données de la BnF.

# Première partie :

# Actualité de la science ouverte et des services à la recherche

## 1.1. Le cadre normatif de la science ouverte, de plus en plus incitatif

### 1.1.1. De l'Open Data à l'Open Science

Si la notion d'*open data* n'est pas neuve, et la préoccupation des bibliothécaires pour l'interopérabilité encore plus ancienne, elle ne saurait se confondre avec l'ouverture des données de la recherche, l'ère du web de données ayant transformé l'approche de cette question. Il ne s'agit plus seulement d'assurer la transparence, mais d'améliorer le fonctionnement des institutions en tirant parti de l'exploitation des données. Le but est de promouvoir un écosystème cohérent auprès des producteurs et des utilisateurs en vue d'encourager la collaboration et l'innovation. Afin de proposer une marche à suivre pour ouvrir les données de manière progressive, Tim Berners-Lee, l'inventeur du web, a proposé un classement en cinq étoiles[5]. Il s'agit d'abord de publier les données sur le web avec une licence ouverte[6], ensuite de le faire avec des formats lisibles par les machines, puis dans des formats ouverts, et enfin éventuellement selon les standards du web de données.

Le mouvement de l'*open data*, qui vise à rendre les données publiques pour qu'elles puissent être exploitées et réutilisées par tout un chacun, a commencé à se dessiner en France à partir de 2011[7], et s'est traduit par une politique gouvernementale qui incite par le biais de décrets les administrations et établissements publics à assurer l'ouverture juridique et technique des données.

La création, en 2014, de la mission Etalab, aujourd'hui pilotée par la Direction interministérielle du numérique (DINUM), a permis de concrétiser la politique française en termes d'ouverture et de partage des données publiques via le portail data.gouv.fr[8]. Les administrations et établissements publics y ont mis en ligne un nombre croissant de données en les plaçant sous la Licence ouverte, laquelle autorise la libre réutilisation y compris à des fins commerciales, à la condition de citer la

---

[5] M. DEKKERS, « Opportunités et défis. Linked (Open) Data », *Arabesques*, n° 64, Agence bibliographique de l'enseignement supérieur, 19 décembre 2019, p. 4-6. http://publications-prairial.fr/arabesques/index.php?id=1397

[6] INIST-CNRS. *Guide des licences ouvertes*. Site DORANum. https://doranum.fr/aspects-juridiques-ethiques/guide-des-licences-ouvertes_10_13143_tv6f-sv31/

[7] L. TELLIER-LONIEWSKI. *Données publiques - Création de « data.gouv.fr»*, portail unique d'accès aux données publiques. https://www.banquedesterritoires.fr/creation-de-datagouvfr-portail-unique-dacces-aux-donnees-publiques

[8] Portail data.gouv.fr. https://www.data.gouv.fr/fr/

source des données. La France a initié en 2015 une série de plans d'action pour un gouvernement ouvert, dont le dernier en date couvre la période 2021-2023[9]. Cette politique nationale d'*open data* porte ses fruits, puisque le récent rapport de *Data Europa* classe la France parmi les premiers pays en matière de développement et d'ouverture des données publiques[10].

Les ministères ont été les premières entités concernées par cette nouvelle politique. Le ministère de la Culture a lancé sa propre plateforme, data.culture.gouv.fr[11] en 2016, avec l'objectif de valoriser ses ressources culturelles numériques ainsi que celles de ses établissements. Parmi ces derniers, la BnF, à travers sa politique de qualité des données, a affirmé tôt son exemplarité, en étant la première institution culturelle, en 2011, à déposer sur le site data.gouv.fr un jeu de données mis à disposition du public sur data.bnf.fr[12].

Elle a également adopté en 2014 la licence ouverte de l'État[13] pour l'ensemble des métadonnées[14] qu'elle produit. Selon les termes de cette licence, les métadonnées de la BnF sont devenues librement et gratuitement réutilisables, quel qu'en soit le format et le protocole de diffusion, pourvu que les réutilisateurs en mentionnent la provenance.

L'*open data* s'inscrit dans le mouvement mondial du libre accès à la connaissance (*open knowledge*)[15] et plus largement de la science ouverte (*open science*), qui considère la science comme un bien commun dont la diffusion est d'intérêt public et général.

---

[9] *Le plan d'action national 2021-2023 pour un gouvernement ouvert*. Site de la DINUM. https://www.modernisation.gouv.fr/outils-et-formations/le-plan-daction-national-2021-2023-pour-un-gouvernement-ouvert

[10] *La France-a-la-premiere-place-en-matiere- dopen-data-en-europe-pour-la-troisieme-annee-consecutive*. Site de la DINUM. https://www.numerique.gouv.fr/espace-presse/la-france-a-la-premiere-place-en-matiere-dopen-data-en-europe-pour-la-troisieme-annee-consecutive/

[11] Plate-forme de données ouvertes du ministère de la Culture. https://www.numerique.gouv.fr/espace-presse/la-france-a-la-premiere-place-en-matiere-dopen-data-en-europe-pour-la-troisieme-annee-consecutive/

[12] *Politique de qualité des données de la Bibliothèque nationale de France*, octobre 2018. Site de la BnF. https://www.bnf.fr/fr/politique-de-qualite-des-donnees

[13] G. ILLIEN, « Le web sémantique, nouveau levier de la valeur pour les services d'information ? ». I2D - *Information, données & documents*, 52, (2015), p. 59-60. https://doi.org/10.3917/i2d.154.0059

[14] Selon le dictionnaire de l'Enssib, École nationale supérieure des sciences de l'information et des bibliothèques, «Les métadonnées sont la carte d'identité d'un document. Elles permettent de l'identifier, de le décrire, d'expliquer l'origine de sa création.» Consulté sur le site de l'Enssib. https://www.enssib.fr/services-et-ressources/questions-reponses/metadonnees-et-indexation

[15] *Comprendre les enjeux de l'ouverture des données publiques culturelles*, Ministère de la Culture, Datactivist. Novembre 2023. https://datactivist.coop/ministere-culture/jour1.html#26

La notion de science ouverte englobe différentes pratiques qui partagent une origine commune dans un mouvement émergent dans les années 1990-2000[16]. Ce mouvement avait pour objectif initial de permettre un accès en ligne gratuit aux publications scientifiques. Deux éléments clés ont contribué à son émergence : Le coût élevé des abonnements aux revues, qui pèse de plus en plus lourdement sur le budget des bibliothèques, et l'avènement d'Internet, qui permet de dématérialiser les articles de revues et de les diffuser gratuitement en ligne.

Le mouvement du Libre Accès qui, en réaction à la prise de conscience quant à la difficulté d'accéder aux résultats scientifiques, vise la mise à disposition gratuite et immédiate des résultats de recherche, est jalonné de déclarations qui sont ensuite venues structurer cette volonté initiale, en actant les principes de la science ouverte et l'engagement progressif d'un nombre toujours plus important d'acteurs de la communauté scientifique : chercheurs, bibliothèques, institutions de recherche, financeurs de la recherche, états, et éditeur. Au cours des dix dernières années, la science ouverte est devenue une réalité technologique et politique en France.

### 1.1.2. Les avancées de la loi pour la République numérique

Le mouvement de la science ouverte en France a été consolidé sur la loi pour une République numérique (LRN), promulguée le 7 octobre 2016, qui vise à promouvoir l'accès libre aux résultats des travaux de recherche publique et à autoriser l'exploration de textes et de données. Les auteurs des travaux de recherche financés à plus de 50 % par des fonds publics ont ainsi la possibilité de mettre en ligne leurs résultats en accès libre, après une période d'embargo de 6 à 12 mois (6 mois pour les sciences, technologies et médecine, et 12 mois pour les sciences humaines et sociales) indépendamment du contrat qui les lie à l'éditeur de la revue qui publie l'article[17]. Cette mesure facilite la diffusion libre des résultats de recherche, qui étaient auparavant souvent restreints et centralisés par les éditeurs. La loi autorise également l'exploration en ligne de textes et de données, une pratique essentielle notamment dans le domaine des sciences humaines et sociales. Cette

[16] *Bref historique de la science ouverte européenne*. Blog de l'Institut Pasteur. https://openscience.pasteur.fr/2022/04/07/bref-historique-de-la-science-ouverte-europeenne/

[17] F. SANTOS, *La loi pour une République numérique*, https://docs.google.com/document/u/0/d/1eVmZ4Fi_X1R_OCCM7hf9xRZ3-bgwx-5iXP-qULUa-5I/edit

pratique, qui n'était jusqu'à présent pas explicitement autorisée, doit permettre de combler le retard français sur la scène internationale en la matière.

### 1.1.3. Plan national de la science ouverte : l'état engagé

Ce renforcement de la dynamique d'ouverture impulsée par la LRN est porté par le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation (MESR) à travers le plan national pour la Science ouverte (PNSO)[18]. C'est au sein d'une bibliothèque, le Service Commun de Documentation, devenu LILLIAD[19], et à l'occasion du congrès annuel de la Ligue des Bibliothèques Européennes de Recherche (LIBER), réunissant plus de 400 bibliothèques universitaires européennes, à l'Université de Lille, que le premier PNSO a été lancé par la ministre de l'Enseignement supérieur et de la recherche en 2018[20]. Le plan impose l'accès ouvert pour les publications et pour les données issues de recherches financées sur projets. Il instaure un Comité pour la science ouverte (CoSO) et apporte son soutien à des initiatives majeures de structuration du paysage concernant les publications et les données. De plus, il prévoit un volet formation et un volet international qui sont essentiels à la mobilisation des communautés scientifiques et à l'influence de la France dans ce paysage en cours de constitution.

Après trois années de mise en œuvre, d'importants progrès ont été réalisés. Le taux de publications scientifiques françaises en accès ouvert a augmenté[21]. La création du Fonds national pour la science ouverte a permis le lancement de deux appels à projets en faveur de la publication scientifique ouverte, ainsi que le soutien à des initiatives internationales structurantes. Désormais, l'Agence nationale de la recherche (ANR) et d'autres organismes de financement exigent l'accès ouvert aux publications et la rédaction de plans de gestion des données pour les projets qu'ils financent. La fonction d'administrateur ministériel des données de la recherche a été

---

[18] *Plan national pour la Science Ouverte*, Site Ouvrir la Science. https://www.ouvrirlascience.fr/bilan-du-plan-national-pour-la-science-ouverte-2018-2021/

[19] LILLIAD (Learning Center Innovation) est le forum des savoirs à caractère scientifique de l'université de Lille, axé sur l'innovation et inauguré en 2016 au sein du campus de la Cité scientifique. https://lilliad.univlille.fr/

[20] P. CORMIER, *Le positionnement des bibliothèques universitaires et de recherche françaises dans les politiques publiques des données de la recherche.*, 2022, p. 51 https://www.enssib.fr/bibliotheque-numerique/documents/70658-le-positionnement-des-bibliotheques-universitaires-et-de-recherche-francaises-dans-les-politiques-publiques-des-donnees-de-la-recherche.pdf

[21] « Deuxième Plan national pour la science ouverte ». Site Ouvrir la science. https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte/

établie, et un réseau est en cours de déploiement dans les établissements. Une vingtaine d'universités et d'organismes de recherche ont adopté une politique en faveur de la science ouverte. Des guides et recommandations ont été publiés afin de concrétiser sa mise en pratique.

Le deuxième plan national pour la science ouverte, lancé le 6 juillet 2021, et qui verra ses effets se déployer jusqu'en 2024, vise à généraliser les pratiques. Il s'inscrit résolument dans une perspective européenne. Son ambition est de faire en sorte que les données produites par la recherche publique française soient progressivement structurées en conformité avec les principes FAIR, préservées, et, quand cela est possible, ouvertes. L'obligation d'ouverture des données de la recherche publique, posée en 2016 par la LRN, doit désormais se traduire dans les pratiques scientifiques grâce à des infrastructures et des services d'accompagnement adaptés[22]. Elle est limitée par les exceptions légitimes encadrées par la loi, par exemple en ce qui concerne le secret professionnel, les secrets industriels et commerciaux, les données personnelles ou les contenus protégés par le droit d'auteur. Dans ces cas, les pratiques de partage des données devront être favorisées à travers la définition de protocoles maîtrisés. Pour soutenir ces ambitions, le budget de la science ouverte est multiplié par trois, passant de 5M€ à 15M€ par an, et la plateforme nationale des données de la recherche, Recherche Data Gouv (RDG), lancée au printemps 2022, est créée[23].

La stratégie nationale de l'IST s'incarne dans des instances et des organisations qui fédèrent largement les acteurs de cette politique d'accès ouvert. Ainsi, les infrastructures de recherche (IR) qui constituent aujourd'hui la colonne vertébrale de la recherche pour un très grand nombre de disciplines, suivent une feuille de route nationale. La cinquième édition de la Feuille de route française des IR met l'accent sur trois leviers d'accélération des pratiques : formaliser les stratégies de science ouverte pour faciliter les dialogues et collaborations, mieux identifier la contribution des infrastructures à la connaissance scientifique à l'aide d'un identifiant unique pour les publications, codes et données qu'elles diffusent, et faciliter la réutilisation et l'intégration des données dans des jeux plus larges en

[22] Ibid.
[23] Site du MESR. 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

unifiant les descriptions des données et en créant de meilleures interfaces entre les IR et les disciplines.

### 1.1.4. Impératifs de science ouverte et sobriété numérique

La transition numérique, tout en offrant des opportunités pour ouvrir la science, représente également un défi pour l'environnement. Si la Loi pour une république numérique d'octobre 2016 impose aux opérateurs publics d'ouvrir les données qu'ils produisent, parallèlement, la loi visant à réduire l'empreinte environnementale du numérique (REEN), adoptée en novembre 2021, cherche à réduire cet impact, qu'on illustre avec la nouvelle mesure du zettaoctet[24] : *« La masse des données générées en ligne croît chaque année de façon exponentielle, avec [..] des prévisions autour de 600 Zo en 2030.* »[25]

La recherche scientifique n'échappe pas à ce mouvement, et le poids des données de la recherche dépasse encore largement celui des publications. Elles subissent différentes transformations tout au long de leur cycle de vie, passant de l'état brut à des données traitées et analysées. Aussi est-il nécessaire de réfléchir à la conservation et au partage de ces différents états, y compris à long terme, ainsi qu'à la gestion du tri et de l'élimination des versions intermédiaires. Dans certains cas, les données, publiées à la fois sur le portail *open data* d'une institution et sur un portail disciplinaire, sont dupliquées, et non synchronisées. L'ouverture représente alors un risque de redondance. Dans un contexte de crise énergétique, le défi est d'identifier les bonnes pratiques de gestion de données, en invitant les déposants à anticiper les étapes du cycle de vie de leurs données, à adopter des pratiques rationnelles et compatibles avec l'exigence de sobriété numérique. La gestion des données de recherche évolue face à cet enjeu majeur : la conciliation de la durabilité et de la sobriété.

Ces enjeux de développement de la science ouverte s'inscrivent plus largement dans un contexte de recherche internationale, et un cadre législatif européen favorable. Le cadre politique national réglementant l'ouverture des données de la

[24] Zetta (symbole Z) est le préfixe du Système international d'unités (SI) qui représente un trilliard (10 puissance 21), soit mille milliards de milliards.

[25] A. MOALIC, É. LEHOUX, C. PION, et C. LASNE. « La science ouverte à l'épreuve de la sobriété.» *Arabesques*. 1 avril 2023. N° 109, pp. 12-14. DOI 10.35562/arabesques.3418.

science s'appuie en effet sur un large éventail de programmes lancés par l'Union européenne.

### 1.1.5. La promotion de la science ouverte à l'échelle européenne

C'est à compter de 2014 que la Commission européenne commence à employer le terme de « science ouverte », Dans la continuité, les textes européens ont imposé l'ouverture des données de la recherche dès lors qu'elles ne sont pas protégées par un droit (d'auteur, des données personnelles, des secrets industriels et commerciaux) et, encouragent vivement les États à mettre en place des politiques nationales similaires.[26]

Parmi les nombreuses initiatives visant à promouvoir la science ouverte en Europe, nous présentons OpenAIRE et DARIAH (Digital Research Infrastructure for the Arts and Humanities) afin de donner les repères nécessaires en vue de l'environnement du scénario de publication proposé en troisième partie, car ils y sont mentionnés. OpenAIRE est une infrastructure européenne née en 2006[27], qui, pour répondre à l'obligation d'assurer le libre accès aux publications issues des recherches qu'il aura contribué à financer, agrège les données en moissonnant des réservoirs de données comme HAL (Hyper Article en Ligne), la plateforme pluridisciplinaire nationale pour le dépôt et la consultation des travaux et résultats de recherches scientifiques, et Zenodo, entrepôt de données du CERN (European Organization for Nuclear Research)[28] issu d'une collaboration avec OpenAire et financé par la Commission Européenne.

Afin de doter l'Union Européenne d'une infrastructure d'échange et de services en commun, DARIAH, est née en 2006 également. Sa principale mission est de doter l'Europe d'un cadre pour créer et coordonner des projets autour des infrastructures en humanités numériques dans toutes les disciplines des sciences humaines et sociales. DARIAH s'appuie sur une structure légale de type ERIC (European Research Infrastructure Consortium). C'est une infrastructure distribuée, en ce qu'elle met en commun les infrastructures existantes de chaque pays engagé.

[26] P. CORMIER, *Le positionnement des bibliothèques universitaires et de recherche françaises dans les politiques publiques des données de la recherche*, Mémoire d'étude, Enssib, 2022. https://www.enssib.fr/bibliotheque-numerique/documents/70658-le-positionnement-des-bibliotheques-universitaires-et-de-recherche-francaises-dans-les-politiques-publiques-des-donnees-de-la-recherche.pdf

[27] Site couperin de la science ouverte en France. https://scienceouverte.couperin.org/quest-ce-quopenaire/

[28] Consulté sur le site de CatOpidor; https://cat.opidor.fr/index.php/Zenodo

La participation française à DARIAH est coordonnée par l'infrastructure de recherche (IR*) Huma-Num depuis janvier 2014.

A l'heure actuelle, c'est Horizon Europe, le programme-cadre de l'Union européenne pour la recherche et l'innovation dans la période allant de 2021 à 2027, qui prend la suite du programme Horizon 2020, et constitue un nouveau souffle pour la recherche européenne, en allant plus loin puisqu'il étend l'obligation de diffusion en accès ouvert aux formes longues de publications, c'est à dire principalement les livres, en *open access* immédiat, sur des archives ouvertes, et avec des licences ouvertes. D'après Marlène Delhaye,[29] *« C'est une différence notable avec la politique de H2020 [..] qui se limitait jusqu'ici aux articles de revues scientifiques. Il édicte également la fin des embargos, met l'accent sur la gestion des données de recherche, qui devient la règle par défaut, dans le respect des principes FAIR, et rend éligibles à un remboursement les coûts de gestion des données. »*

Le programme lance également un chantier majeur pour les infrastructures de recherche : l'EOSC (*The European Open Science Cloud*), initié en 2015 et ouvert en 2018, qui vise à fournir un réseau de données FAIR et de services pour la recherche en Europe en reliant les infrastructures européennes et nationales existantes.[30]

La BnF, engagée depuis ses débuts dans la recherche, est mobilisée continuellement dans de nombreux partenariats aux côtés d'autres acteurs scientifiques œuvrant à une politique d'ouverture des données. Elle cherche aujourd'hui à s'impliquer dans les réseaux européens, et cet engagement est porté par son jeune laboratoire et service à la recherche, le DataLab.

## 1.2. Le BnF Data Lab, un laboratoire pour la recherche

### 1.2.1. Les services à la recherche en bibliothèque, faiseurs de science ouverte

Les bibliothèques, qu'elles soient universitaires ou patrimoniales, constituent des gisements de données considérables, en particulier pour les chercheurs en sciences humaines. Introduire les humanités numériques dans la bibliothèque est un

[29] M. DELHAYE, 2021. « Science ouverte : qu'est-ce qui change avec Horizon Europe ? » *Le réservoar*. 5 février 2021. Consulté sur le carnet Hypothèses Le Reservoar. https://oaamu.hypotheses.org/2722 :

[30] L. DE LUCA, Pascal LIEVAUX. « Un projet européen de Cloud collaboratif pour le patrimoine culturel. » *Culture et recherche*, 2023, 144. ⟨hal-04146822⟩

défi de mieux en mieux relevé, qui nécessite à la fois de nouvelles compétences métiers (ingénieurs, experts, etc.) et de nouvelles infrastructures en termes d'espaces et d'équipements : serveurs, machines virtuelles, logiciels, et déploiement d'API.

Une initiative notable a encouragé ce développement des Humanités numériques en 2017 ; il s'agit de la déclaration de Santa Barbara[31] sur les « *collections as data* » ou « collections en tant que données » en français, idée selon laquelle les collections numériques des bibliothèques peuvent être exploitées comme des données, informatiquement, ce qui inclut la fouille de texte, la visualisation de données, la cartographie, l'analyse d'images, de sons et de réseaux. Des collections définies par leur usage, une idée qui a mobilisé les professionnels de l'IST dans la création de Digital Humanities center et de « *datalabs* » développés au sein de grandes bibliothèques universitaires et nationales, parmi lesquelles l'Alan Turing Institute et le British Library Labs, et de services d'archives tels que le Lab de l'INA (Institut national de l'audiovisuel)[32] ou celui des Archives nationales.

En France, ce sont les bibliothèques universitaires (BU), par leur proximité avec les chercheurs, qui ont pu prendre la mesure de la mutation à l'œuvre, à l'exemple de l'université Bordeaux-Montaigne, dont le service commun de documentation (SCD) se présente comme « fournisseur de services » sur des projets ayant trait aux humanités numériques. À l'université de Lille, c'est une BU d'un nouveau genre que propose LILLIAD, « forum des savoirs » avec des espaces dédiés aux apprentissages. Avec son DataLab dont l'ouverture a eu lieu en 2021, la Bibliothèque nationale et universitaire de Strasbourg (Bnu) s'est engagée elle aussi dans un nouveau service d'accompagnement de la recherche dans les humanités numériques. Marie Carlin et Arnaud Laborderie respectivement coordinatrice du BnF Datalab et chef de projet au service de la Coopération numérique et de Gallica à la BnF, observent : « *En intégrant en leur sein des learning centers, les BU renouvellent la relation entre bibliothèque et formation par l'articulation entre l'enseignement (teaching), l'acquisition de connaissances (learning), la documentation et la formation aux technologies (training).* »[33]

---

[31] The Santa Barbara statement on Collections as data, 2017. https://collectionsasdata.github.io/statement/.

[32] Carnet hypothèses INA Le Lab. https://inalelab.hypotheses.org/

[33] M. CARLIN et A. LABORDERIE, « Le BnF DataLab, un service aux chercheurs en humanités numériques », *Humanités numériques*, vol. 4, 1er décembre 2021. https://hal-bnf.archives-ouvertes.fr/hal-03285816

En effet, dès le début des années 2000, les bibliothèques universitaires, confrontées à l'augmentation continue du coût des abonnements aux revues proposées à leurs publics, ont apporté leur soutien au mouvement en faveur de l'accès ouvert aux publications scientifiques. Depuis, leur implication n'a eu de cesse de se renforcer, comme le démontre leur rôle dans l'intégration à l'environnement local de l'archive nationale HAL, avec la multiplication de portails institutionnels qu'elles animent. Dans une première phase, leur principal objectif était de promouvoir et de gérer des archives ouvertes, ainsi que de sensibiliser les chercheurs et les instances de gouvernance des établissements à l'accès ouvert. L'extension de la science ouverte à l'ensemble du processus scientifique a incité les bibliothèques universitaires à réfléchir à leur rôle spécifique dans l'assistance aux chercheurs pour la gestion de leurs données.

Les résultats de l'enquête[34] mise en place dans le cadre d'une étude réalisée sur la place des bibliothèques universitaires dans le développement de la science ouverte, montrent qu'elles ont pour la majorité diversifié leur offre de services, mais se heurtent toutefois à des déficits tant en termes de ressources humaines que de compétences techniques et juridiques nécessaires pour accompagner les chercheurs dans les particularités de leur domaine de recherche, faire évoluer leur politique documentaire en faveur des contenus ouverts, améliorer la visibilité de ces contenus et évaluer leur utilisation et leur impact.

A ce stade, elles ont avant tout besoin d'une volonté politique ferme et de stratégies d'établissements affirmées et soutenues par des ressources pérennes. Cette démarche des tutelles implique incontestablement un décloisonnement entre les différentes professions et la mise en synergie des compétences au sein de groupes de formation et de réflexion ouverts tant aux bibliothécaires qu'aux chercheurs, de collaborations entre équipes de recherche mixtes, de guichets de services regroupant les compétences des différents acteurs, ainsi que de pôles éditoriaux de proximité.

[34] C. LETROUIT et al., *La place des bibliothèques universitaires dans le développement de la science ouverte*. Rapport de l'Inspection générale de l'éducation, du sport et de la recherche, février 2021. https://www.education.gouv.fr/la-place-des-bibliotheques-universitaires-dans-le-developpement-de-la-science-ouverte-322815

### 1.2.2. La BnF, une bibliothèque patrimoniale et de recherche engagée

L'histoire institutionnelle française a prioritairement confié le patrimoine documentaire et archivistique en héritage à l'actuelle BnF, le plaçant ainsi sous la responsabilité du ministère chargé de la Culture. C'est ainsi que les politiques en faveur du patrimoine ont été d'abord formalisées à partir de la BnF et de ce ministère plutôt qu'à partir du secteur universitaire, alors même que les publics du patrimoine sont majoritairement universitaires. Le développement des catalogues collectifs, leur accès en ligne et les programmes de numérisation ont ensuite continuellement rapproché ses fonds des chercheurs.

Sous tutelle du Ministère de la Culture dont les missions de sauvegarde, de protection et de mise en valeur du patrimoine culturel ne sont pas tournées vers les chercheurs, la BnF affirme, elle, dès son décret de création du 3 janvier 1994, une mission de recherche, ce qui représente un défi car elle ne relève pas des politiques du MESR. Ce statut singulier l'amène continuellement à interroger sa juste place, au point de convergence de ces missions.

Un plan quadriennal de la recherche, dispositif qu'elle a engagé depuis 1994, permet jusqu'à aujourd'hui de conduire des programmes de recherche autour des sciences du livre et des bibliothèques.. Partenaire d'organismes et groupements de recherche nationaux et internationaux (Labex, Equipex, communauté d'universités et établissements, etc.), elle contribue à l'élaboration des hypothèses scientifiques et des méthodologies des projets dans lesquels elle choisit de s'impliquer ; elle est également à l'initiative de certains.

D'un partenariat fondé sur les disciplines « cœur de métier » de la BnF, la collaboration s'est élargie à d'autres types de laboratoires pour favoriser des collaborations de recherche basées sur les collections numériques dans leur ensemble, qu'elles soient intrinsèquement numériques (comme le dépôt légal de l'internet) ou issues de la numérisation (comme les collections de Gallica traitées par reconnaissance optique de caractères), sur les nouveaux usages de lecture numérique, d'interactions sur les réseaux sociaux, et en sciences de la donnée.

### 1.2.3. Du projet Corpus au BnF DataLab

Le département de la Découverte des collections et de l'accompagnement à la recherche (DCA), auparavant nommé département de l'Orientation et de la recherche

bibliographique, est l'héritier de la Salle des catalogues et des bibliographies de l'ancien département des Imprimés de la Bibliothèque nationale. Les professionnels du DCA retracent ainsi l'histoire des lieux d'accueil de l'orientation bibliographique :

*« La salle X a pour ancêtre la salle des catalogues et des bibliographies, alors sur le site Richelieu, aménagée par l'architecte Michel Roux-Spitz sous la principale salle de travail de la Bibliothèque Nationale. Née de l'ambition de Suzanne Briet, alors responsable du bureau des recherches, et de Julien Cain, administrateur de la Bibliothèque Nationale, qu'elle a su convaincre très tôt de la nécessité d'un lieu pluridisciplinaire où seraient rassemblés les outils bibliographiques nécessaires non seulement pour le repérage dans les collections de la bibliothèque mais aussi, plus largement, pour toutes recherches scientifiques, la salle des catalogues a ouvert au public le 23 avril 1934.»*[35]

Au sein de ce Département (DCA), le service de l'assistance à la Recherche et de la formation des usagers à la recherche (SAR) abrite depuis le 18 octobre 2021 le BnF DataLab, un lieu conçu pour le travail individuel et collaboratif, qui offre un ensemble de services destinés à accompagner les chercheurs souhaitant travailler sur les collections numériques de la BnF. Il a été conçu au cœur du projet Corpus qui a permis de mener une étude prospective de besoins auprès d'usagers potentiels afin de définir les contours du futur laboratoire.

C'est face à la croissance de projets centrés sur les collections numériques et les données, et dans le cadre du plan quadriennal de recherche de la BnF pour 2016-2019, que ce projet Corpus[36], piloté par Emmanuelle Bermès alors conservatrice à la Direction des services et des réseaux, en collaboration avec Eleonora Moiraghi, s'est donné pour objectif de simplifier l'accès des chercheurs aux collections numériques de la BnF, en facilitant leur exploitation sous forme de données : analyse de textes et d'images, visualisation de données, réutilisation et alignement de référentiels, etc.

Trois années d'expérimentation, autour de 4 collections numériques : archives du web, numérisations, métadonnées, images et cartes, ont révélé que l'analyse de données de type *Text and data mining* (TDM) était en passe de se généraliser dans toutes les disciplines, bien que le rythme et l'ampleur de cette évolution soient

[35] C. ÉLOI, E. MOIRAGHI et V. ROSE, « Un espace pour les humanités numériques à la BnF », *Bulletin des bibliothèques de France (BBF)*, 2019, n° 17, p. 90-95.https://bbf.enssib.fr/consulter/bbf-2019-17-0090-009 ISSN 1292-8399

[36] E. MOIRAGHI, *Le projet Corpus et ses publics potentiels : Une étude prospective sur les besoins et les attentes des futurs usagers*, *op. cit.*

difficiles à prévoir. En ce qui concerne l'aménagement de l'espace, cette enquête a souligné la nécessité de prendre en compte de nouvelles contraintes temporelles (les traitements de données s'étendant sur plusieurs heures) et de concevoir des espaces dédiés non seulement au travail individuel, mais aussi au travail collaboratif, à la formation, au partage d'expérience et à la convivialité.

Suite à cette enquête et à d'autres expérimentations, la BnF a lancé un projet de réaménagement de la salle X, située au rez-de-jardin et actuellement dédiée à la recherche bibliographique. Emmanuelle Bermès explique que « *la proximité des experts de la BnF est perçue comme la principale valeur ajoutée [du] lieu physique.*»[37] Le DataLab est pensé comme un lieu de dialogue qui permet aux chercheurs de faire émerger leurs besoins et de se former aux côtés des bibliothécaires qui peuvent de cette façon apporter une expertise sur les collections, les questions juridiques et les aspects techniques, accompagner les chercheurs dans leur compréhension de la structure des catalogues et des données, les aider à identifier les outils pertinents pour leur recherche.

Autre étape ambitieuse du projet Corpus, « *Le site Api et jeux de données, ouvert en 2017 à l'occasion du 2e hackathon de la BnF, apporte le volet numérique de l'offre, en documentant les API d'accès aux collections numériques et en redistribuant les données enrichies par les chercheurs* ». Enfin, « *la dernière étape de la construction de ce nouveau service aux chercheurs réside dans l'élaboration de partenariats. Qu'il s'agisse du CNRS (via l'INSHS, la TGIR Huma-Num ou des laboratoires comme le Lattice) ou d'établissements d'enseignement supérieur (Sorbonne Université, Ecole Polytechnique de Lausanne).* »[38]

Eleonora Moiraghi explique ainsi les enjeux techniques et de compétences que représentent ce partenariat pour la BnF : « *La TGIR Huma-Num constitue un partenaire privilégié en raison de son expérience dans le stockage, la gestion et le traitement des données de la recherche, mais aussi en raison de l'infrastructure mise en place qui comprend plusieurs machines virtuelles, un firewall imposant et la possibilité de demander l'installation de logiciels à distance.* »[39]

---

[37] E. BERMÈS, « BnF : des métadonnées au service de projets de recherche innovants », *Arabesques*, 95 | 2019. https://publications-prairial.fr/arabesques/index.php?id=1302

[38] Ibid.

[39] E. MOIRAGHI, *Le projet Corpus et ses publics potentiels : Une étude prospective sur les besoins et les attentes des futurs usagers*, *op. cit.*

### 1.2.4. Le BnF DataLab, entre service à la recherche et Laboratoire R&D

Le DataLab constitue un lieu de travail, d'échange, de résidence pour les chercheurs, une offre de services sur place et à distance, pour accompagner les usagers, et intègre également une fonction « Recherche & Développement » qui vise à innover en expérimentant des technologies nouvelles afin de passer de la recherche à une faisabilité avérée. Le DataLab prend également cette forme de laboratoire qui permet de tester des idées à côté des schémas existants, et d'explorer des terrains inconnus grâce à la mise en place de partenariats scientifiques.

Son offre sur place et à distance se répartit autour de trois axes ; d'abord l'accueil et l'orientation des chercheurs avec la définition d'un parcours, ensuite l'identification des collections (accès aux données, aide à la constitution des corpus), et enfin le travail sur les corpus, avec la mise à disposition dans l'espace physique du DataLab d'un environnement de travail, l'animation de formations à l'intention des chercheurs, et le suivi de projet avec des experts BnF.

L'objectif du DataLab est de poursuivre la co-construction des services et des outils avec les équipes de recherche, à travers une coopération et valorisation scientifique qui continue de s'interroger sur ses formats, son modèle économique et ses moyens. Son offre de services s'appuie sur une activité de collaboration avec les équipes de recherche qui sont invitées à contribuer à l'enrichissement des ressources numériques, qu'il s'agisse d'intégrer des données enrichies dans Gallica, par exemple en produisant un OCR (*Optical Character Recognition*) corrigé, en transformant un corpus en TEI (*Text encoding Initiative*), ou de mettre à disposition des scripts.

Marie Carlin et Arnaud Laborderie expliquent ainsi cet axe de développement du DataLab : « *L'aspect collaboratif est une dimension majeure de l'animation du BnF DataLab : les outils, développés dans le cadre de projets de recherche, seront reversés dans une boîte à outils et pourront bénéficier à d'autres projets mais aussi aux programmes internes, comme la reconnaissance d'images, le séquençage de la presse ancienne, la détection automatique des styles littéraires, etc. Les outils et les jeux de données ainsi réalisés devront se conformer aux recommandations*

*d'Huma-Num (open access, FAIR data, interopérabilité, etc.). Ils seront également accessibles et documentés sur GitHub*[40]*.* »[41]

Cet aspect de développement comme laboratoire favorisant l'expérimentation et la *R&D* a été particulièrement fructueux les deux premières années d'existence du DataLab. Cependant, si certains travaux conduits en son sein peuvent intéresser la BnF pour faire évoluer ses propres outils, sa finalité première n'est pas d'être un laboratoire interne de *R&D*. Il a été conçu pour apporter un service nouveau aux chercheurs désireux d'exploiter informatiquement les importantes collections et données dématérialisées de la BnF, service d'autant plus nécessaire qu'une partie de ces collections « protégées » (croissante avec le développement du dépôt légal numérique[42]) n'est consultable que dans les emprises de la BnF. Aussi le DataLab cherche-t-il aujourd'hui à consolider une offre plus classique de formation à la recherche, afin d'assurer le développement d'un public de chercheurs en humanités numériques autour de ses collections.

Favoriser l'expérimentation et la *R&D* en partenariat implique une ouverture sans cesse renouvelée aux acteurs de l'enseignement supérieur à l'échelle nationale et européenne. C'est pourquoi, dans la continuité de son engagement avec Huma-Num, la BnF intensifie encore ses partenariats scientifiques en adhérant à l'ERIC DARIAH au mois de mai 2023. Le DataLab, fer de lance de cette adhésion, y voit l'opportunité de rejoindre un réseau européen d'experts, partenaires potentiels pour monter et rejoindre des projets, et de fournir au sein de cette communauté une expertise sur ses données. Pour incarner ce partenariat, il envisage entre autres actions, de contribuer au *marketplace* du SSHOC, soutenu par DARIAH en offrant ainsi aux outils et jeux de données de la BnF une visibilité européenne. Ce marketplace[43] est un portail de découverte mutualisant et contextualisant, à destination des communautés scientifiques européennes, des outils, services, matériels de formation, jeux de données, publications et *worflows*[44]. La démarche d'adhésion à DARIAH ne requiert

---

[40] GitHub est un service web d'hébergement et de gestion de développement de logiciels. https://fr.wikipedia.org/wiki/GitHub.

[41] M. CARLIN, A. LABORDERIE, 2021. « Le BnF DataLab, un service aux chercheurs en humanités numériques. » *Humanités numériques* [en ligne]. 1 décembre 2021. Vol. 4. [Consulté le 13 avril 2023]. url : https://hal-bnf.archives-ouvertes.fr/hal-03285816

[42] Site de la BnF. https://www.bnf.fr/fr/le-depot-legal-numerique

[43] Site du Social Sciences & Humanities Open Marketplace. https://marketplace.sshopencloud.eu/

[44] On appelle *workflow* (ou flux de travail) la modélisation et la gestion informatique de l'ensemble des tâches à accomplir. Un workflow documentaire formalisera des processus métiers précisant les acteurs impliqués et rôles, les délais, modes de validation, les suites de tâches.

pas de financement, car la cotisation est intégrée au budget d'Huma-Num qui coordonne la participation française dans DARIAH. Elle requiert plutôt une nouvelle mobilisation de moyens pour faire un état des lieux de la gouvernance des données de la BnF. Il s'agira de trouver les termes d'une coordination entre les différents producteurs de données à la BnF, ainsi qu'avec Huma-Num. Identifier la distribution des rôles de chacun permettra d'assurer les activités de gestions de données requises en vue de la publication, comme par exemple le formatage des jeux de données en anglais, ou la migration de fichiers vers d'autres formats. La démarche implique également de déterminer les actions de coordination à mettre en place au sein de la BnF, le temps de travail, les ressources nécessaires, et le développement des compétences que représentent ces versements dans les outils de référencement de DARIAH[45]. Les partenariats avec les infrastructures de recherche constituent également des opportunités de valoriser les actions menées dans le BnF DataLab : manifestations, restitutions des travaux des équipes de recherche, et publications.

La route menant au libre accès à la recherche scientifique a été longue. Les bibliothèques, touchées de plein fouet par l'explosion des tarifs d'accès aux publications scientifiques, ont été parmi les premières à prendre conscience du danger que les stratégies de fermeture des éditeurs spécialisés faisaient courir à la liberté de la recherche et à l'exercice de la fonction documentaire. Les professionnels des bibliothèques ont su, par leur implication et leur expertise, s'imposer dans le débat et s'y faire reconnaître comme des acteurs légitimes. En 2016 la donne a changé avec la LRN, autorisant les chercheurs financés sur fonds publics à déposer leurs articles dans une archive ouverte, et en 2018 le combat a changé d'échelle avec le lancement de l'ambitieux PNSO. Ces appuis forts doivent permettre aux bibliothécaires de continuer à élargir leur réflexion. A l'échelle de la BnF, la création du DataLab constitue une véritable opportunité de poursuivre le chemin vers l'ouverture des données. Ses partenaires du monde de la recherche sont partie prenante de ses interrogations autour de la diffusion des données sur des entrepôts de données de la recherche.

[45] Ces informations ont été recueillies au cours d'entretiens réalisés avec l'équipe du DataLab.

# Deuxième partie :

# Données, entrepôts, et pratiques de recherche

# 2. 1. FAIRiser les données de la recherche

## 2.1.1. Le partage des données, au coeur de la recherche

Les données de recherche font référence à l'ensemble des informations recueillies, observées ou créées par les chercheurs dans le cadre d'un projet de recherche, qu'elles soient sous forme numérique ou non[46]. Ces données servent de fondement à l'élaboration des hypothèses scientifiques. Les archivistes, au sein de l'Association des archivistes français, en donnent cette définition[47] : « *Les données de la recherche sont des informations, spécimens et matériaux produits, recueillis et documentés. Elles sont collectées ou exploitées à des fins de recherche et de preuves par les chercheurs et leurs équipes. À ce titre, elles constituent une partie des archives de la recherche.* » L'archiviste a donc un rôle à jouer dans leur gestion. Programme

Ainsi, les données de la recherche constituent à la fois le produit de la recherche et un élément essentiel de la communication scientifique, englobant une grande variété de sources et de matériaux. Différents types de données peuvent être distingués. Les données primaires ou brutes comprennent les données empiriques, observées ou mesurées. Certaines de ces données ne sont pas destinées à être stockées ou partagées. Les données secondaires, elles, dérivées des données primaires, sont annotées, enrichies et interprétées, apportant ainsi une valeur supplémentaire aux données initiales. Leur traitement et leur analyse peuvent impliquer la participation d'autres acteurs. Les métadonnées, quant à elles, jouent un rôle dans la structuration, la gestion et la facilitation de l'accessibilité des données primaires et secondaires. Elles fournissent également des informations sur les conditions de partage. Les données présentent une grande diversité de nature (statistiques, mesures, résultats d'expériences, objets de fouille, observations de terrain, transcription d'entretiens, archives) et de formes (texte, sons, vidéos, images, algorithmes, codes sources, scripts).

Dans le contexte de la science ouverte, il est désormais obligatoire pour les chercheurs de partager ces données[48]. Le principe fondamental de leur ouverture

[46] Site de Recherche.data.gouv. https://recherche.data.gouv.fr/fr/page/quelles-donnees-de-recherche

[47] L. Sbeih et al., « L'archivage des données de la recherche à l'Inra. Eléments de réflexion, démarche et perspectives », *Cahier des Techniques de l'INRA*, (4 avril 2020).

[48] Site du Programme prioritaire de recherche (PPR) Autonomie.

repose sur la maxime selon laquelle elles doivent être « *aussi ouvertes que possible, tout en restant aussi fermées que nécessaire* » selon l'expression de la communauté européenne[49]. En effet, il ne s'agit pas de les rendre accessibles systématiquement à tous les publics. Bien que l'ouverture sans restriction puisse être une option envisagée, elle n'est qu'une parmi plusieurs autres possibilités. Outre les contraintes juridiques liées au partage de certaines données, il existe de nombreuses raisons qui peuvent justifier la décision d'en restreindre l'accès : confidentialité liée aux personnes interrogées et relevant du Règlement général sur la protection des données (RGPD), protection des données liées aux brevets et aux marques pour des raisons économiques, besoin de produire d'abord des résultats d'analyses primaires.

La diffusion des données relève tout d'abord d'arguments épistémologiques ; le partage des données de recherche permet en effet de fournir à la communauté scientifique les éléments empiriques nécessaires pour étayer un raisonnement, et d'autre part il facilite l'identification des limites des données, ce qui permet aux producteurs de prendre en compte ces limites, et d'améliorer la collecte de données à l'avenir. Ainsi, la diffusion des données favorise une approche plus transparente, critique et évolutive de la recherche.

Par ailleurs, la recherche étant un processus collaboratif et cumulatif, la mise à disposition à l'ensemble de la communauté des données utilisées dans les travaux de recherche tend à permettre, dans la mesure du possible, à d'autres équipes de tirer parti de jeux de données auxquels elles n'auraient pas pu accéder en raison des coûts de collecte ou des méthodes utilisées.

En outre, le dépôt constitue un moyen de stockage sécurisé des données, et permet un contrôle de leur diffusion aux équipes de recherche autorisées. Ainsi, cela garantit la protection des données sensibles, et facilite une utilisation responsable et contrôlée par les chercheurs qualifiés.

L'objectif de la diffusion de données est également d'augmenter la visibilité de la production scientifique d'un collectif de recherche, et d'être repéré comme partenaire potentiel en vue d'un projet de réutilisation de ces données ou d'un projet potentiel sur des sujets avoisinants.

---

https://www.miti.cnrs.fr/wp-content/uploads/2021/06/la_diffusion_des_donnees.pdf

[49] DIRECTIVE (UE) 2019/1024 DU PARLEMENT EUROPÉEN ET DU CONSEIL du 20 juin 2019 concernant les données ouvertes et la réutilisation des informations du secteur public.
https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32019L1024&rid=1

Enfin, les motivations sont également archivistiques ; la diffusion des données est aussi une garantie qu'elles restent accessibles pour le projet qui les a créées, au-delà des changements de personnes, et pour la personne qui les a créées, au-delà de l'obsolescence technologique.

### 2.1.2. Rendre les données FAIR

On l'a vu, intégrer la gestion de données scientifiques aux activités de recherche constitue désormais une condition majeure d'ouverture de la science. Cette gestion est active tout au long du cycle de vie des données, généralement décomposé en six étapes[50] : création ou collecte, traitement, analyse, conservation, accès, réutilisation. Les principes FAIR ont été imaginés par une communauté composée de chercheurs, d'éditeurs, de sociétés savantes, d'universités, de bibliothécaires, d'archivistes, préoccupés d'élaborer des pratiques de gestion de données ouvertes.

Ces principes ont ensuite été repris et énoncés en 2016 par un groupe de chercheurs de différents laboratoires à travers le monde dans la revue *Scientific Data*[51]. A chaque lettre du mot FAIR sont liées des bonnes pratiques de gestion des données : **F**aciles à trouver**, A**ccessibles**, I**nteropérables et **R**éutilisables. Chaque principe FAIR se décline en un ensemble de caractéristiques que doivent présenter les données et les métadonnées pour faciliter leur découverte et leur utilisation par les hommes mais aussi par les machines.

Le principe de données faciles à trouver se traduit par leur indexation dans un dispositif permettant de les rechercher et leur identification par un identifiant global unique et pérenne, ainsi que par la richesse des métadonnées les décrivant.

Selon le deuxième principe FAIR, elles se doivent ensuite d'être accessibles, par leur identifiant et via un protocole de communication standardisé ouvert, libre, qui peut être implémenté de manière universelle, et permet l'authentification et l'autorisation si besoin. Les métadonnées sont accessibles même quand les données ne le sont plus.

Le respect du troisième principe d'ouverture implique d'avoir des données interopérables qui doivent pouvoir être partagées et rediffusées entre différents

[50] Site de l'Université Paris Saclay. https://www.universite-paris-saclay.fr/recherche/science-ouverte/le-cycle-de-vie-des-donnees

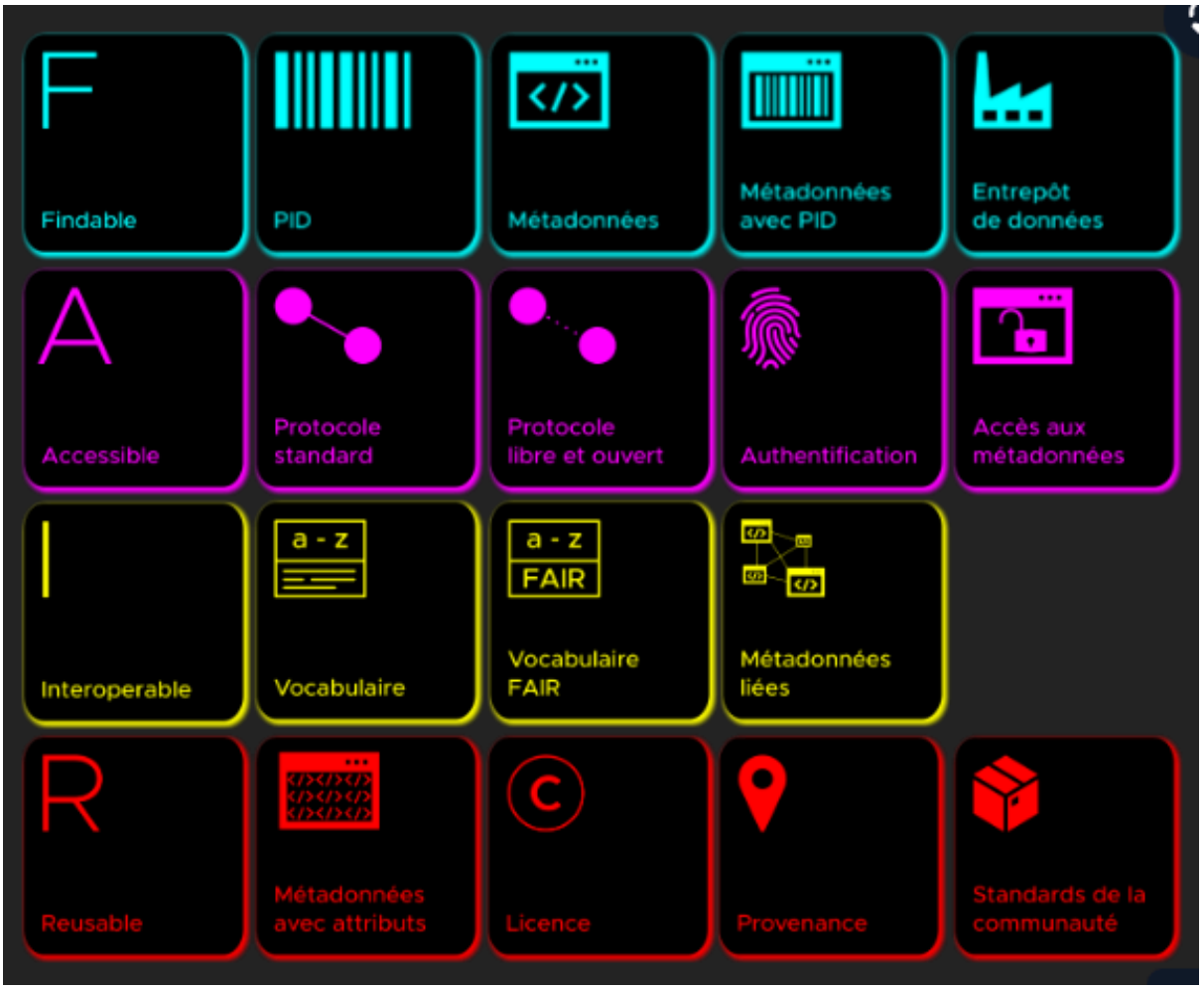
[51] Site de la revue Scientific Data. https://www.nature.com/articles/sdata201618

systèmes informatiques, grâce à l'utilisation de formats ouverts, d'identifiants, et à l'application de normes spécifiques.

Enfin, des données FAIR sont réutilisables pour de futurs travaux de recherche, d'enseignement, de reproduction. Pour cela, elles sont mises à disposition selon une licence explicite et accessible[52], associées à leur provenance, et correspondent aux standards des communautés indiquées.

Ce schéma proposé par la plateforme DoRANum (Données de la recherche : Apprentissage Numérique[53], qui propose des ressources pédagogiques d'auto-formation pour accompagner la communauté scientifique dans la gestion et le partage des données de la recherche, a pour objectif d'expliquer de façon simplifiée chaque item des principes FAIR.

Figure 1 : Schéma FAIR de la plateforme DoRANum




[52] Site de l'Université de Rennes 2.
https://socle.univ-rennes2.fr/vos-besoins/savoir-utiliser-licences

[53] Site de DoRANum.
https://doranum.fr/enjeux-benefices/principes-fair_10_13143_z7s6-ed26/

### 2.1.3. Favoriser la réutilisation par la diffusion de jeux de données

Afin de donner les clés pour faire comprendre ses données et les rendre pleinement réutilisables, la mise à disposition sous la forme de jeux de données (ou *dataset*) constitue un levier important. Le jeu de données peut être défini comme l'agrégation, sous une forme lisible, de données brutes ou dérivées présentant une certaine « unité »[54]. Cet ensemble de données prend forme dans des tableaux avec des lignes et des colonnes. Chaque colonne décrit une variable particulière, et chaque ligne correspond à un élément donné de l'ensemble de données. Le nombre de fichiers peut être très important dans un jeu de données, et il est citable comme un objet unique.

Les jeux de données ont vocation à être partagés, que ce soit en interne ou vers l'extérieur. Il doivent donc être accompagnés d'une série d'éléments et d'outils permettant leur réutilisation : métadonnées, datavisualisations et API. Les métadonnées associées au *dataset* renseignent les informations relatives à la nature de ses données : licence, dates de création et de modification, producteur, modèle de donnée utilisé, etc. Ces informations permettent de rassurer le réutilisateur sur la fiabilité du jeu de données. Plusieurs normes et schémas de métadonnées existent pour décrire les différents types de données et pour répondre aux besoins qui varient d'une discipline à l'autre. DCAT et *Dublin Core*[55] sont des standards utilisés pour des données générales, DDI (*Data Documentation Initiative*) est un standard pour les Sciences sociales.

Ensuite, étant donné que sous sa forme brute un *dataset* peut être difficile à analyser, des datavisualisations peuvent être proposées, directement ou via des outils mis à disposition. Elles permettent la lecture des données à travers des cartes ou graphiques, ou des formats plus avancés tels que des *dashboards* ou *data stories*.

---

[54] R. GAILLARD, *De l'Open data à l'Open research data : quelle(s) politique(s) pour les données de recherche ?* - Enssib, 2014. https://www.enssib.fr/bibliotheque-numerique/documents/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche.pdf

[55] Site de Wikipedia. https://fr.wikipedia.org/wiki/Dublin_Core

Enfin, indispensables pour récupérer des grands ensembles de données en temps réel, les API sont généralement fournies par les producteurs des *datasets*. Une fois connectées, elles permettent de récupérer des informations actualisées.

Les jeux de données peuvent être utilisés de nombreuses façons. Par exemple, en vue de susciter des usages de recherche, la publication de *datasets* spécifiques et l'autorisation de les utiliser pour des hackathons ou des concours ouvrent la porte à une innovation au sein d'un écosystème. Ils peuvent être également mis à disposition via des portails *open data* pour communiquer en toute transparence sur diverses thématiques, et favoriser en général la découverte des données de l'organisation dans toutes leurs latitudes.

Pour conclure, si le partage de données pose de nombreuses questions épistémologiques ou éthiques, il a un intérêt certain pour la recherche, et nécessite des infrastructures vouées à la diffusion. Ces plateformes permettront aux chercheurs de gagner du crédit par leurs données et de réutiliser des jeux existants.

## 2.3. Entrepôts de données : critères pour le choix

### 2.3.1. Notions, caractéristiques et fonctionnalités des entrepôts

Les enjeux liés à la gestion et au partage des données de la recherche nécessitent des outils appropriés communément appelés « Entrepôts de données ». Selon Wikipedia, « *Un entrepôt de données (ou base de données décisionnelle ; en anglais, data warehouse) désigne une base de données utilisée pour collecter, ordonner, journaliser et stocker des informations provenant de base de données opérationnelles? et fournir ainsi un socle à l'aide à la décision en entreprise.* » Ici on mettra de côté les plateformes telles que data.gouv et data.culture.gouv qui exposent les données publiques des administrations et vers lesquelles la BnF s'est naturellement d'abord tournée pour publier, afin de ne retenir dans notre étude que les plateformes qui publient les données de la recherche.

Les entrepôts de données de la recherche ou *Data repositories* sont des services en ligne permettant le dépôt, la description, la conservation, la recherche et la diffusion de jeux de données. Leur rôle principal est de faciliter le dépôt ou la collecte de données, leur description, leur accès et leur partage en vue de leur réutilisation ultérieure.

Il existe différentes catégories d'entrepôts de données : institutionnels, pluridisciplinaires, disciplinaires et thématiques. Un entrepôt institutionnel peut être mis en place et géré au niveau d'un établissement de recherche ; on peut citer en France Dataverse CIRAD, DataSuds (Entrepôt de l'IRD), data.sciencespo (Sciences Po). Parmi les entrepôts multi-disciplinaires figurent Figshare, Zenodo (développé dans le cadre du projet européen OpenAIRE), Dryad, Nakala (structure pour les Sciences humaines et sociales, mise en place par l'IR* Huma-Num) ou Progedo (dédié aux enquêtes sociales). Depuis juillet 2022, la France dispose d'un entrepôt de données généraliste à l'échelle nationale, Recherche Data Gouv. Il s'agit d'une solution souveraine pour le partage et l'ouverture des données de recherche produites par les communautés qui ne disposent pas d'un entrepôt disciplinaire reconnu.

Chaque entrepôt de données a sa propre politique concernant le dépôt, la description et la diffusion des données. Différents modèles de dépôt se distinguent, avec parfois des frais de publication. Le premier consiste à accepter tous les dépôts, c'est notamment le cas de Zenodo. D'autres modèles n'acceptent des jeux de données que s'ils sont en lien avec une publication, comme Dryad[56], ou bien acceptent uniquement certains types de données, en se concentrant sur une discipline, un domaine de recherche ou encore un projet particulier.

Le choix d'un entrepôt de données sera guidé par les pratiques d'une communauté scientifique, par la politique d'un établissement et la disponibilité ou non d'un entrepôt institutionnel dédié[57]. Les caractéristiques et fonctionnalités à prendre en compte sont diverses. Il convient en premier lieu de choisir un entrepôt en fonction du public que l'on vise. Pour déposer des données de la recherche, on se tournera plutôt vers l'un des entrepôts dédiés que vers les entrepôts diffusant des données publiques administratives tels que data.gouv. L'entrepôt répondra par ailleurs aux recommandations de l'institution dans laquelle la publication est faite, il sera reconnu dans la discipline et par la communauté scientifique par le biais d'une certification.

Les fonctionnalités centrales relèvent d'abord du dépôt et de la conservation des données, c'est à dire d'une part l'import de fichiers (formats et taille des fichiers,

---

[56] Entrepôt Dryad. https://datadryad.org/stash

[57] M. PAIN, *Les données de la recherche et leurs entrepôts, de la documentation à la réutilisation : étude de cas pour l'archive HAL*, Enssib, 2016. https://memsic.ccsd.cnrs.fr/mem_01374509.

téléchargement via une API), d'autre part l'organisation des données en collections, et enfin la conservation en vue du stockage sécurisé et de l'archivage à long terme.

Autre caractéristique d'un entrepôt de données : l'identification pérenne des données qu'il propose, c'est-à-dire l'attribution, au jeu dans son ensemble, ou à chaque version du jeu de données d'un identifiant, au moment du dépôt (voir paragraphe 3.1.5).

Parmi les caractéristiques relevant de la préparation et de la publication, la description des données est une étape importante. La précision des métadonnées doit pouvoir reposer sur un schéma générique ou spécifique à un domaine proposé par l'entrepôt, leur qualité doit pouvoir s'appuyer sur des vocabulaires contrôlés[58], et leur saisie être réalisée via des API. La documentation constitue également une description de données complémentaire indispensable.

En matière d'accessibilité, le contrôle des droits d'accès aux données, les conditions d'utilisation et licences offrent selon les entrepôts, des possibilités diverses : téléchargement libre (*open*) fermé (*closed*), embargo, demande d'accès (restreint), génération d'url privée parfois proposé, attribution d'une licence à chaque jeu de données en saisie libre ou liste fermée.

Les fonctionnalités de recherche, d'affichage et d'export des métadonnées doivent être prises en compte. La recherche dans les métadonnées et les données peut être simple et/ou avancée, ou affinée par facettes. L'affichage des métadonnées doit permettre de faire le lien avec d'autres *datasets*, avec des publications, et de générer de la citation. L'affichage des données par prévisualisation selon le format, est avantageuse quand elle existe. Les formats d'exports de métadonnées et de données en vue du téléchargement sont aussi des critères importants. Une autre fonctionnalité importante réside dans l'exploration et la visualisation des données. Des outils dédiés à ces fonctions peuvent être mis à disposition directement via l'entrepôt.

Enfin, les entrepôts élargissent la visibilité des données car ils sont en lien avec les autres dispositifs numériques tels que les archives de données, annuaires de données et intégrateurs. Ils sont scannés par des outils de recherche spécifiques comme *Data Cite search*, *Data Citation Index, Google Dataset Search*, moissonnés par des catalogues, intégrateurs, infrastructures européennes de donnée de plus en

[58] Le vocabulaire contrôlé permet de représenter un concept par un seul mot ou une seule expression sélectionnée dans une liste prédéfinie.

plus nombreux tels que l'EOSC[59] et OpenAire, et peuvent diffuser leurs données via le protocole d'échange standard OAI-PMH. Ils répondent aux besoins d'évaluation de l'usage des données en mettant en place des statistiques de consultation et de téléchargement des *datasets*[60].

### 2.3.2. Les entrepôts FAIR

Le 2ème axe du PNSO, « Structurer et ouvrir les données de la recherche », énonce la recommandation aux chercheurs de déposer leurs données dans des entrepôts certifiés dont la gouvernance et les règles de propriété intellectuelle seront conformes aux bonnes pratiques. Ces entrepôts qui appliquent les principes directeurs de la science ouverte et qu'on peut à ce titre qualifier de FAIR, visent une certification de type *CoreTrustSeal* qui atteste de la confiance que peuvent accorder à l'entrepôts utilisateurs et gestionnaires de données. Cette certification permet de garantir un bon niveau de compatibilité des données avec les principes de trouvabilité, d'accès, d'interopérabilité et de réutilisation, les prérequis de la certification étant compatibles avec les principes FAIR[61].

Pour les entrepôts ne disposant pas de certification, certains critères permettent d'évaluer leur niveau de compatibilité avec les principes FAIR[62]. Il s'agit en premier lieu de l'attribution d'identifiants uniques et pérennes comme le DOI (voir paragraphe 3.1.5) aux jeux de données et/ou aux fichiers composant les jeux de données. Un autre critère réside dans la possibilité offerte de documenter les données avec des métadonnées (auteurs, description du contenu du jeu de données, publications associées, etc.) ainsi qu'avec des informations permettant de mieux comprendre et utiliser les données (définition des variables, logiciels associés, provenance, etc.), et encore d'assurer une bonne indexation des métadonnées pour permettre leur recherche. Un entrepôt FAIR donnera également la possibilité de mentionner clairement la licence (licence ouverte, CC BY, etc.) ou les conditions

---

[59] European Open Science Cloud

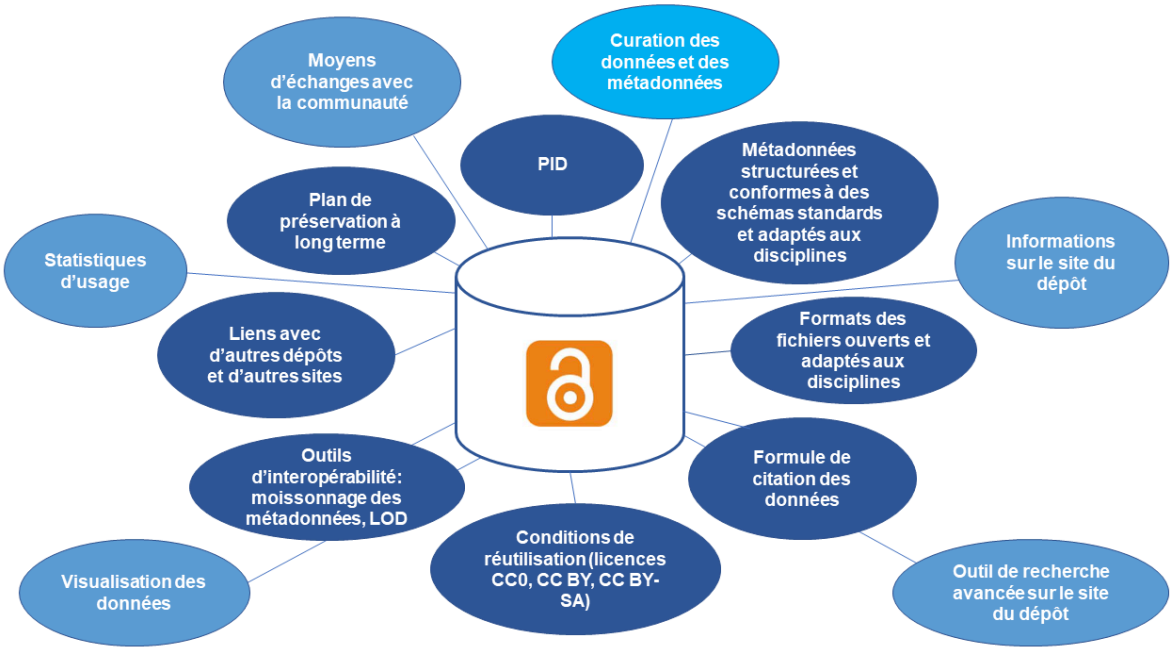
[60] J. DESCONNETS, *Les entrepôts de données ou comment rendre les données trouvables, accessibles et réutilisables ?* Site de SIST le réseau des gestionnaires de données d'observation. https://sist.cnrs.fr/wp-content/uploads/2021/12/SIST20-21-02-J-Desconnets-Entrepots-de-donnees.pdf

[61] *Certification des entrepôts et services de données*. Site Ouvrir la science. https://www.ouvrirlascience.fr/certification-des-entrepots-et-services-de-donnees/?menu=4

[62] *Les entrepôts FAIR*, Site de l'INRAE, Institut national de recherche pour l'agriculture, l'alimentation et l'environnement. https://datapartage.inrae.fr/Produire-des-donnees-FAIR/Comment-FAIR-en-pratique/Les-entrepots-FAIR

spécifiques sous lesquelles les données sont utilisables. Il rendra accessibles publiquement les citations et les métadonnées, même dans le cas de jeux de données dont les fichiers associés sont à accès restreint, il utilisera un standard de métadonnées reconnu, et disposera d'un plan de préservation à long terme des données.

Figure 2 : Les outils et services d'un dépôt de données de recherche ouvertes[63]




## 2.3.3. Au-delà du stockage, pourquoi et comment exposer dans un entrepôt ?

Déposer ses données dans un entrepôt assure leur préservation, leur visibilité et leur accès, facilitant ainsi leur partage et leur réutilisation. En outre, le crédit reçu en tant qu'auteur ou contributeur crédité des données diffusées, permet d'accroître sa notoriété. Le dépôt apporte ainsi de nombreux avantages tels que la conservation des données dans un environnement sécurisé, leur visibilité et leur accès facilité pour

[63] M. GUIRLET, *Guide décisionnel et vade-mecum pour la mise à disposition d'un dépôt de données de recherche ouvertes en Suisse.* – Mémoire d'étude version révisée, 18 décembre 2020. https://zenodo.org/records/4357134

les moteurs de recherche. Il permet également l'interopérabilité des données grâce à l'utilisation de standards de métadonnées, la découverte, la réutilisation et la citation du jeu de données facilitées par son identifiant pérenne. Enfin il offre la possibilité de gérer les modalités de partage des données par l'attribution de licences de diffusion, le respect des recommandations des financeurs et institutions sur l'ouverture des données, une intégrité et une validation scientifique améliorées, ainsi que la valorisation des données par leur réutilisation dans de nouvelles études et innovations[64].

Lors du dépôt, les informations supplémentaires que sont les métadonnées sont saisies ou collectées. Elles sont destinées à faciliter la compréhension et l'interprétation des données déposées, par exemple en précisant leur couverture géographique et temporelle. En plus des métadonnées standards telles que celles du format *Dublin Core*, qui permettent de décrire l'auteur, le titre, l'année de création, d'un jeu de données, un entrepôt propose également des métadonnées spécifiques au sujet, au thème ou à la discipline des données qu'il accueille. Par exemple, cela peut concerner des données biologiques, astronomiques, ou environnementales.

Lorsque les jeux de données sont déposés dans un entrepôt, les responsables conservent leurs droits moraux d'auteur pour caractériser ces jeux de données. En plus des métadonnées, il est recommandé de fournir une documentation complémentaire (comme un fichier *lisez-moi* ou *read me*[65], un sommaire, le contexte de collecte, la méthodologie, le traitement, et les limites des données), et d'effectuer une indexation en utilisant des mots-clés provenant de référentiels. De plus, lorsque les jeux de données sont nombreux, il est utile d'organiser les fichiers selon des normes préétablies pour faciliter la navigation dans la documentation. Cela peut impliquer l'adoption de conventions pour la structure des fichiers, la transcription des entretiens ou des observations, ainsi que le nommage cohérent des fichiers et des dossiers. L'organisation des fichiers peut également améliorer l'intelligibilité des données en les distinguant selon leur type, leur temporalité, leur origine géographique, etc.

Enfin, il est conseillé aux chercheurs d'envisager dès le début du projet la diffusion des données, notamment dans le cadre d'un Plan de Gestion des Données

[64] *Déposer ses données de recherches dans un entrepôt*. Site du CIRAD. https://coop-ist.cirad.fr/gerer-des-donnees/deposer-des-donnees-dans-un-entrepot/3-pourquoi-deposer-des-donnees-dans-un-entrepot

[65] *#RechercheDataGouv | Bien décrire ses données pour les valoriser : un modèle de README à votre disposition*. Site CNRS INIST. https://lalist.inist.fr/?p=60037

(PGD), afin d'anticiper au mieux tous ces aspects. Les professionnels de l'IST sont amenés à contribuer, en lien avec les ressources qu'ils mettent à disposition, à ces PGD.

Certains entrepôts de données prennent en charge tout ou partie de la documentation des jeux de données, tandis que d'autres accompagnent les chercheurs dans le processus de documentation et vérifient la conformité à la législation lors de la diffusion. Certains entrepôts pratiquent également l'auto-dépôt, ce qui signifie que les producteurs sont libres de partager leurs jeux de données sans qu'un examen préalable soit nécessaire de la part de l'entrepôt.

## 2.4. Des riches données de la BnF aux besoins et pratiques des chercheurs.

### 2.4.1. Des données riches et ouvertes

Les collections de la BnF sont d'une grande diversité, comprenant des documents numérisés, des documents nativement numériques, des supports numériques tels que des CD, DVD, jeux vidéo, des archives de l'Internet, ainsi que des métadonnées bibliographiques et des données d'autorité. Les métadonnées, issues de ressources variées, se distinguent également par leur structuration et leur format, selon les collections qu'elles décrivent et les usages prévus. « *Par exemple, Intermarc est utilisé pour le Catalogue général, XML-EAD*[66] *pour le catalogue BnF Archives et manuscrits, Dublin Core*[67] *pour la bibliothèque numérique Gallica, RDF*[68] *pour le portail data.bnf.fr, et WARC pour les archives du web.* [..] »[69]. D'autres métadonnées constituent à présent des données de la recherche, il s'agit des logs de connexion aux serveurs de Gallica, qui enregistrent l'activité des utilisateurs et permettent de suivre leurs actions. Elles relèvent du « Règlement Général sur la

[66] L'EAD (Encoded Archival Description), est un format basé sur le langage XML qui permet de structurer des descriptions de manuscrits ou de documents d'archives.

[67] Le format Dublin Core, est un format de description simplifié (15 champs) et un langage du web sémantique (format obligatoire dans le cadre du protocole OAI-PMH) qui permet d'exprimer les données dans un modèle RDF. Le Dublin Core est utilisé par Gallica et RDF par DataBnF.

[68] Langage de base du web sémantique, RDF (Resource Description Framework) est un modèle de données destiné à décrire les ressources web et leurs métadonnées sous forme de triplet (sujet, prédicat, objet).

[69] A. LABORDERIE, et F. TFIBEL, 2023. « Ouvrir les données de la Bibliothèque nationale de France pour la recherche.» *Culture et recherche*. 1 juin 2023. N° 144. https://bnf.hal.science/hal-04074665

Protection des Données » (RGPD), aussi ont-elles été anonymisées pour des raisons juridiques et éthiques, afin de les rendre accessibles aux chercheurs.

A son lancement en 1997, Gallica proposait quelques milliers d'œuvres. Depuis lors, les collections numériques de la BnF dont le volume estimé à plus de 6 pétaoctets[70] explose aujourd'hui relèvent de conditions d'utilisation particulières[71] qui leur permettent d'être accessibles et téléchargeables gratuitement, partagées et utilisées librement dans un cadre non commercial sous réserve de la mention de la source, ce qui revient à couvrir de très nombreux usages. Seule la réutilisation commerciale entraîne une redevance. Dès 2014, la BnF s'était engagée dans cette démarche d'ouverture en mettant toutes ses métadonnées sous licence Etalab, la licence publique de l'État, qui permet de déclarer un document comme appartenant au domaine public, et donc revient à abandonner tous droits de propriété intellectuelle le concernant. L'utilisation de ces métadonnées est libre et gratuite sous réserve du maintien de la mention de leur source et de l'indication de leur date de récupération[72].

L'ouverture de la licence sur les contenus de Gallica a reposé sur une politique d'*open data* qui maintenait en France le caractère exceptionnel des données culturelles en autorisant des redevances pour la fourniture des données culturelles. La BnF a de nouveau marqué son engagement à partager ses collections en faisant évoluer, en 2019,[73] la politique de diffusion et d'utilisation de ses millions d'images numérisées. Elle l'a assouplie, en exonérant les chercheurs de la redevance d'utilisation commerciale, en permettant la récupération des fichiers en haute définition, et en proposant une grille tarifaire repensée à la baisse. La BnF a également répondu aux nouvelles pratiques nées de l'ère du partage numérique en permettant la récupération des fichiers HD libre et gratuite via les API de Gallica.

Les contenus accessibles sur le site Gallica sont pour la plupart des reproductions numériques d'œuvres tombées dans le domaine public provenant des collections de la BnF. Une autre part des données d'intérêt pour la recherche réside dans les documents sous droits. Gallica *intra muros,* qui ajoute aux ressources disponibles en ligne sur Gallica plusieurs centaines de milliers de contenus protégés

[70] Ibid.

[71] Conditions d'utilisation des contenus de Gallica. https://gallica.bnf.fr/edit/und/conditions-dutilisation-des-contenus-de-gallica

[72] Conditions de réutilisations des données de la BnF. Site de la BnF https://www.bnf.fr/fr/conditions-de-reutilisations-des-donnees-de-la-bnf

[73] *La BnF change sa politique d'accès à ses images.* 2 octobre 2019. https://www.bnf.fr/sites/default/files/2019-10/CP_politique_diffusion_images_BnF.pdf

au titre de la propriété intellectuelle, numérisés par la BnF et ses partenaires, n'est à ce titre consultable que dans les salles de lecture de la BnF. Les archives de l'Internet français, collectées par la BnF depuis 2006, et pour la plupart librement accessibles en ligne, acquièrent, dès lors qu'elles sont collectées par la BnF dans le cadre du Dépôt légal, un statut de données patrimoniales avec des restrictions d'accès. Aussi ne sont-elles consultables que dans les salles de lecture BnF ou dans les bibliothèques de dépôt légal imprimeur (BDLI).

Chacune des catégories de données de la BnF présente des structures, des formats, des qualités, des contextes de production, des fonctions et des contenus différents, ce qui exige des traitements spécifiques et nécessite des compétences et des expertises particulières. La BnF s'attache donc non seulement à mettre à disposition ces collections, mais également à faciliter leur compréhension et leur manipulation en mettant en place une gamme d'outils dédiés : interface de recherche permettant de constituer des corpus selon la proximité, l'occurrence ou la similarité des documents, portail API et jeux de données pour l'extraction des documents, et recherche avancée dans les métadonnées via data.bnf.fr et le SPARQL endpoint[74].

Il s'agit, grâce à ces efforts d'accessibilité, d'inciter les auteurs à diffuser librement eux aussi leurs travaux dans l'esprit de la science ouverte.

### 2.4.2. Besoins et pratiques des communautés de recherche

Émanant des préoccupations sans cesse renouvelées des bibliothécaires de resserrer les liens avec les chercheurs et ajuster leurs services, les enquêtes menées auprès des communautés de recherche font remonter une grande variété de pratiques de gestion et de partage de données.

Tout d'abord, les questions au sujet de leur parcours dans la collecte et de dépôt de données ne font pas apparaître d'habitudes suffisamment récurrentes de leur part pour qu'on puisse en déduire les endroits stratégiques où diffuser des jeux de données. Ensuite, ils évoquent davantage le dépôt de leurs publications, et envisagent le dépôt de leurs données plutôt pour des raisons de stockage que dans une perspective de réutilisation. La diffusion de données dans un mode plus ouvert reste à développer dans leurs pratiques. Violaine Rebouillat, dans une thèse qu'elle a

[74] Le SPARQL est un langage de requête permettant d'interroger le RDF. Le SPARQL endpoint de la BnF permet d'interroger sa base de données data.bnf.fr.
https://api.bnf.fr/fr/sparql-endpoint-de-databnffr

consacrée au sujet, fait apparaître que « *L'évaluation et la reconnaissance des activités scientifiques étant principalement fondées sur la publication d'articles et d'ouvrages, la gestion et l'ouverture des données ne sont pas considérées comme prioritaires par les chercheurs.* »[75]

Autre difficulté questionnée à travers les entretiens menés au coeur de ces enquêtes, les services d'ouverture des données sont fournis par des acteurs qui ne sont pas directement liés aux communautés de recherche ; il se pourrait que les chercheurs soient plutôt influencés par des réseaux spécifiques à leur domaine d'étude, tels que les revues scientifiques ou les infrastructures spécialisées. Ces conclusions amènent à réévaluer la question de la médiation.[76] Le problème de communication et de temporalités différentes entre bibliothèques et chercheurs, bien identifié, rencontre de nombreuses difficultés : temps du dialogue chronophage, mobilité géographique inhérente à la professiondes d'enseignants-chercheurs, instabilité des postes croissante. Pourtant, le nomadisme du bibliothécaire, sa capacité à dispenser des formations dans les laboratoires ou les bureaux des chercheurs, est bien ce qui permet aujourd'hui le plus souvent de construire la relation avec le public de niveau recherche.

Les besoins en formation aux outils de recherche documentaire et d'extraction, eux, sont clairement identifiés par les chercheurs. L'étude de besoin des publics réalisée dans le cadre du projet Corpus[77] l'avait nettement mis en évidence, et la dernière enquête réalisée en 2020 auprès de l'ensemble des publics de la BnF l'a montré à nouveau. Il y est ressorti qu'un tiers des répondants s'inscrivait dans une démarche académique (étudiants, doctorants, professionnels de la recherche en activité ou en retraite), et que la connaissance des usages avancés de Gallica par les chercheurs restait faible. « *[..] le rapport à Gallica est marqué par une double dimension de l'expertise liée, d'une part, à la recherche documentaire (le Catalogue comme point d'entrée, la constitution de corpus et la recherche systématique) et, d'autre part, à la maîtrise du numérique, à la capacité à explorer l'interface (utilisation experte du moteur de recherche, formulation pertinente puis affinage des requêtes et le fait d'avoir des routines de recherche).* »[78]

---

[75] V. REBOUILLAT, *Ouverture des données de la recherche : de la vision politique aux pratiques des chercheurs*. thèse de doct. - CNAM. 2019.https://theses.hal.science/tel-02447653

[76] P. CORMIER, *Le positionnement des bibliothèques universitaires et de recherche françaises dans les politiques publiques des données de la recherche*. op. cit.

[77] Voir paragraphe 1.2.3

[78] I. BASTARD et A. LABORDERIE, 2023. « La découvrabilité des collections numériques patrimoniales sous l'angle des usages de Gallica », Bulletin des bibliothèques de France (BBF), 14 juin

Le projet Corpus a montré les attendus des chercheurs, et permis de les mettre en place. L'étude a mis à plat le besoin qu'ils ont exprimé d'un système combinant des machines virtuelles et une plateforme sécurisée, accessible depuis les locaux de la bibliothèque. Cette plateforme vouée à offrir une gamme d'outils, capable de gérer différents niveaux d'autorisations, pour permettre l'installation de logiciels, fournissant une documentation complète, des tutoriels sous forme de MOOC, des exemples d'utilisation, ainsi qu'un accès via des connecteurs à des corpus pré-constitués tels que des API et des jeux de données, est désormais à disposition, et l'essentiel reste aujourd'hui de la faire connaître. En effet, les chercheurs ou ingénieurs d'étude qui connaissent les outils d'accès aux données de la BnF l'affirment : c'est à travers leur présence régulière au DataLab et la fréquentation côte à côte des professionnels bibliothécaires impliqués dans l'accompagnement à la recherche, qu'ils ont pu découvrir les outils et apprendre à s'en emparer.

Au-delà de ces besoins de formation, d'accompagnement et de mise à disposition d'outils, le projet Corpus avait permis de poser une autre question d'importance, celle de savoir si la BnF devait ingérer les résultats de la recherche. Une partie des universitaires et des agents de la BnF interviewés y avait vu l'intérêt réciproque de la conservation des résultats en tant qu'enrichissement ou correction des données existantes, tels que l'ingestion d'un OCR corrigé, d'une édition numérique encodée en XML/ TEI, d'annotations d'images.

« *Cela demanderait à la Bibliothèque non seulement l'établissement de nouvelles chaînes d'entrée, mais aussi une validation. Pour alléger le travail chronophage d'un processus de validation, un fonctionnement par calques permettrait par exemple d'ingérer dans Gallica les données produites par les laboratoires de recherche ou d'autres institutions mais en les identifiant clairement comme produites à l'extérieur de l'établissement et en les proposant manifestement comme ajouts hors BnF.* [79]»

Pour la BnF, réintroduire ces enrichissements dans ses propres données reste encore un défi. Les partenariats avec Huma-Num et DARIAH marquent la volonté de valoriser ces résultats de la recherche. Les outils développés dans le cadre de programmes de recherche ou d'appels à projets sont conservés dans une sorte de

---

2023. https://bbf.enssib.fr/matieres-a-penser/la-decouvrabilite-des-collections-numeriques-patrimoniales-sous-l-angle-des-usages-de-gallica_71295l

[79] E. MOIRAGHI, *Le projet Corpus et ses publics potentiels : Une étude prospective sur les besoins et les attentes des futurs usagers*, *op. cit.*

boîte à outils et réutilisables pour d'autres projets. Le portail api.bnf, qui offre une réponse à ces attentes depuis 2017 en publiant des données enrichies (correction, annotation, modèles, etc.) produites par des projets de recherche, parvient-il à répondre à ces besoins ?

## 2.6. Le portail api.bnf, état des lieux et perspectives

### 2.6.1. Une mine d'informations, qui restent aujourd'hui peu visibles

Suite à l'adoption en 2014 de la Licence ouverte pour l'ensemble de ses métadonnées, la BnF a poursuivi sa stratégie d'ouverture des données en lançant son portail API et jeux de données en 2017 à l'occasion du deuxième hackathon[80]. Le portail a pour vocation de « simplifier l'accès à ces données et de susciter de nouveaux usages (alimentation de catalogues, création d'applications innovantes, fouille de données, datavisualisation, etc.) auprès de publics professionnels diversifiés (développeurs et développeuses, entrepreneurs et entrepreneuses, acteurs et actrices de la culture et de la chaîne du livre, chercheurs et chercheuses, *digital humanists*) ou tout simplement des amateurs et amatrices de culture. »

Le portail présente d'abord un accès par Sources de données qui fait découvrir le panorama complet des produits issus de chacune des huit grandes sources de données. Un onglet *Utilisations* permet de présenter des projets de recherche partenaires, de valoriser des *expérimentations* produites dans ce cadre, et donne également un accès *Tutoriels et outils,* contenant la documentation de tous les formats de diffusion : les API IIIF d'affichage et de récupération des images de Gallica, les API d'interrogation des métadonnées de Gallica et du Catalogue général (SRU[81], Z39.50[82], OAI-PMH) et le SPARQL endpoint de data.bnf.fr, ainsi que des jeux de données qui constituent des extractions d'ensembles cohérents des catalogues (produits bibliographiques, *dumps*[83] de data.bnf.fr, du CCFr, de Mandragore[84] et Reliures[85]), des corpus documentaires réalisés et enrichis dans le

[80] Portail api.bnf.fr. https://api.bnf.fr/fr/a-propos

[81] Le service SRU (Search/Retrieval via URL) du Catalogue général permet d'interroger le Catalogue général de la BnF. https://www.bnf.fr/fr/service-sru-catalogue-general-de-la-bnf

[82] Le serveur Z39.50 de la BnF permet de récupérer des métadonnées du Catalogue général de la BnF. url

[83] Le mot *dump* est un anglicisme qui désigne, selon la définition du Wiktionnaire, un cliché ou vidage de mémoire, une copie du contenu d'une mémoire sur un autre support.

[84] Base iconographique pour les manuscrits de la BnF. https://mandragore.bnf.fr/

[85] Base des reliures numérisées de la BnF. https://reliures.bnf.fr/

cadre de projets de recherche, et enfin des pages éditoriales destinées à guider les utilisateurs dans la découverte et le choix des modes de récupération des données qui conviennent à leurs besoins.

Le Département des métadonnées, en charge de la coordination du portail, réunit une à deux fois par an lors d'un comité éditorial différents contributeurs issus de différentes directions, parmi lesquels figurent entre autres le Département de la coopération, le Département du dépôt légal, le Service du dépôt légal numérique et le Département des manuscrits. D'autres contributeurs, tels que les Départements scientifiques l'alimentent ponctuellement et diversifient la production de contenus, ce qui pose des questions de charge de travail et d'organisation, et implique le maintien d'une convergence de visions.

Fruit d'une réflexion collective vouée à améliorer le site, une deuxième version d'api.bnf.fr, mise en ligne en 2020, a offert une ergonomie repensée, avec notamment l'ajout de critères (sources de données, formats, sujets) destinés à faciliter la recherche. Le portail utilise depuis cette deuxième version le standard DCAT qui vise à normaliser la description des catalogues d'informations publiques. Cette implémentation rend aujourd'hui le portail api.bnf potentiellement interopérable avec d'autres plateformes de données[86].

Via l'outil Piano qui mesure l'audience des sites web de la BnF, des statistiques de fréquentation par page, disponibles depuis 2021, indiquent un trafic direct via l'url du site majoritaire, ce qui donne à penser que le site est encore méconnu. Le diagramme ci-dessous, fait état de la fréquentation d'une dizaine de jeux de données de nature différentes, et permet de dégager quelques catégories de contenus davantage consultés. Les *dumps* de data.bnf[87] et la liste des adresses url des collectes[88] sont les plus consultées.

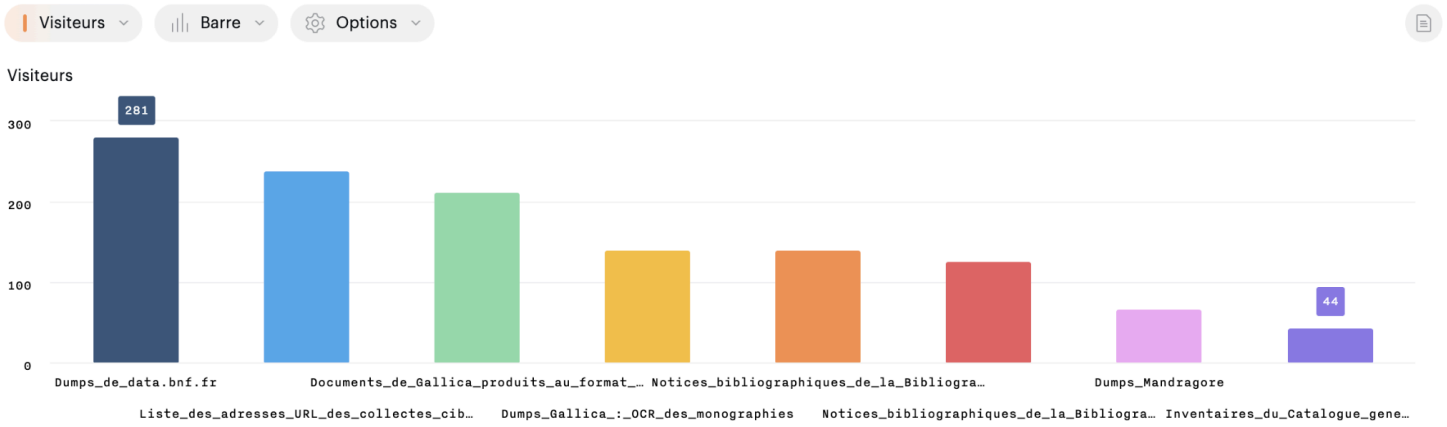
Figure 3 : Statistiques de fréquentation de quelques jeux de données [89]

[86] Voir paragraphe 3.4.1

[87] Portail api.bnf.fr. *Dumps-de-databnf*. https://api.bnf.fr/fr/dumps-de-databnffr

[88] Portail api.bnf.fr. *Liste des-adresses url des collectes ciblees du web francais* https://api.bnf.fr/fr/liste-des-adresses-url-des-collectes-ciblees-du-web-francais-par-la-bnf

[89] Diagramme conçu avec l'outil statistique Xiti/Piano à partir des données de fréquentation du portail api.bnf.fr.




**Légende :**

Dumps de data.bnf

Liste des adresses url des collectes ciblées

Documents de Gallica produits au format TEI

Dump Gallica : ocr des monographies

Notices_bibliographiques_de_la_Bibliographie_nationale_francaise

Dumps Mandragore

Inventaire du Catalogue général des manuscrits

Le site globalement reste trop peu visité et peu connu des publics, plusieurs éléments faisant obstacle à sa fréquentation. Sa barre de recherche n'est pas fonctionnelle, et l'accumulation croissante de jeux de données sans révision de l'éditorialisation génère un manque de lisibilité. Un onglet dédié aux services se résume pour le moment à la diffusion de l'adresse contact à laquelle il est possible de demander une aide à la requête dans les données, ainsi que des produits à façon.

Le site est créé sous Drupal, un système de gestion de contenu flexible qui permet d'obtenir un format fonctionnel et évolutif sans recourir systématiquement à la programmation. Cette enveloppe éditoriale présente des limites importantes comparé aux solutions d'exposition offertes par les entrepôts de données. Les fichiers sont eux mêmes déposés sur la plateforme d'échange de fichiers (PEF) de la BnF, conçue non comme pas un espace de stockage pérenne mais comme un outil de

diffusion pour transférer les fichiers. Ce dispositif a posé problème, le téléchargement depuis le portail API n'étant plus fonctionnel depuis un an car le protocole FTP de la PEF n'était plus supporté par les navigateurs. La réponse à ce problème est en train de se mettre en place par le moyen de licences GoDrive attribuées à chaque département producteur de jeux de données. Il s'agit d'une plateforme sécurisée qui crypte les données stockées et permet de mettre à disposition les fichiers sous la forme de liens via le protocole HTTPS.

Afin de répondre à sa faible utilisation actuellement, plusieurs pistes d'évolution du portail sont actuellement à l'étude. La première consiste à faire évoluer la charte éditoriale du site. Ce travail reposera sur l'évaluation des besoins des différentes équipes contributrices pour s'accorder sur un périmètre prévisionnel, interroger l'articulation avec les autres lieux de diffusion, puis mettre en commun pour formaliser l'évolution de la charte documentaire souhaitée, afin de créer ensuite les nouveaux contenus et modèles. Par exemple est envisagé l'ajout de nouveaux contenus attendus par les bibliothèques utilisatrices des produits bibliographiques, les documentations de conversion de formats, ou bien la documentation de format de données spécifiques aux archives du web.

L'amélioration de la navigation, et de la recherche, sont souhaitées également, avec une possibilité de recherche avancée dans les métadonnées, des datavisualisations, des fonctions de recherche groupée dans toutes les listes API, la mise en place d'un bac à sable pédagogique pour faire des recherches en accès libre.

### 2.6.3. Identification et préservation des jeux de données : quelle articulation ?

Autre évolution envisagée, l'attribution d'ARK aux jeux de données permettrait d'augmenter leur visibilité. Les ARK sont les identifiants pérennes utilisés pour identifier les ressources du type documents numériques et notices (bibliographiques, d'autorité, etc) de la BnF. Une instruction technique préalable reste à mener afin d'identifier les jeux de données concernés et ceux qui ne le seraient pas. Un tableau qui fait état de tous les jeux de données existant sur le portail a vocation à recenser leur emplacement (PEF, data.gouv, data.culture.gouv,) leurs modalités d'accès (adresse url, fixe ou non...), leur durée de vie estimée, l'étape du cycle de vie dans laquelle ils se trouvent (suppression, scission ou fusion,

redirection, versions et devenir des anciennes versions…), le niveau de granularité de citation souhaité le cas échéant (identification pérenne du jeu de données, dans certains cas du fichier), leurs droits associés, les métadonnées associées souhaitées (informations minimales d'identification), leur besoin d'être préservés dans SPAR, le Système de Préservation et d'Archivage Réparti, un outil dont s'est dotée la BnF pour garantir la préservation des documents numériques.

De nombreuses questions d'une technicité élevée se posent autour de la pertinence d'identifier de façon pérenne et de préserver certains jeux de données, telles que l'articulation entre la politique de préservation dans SPAR et l'attribution d'ARK. Les deux besoins ne vont pas nécessairement de pair ; les données des *dumps* de la collecte du web sont déjà conservées dans SPAR, mais pourraient avoir besoin d'ARK pour être citables. Les limites techniques potentielles liées à l'intégration dans SPAR de données de poids très lourd comme les images, et de données aux formats peu pérennes, posent également question.

Pour résumer, trois grands critères de regroupement des jeux de données ont été proposés pour évaluer la pertinence de leur attribuer des ARK. Il s'agit d'abord de la durée de conservation envisagée par le responsable du jeu de données et de la gravité d'une éventuelle perte des données. Si elle est illimitée, le jeu a vocation à rentrer dans SPAR. Ensuite c'est l'originalité qui fait également la différence. Par exemple, s'il s'agit d'ensembles de notices ou de dumps de collectes du web, il est normalement possible de les générer à nouveau à partir des données de SPAR, l'originalité du jeu est donc dans ce cas moins importante. Dernier critère, les informations concernant la périodicité des mises à jour et le sort des versions antérieures permettront d'interfacer l'api ARK en gérant les versions, et de déterminer les développements nécessaires du résolveur, notamment la création d'un préfixe spécifique pour les jeux de données. Le résolveur est un site Web spécialisé dans la réorientation d'identifiants entrants vers les sites Web actuellement les mieux à même de les traiter. Ce transfert est généralement appelé
« résolution » et une des étapes de ce processus de résolution est la « redirection ». Pour qu'un résolveur fonctionne, son nom d'hôte - « ark.bnf.fr » pour les ARKS de la BnF -, doit être soigneusement choisi afin qu'il ne soit plus nécessaire de le changer.[90]

[90] Site de la BnF. https://www.bnf.fr/fr/lidentifiant-ark-archival-resource-key

Aujourd'hui, pour la BnF qui fait face à de nombreux projets de développement numériques de ses fonds, faire connaître les jeux de données primaires ou dérivés produits à partir de ses collections, peut relever de différentes stratégies complémentaires entre elles, et la diffusion dans un entrepôt de données extérieur est une solution à approfondir.

## 2.7. Une sélection d'entrepôts pour diffuser les jeux de données de la BnF

### 2.7.1. Entrepôts *open data* de données publiques

Dans la perspective de rechercher un entrepôt adapté pour publier les données de la BnF, une sélection d'entrepôts a fait l'objet d'une étude plus particulière. Deux types d'entrepôts de données y figurent ; les premiers, - data.gouv et data.culture -, sont dédiés aux données publiques françaises, et les suivants — Nakala, Zenodo, Recherche.data.gouv — sont dédiés aux données de la recherche.

La plateforme de diffusion de données publiques *open data* de l'État français data.gouv, développée par la mission Etalab en 2011, est un catalogue des données publiques de l'administration. Vouée à offrir un matériau de données numériques permettant d'éclairer les politiques locales, dans l'idée des *smart cities* ou « villes connectées », elle autorise le dépôt et le référencement de données issues de ministères comme de collectivités locales. Il s'agissait du premier outil proposé par les politiques publiques en matière d'*open data* pour ouvrir les données publiques, aussi la BnF s'est-elle naturellement tournée vers cette plateforme pour y exposer ses jeux de données.

La plateforme data.culture lancée par la suite par le ministère de la Culture en lien avec data.gouv.fr, est un espace de valorisation des ressources culturelles numériques du ministère et de ses établissements, permettant une visualisation lisible via des cartographies et graphiques. Les données de data.culture.gouv concernent aussi bien l'activité des institutions culturelles (fréquentation des musées) que des aspects économiques (aides à la presse), des ressources iconographiques (fonds de la guerre 14-18) ou encore des évènements culturels (Journées européennes du patrimoine). Il s'agit d'un produit *Open Data soft*[91]

[91] Plateforme Opendatasoft. https://www.opendatasoft.com/fr/

payant, avec un abonnement qui dépend de la volumétrie des jeux de données. La BnF a commencé à déposer ses données dans les débuts de la plateforme, puis le ministère de la Culture a voulu maîtriser les coûts et décidé, en 2018, que l'entrepôt était réservé aux seules administrations centrales et aux services à compétence nationale. Il est demandé aux établissements publics sous tutelle du ministère comme la BnF de déposer sur data.gouv, sachant que les jeux publiés sur data.gouv sont référencés automatiquement sur data.culture via un moissonnage. Aujourd'hui une quinzaine de jeux de données BnF sont déjà présents sur les deux plateformes. Il s'agit de *dumps* de data.bnf, des collectes du web, du catalogue général, de la base Palme (manuscrits littéraires), base des reliures numérisées, indicateurs du dépôt légal.

Poursuivre la publication de certains jeux de données BnF sur data.gouv s'avère intéressant, car avec l'implémentation du schéma DCAT dans la deuxième version d'api.bnf, une publication automatique de ces jeux est désormais envisageable, à la condition de créer un moissonneur sur le site[92]. Une fois la configuration du moissonneur validée par l'équipe en charge de data.gouv.fr, le moissonneur de data.gouv.fr vient automatiquement récupérer les données de la plateforme demandeuse. Par ailleurs, les perspectives d'évolution de la plateforme en 2023 permettent d'envisager dans un avenir prochain un meilleur accès aux métriques importantes pour les usagers[93], en fournissant directement des statistiques d'usage sur les pages de jeux de données, de réutilisations et d'organisations, ainsi que la mise à disposition d'un outil de publication assisté qui orientera notamment les producteurs dans la documentation de leurs données afin d'augmenter la qualité des données déposées. La plateforme data.culture.gouv, elle, n'envisage la possibilité de publier directement sur son site pour les d'établissements publics sous tutelle, qu'à la condition de mettre en place avec eux une gouvernance prévoyant la mutualisation de contributions financières de chacun au produit *Opendatasoft*.

### 2.7.2. Entrepôts *open science* de données de la recherche

Au vu du type de données diffusées par la BnF et des usages qui en sont attendus, ce sont les entrepôts de données de la recherche qui nous ont davantage

[92] *Demander à data.gouv.fr de moissonner votre site*. Plateforme data.gouv.fr. https://doc.data.gouv.fr/jeux-de-donnees/demander-a-datagouvfr-de-moisonner-votre-site/

[93] *De nouvelles statistiques d'usage sur data.gouv.fr*. Plateforme data.gouv.fr. https://www.data.gouv.fr/fr/posts/de-nouvelles-statistiques-dusage-sur-data-gouv-fr/

intéressée dans le cadre de ce mémoire. Notre attention s'est portée sur des entrepôts généralistes et dédiés aux données de la recherche en Sciences humaines et sociales. Seuls nous intéressent les entrepôts multidisciplinaires car les données patrimoniales de la BnF ne relèvent pas d'une discipline en particulier. L'enjeu de cet examen des entrepôts s'est porté également sur la volonté de diffuser les données de la BnF au sein de ses communautés partenaires, celles de l'IR* Huma-num et de l'ERIC européen Dariah.

C'est en premier lieu l'entrepôt Nakala[94] qui a fait l'objet de notre attention. Mis en œuvre par Huma-Num, il est l'entrepôt le plus connu et identifié par la communauté des sciences humaines et sociales. Il offre un service de stockage sécurisé avec identifiant pérenne et accès interopérable OAI-PMH mais sans moteur de recherche. L'entrepôt souffre aujourd'hui d'un retard concernant ses fonctionnalités de recherche, et de l'absence de curation des données qui y sont déposées. La description de données y est restreinte, étant basée sur un ensemble minimal de champs de métadonnées inspirés du Dublin-Core. Par ailleurs, son indexation par le portail de découverte européen OpenAire qui moissonne des entrepôts en Open Access, ne fonctionne plus à ce jour.

Autre entrepôt de la sélection, Zenodo[95], hautement recommandé dans le cadre européen, est un répertoire de travaux de recherche et de données créé par OpenAIRE et le CERN. Grâce à OpenAIRE, il intègre les résultats de la recherche dans les lignes de reporting exigibles par les agences de financement telles que la Commission européenne. Les informations de citation sont également transmises à *DataCite* et aux agrégateurs universitaires.

C'est enfin l'entrepôt national Recherche Data Gouv (RDG)[96] ouvert au printemps 2022, et voué à accueillir toutes les données issues de la recherche française, qui nous a permis d'envisager une exposition d'un niveau différent. RDG est voué à accueillir des jeux de données en accès ouvert ou restreint et à référencer l'ensemble des données de recherche produites en France tout en créant des ponts avec les infrastructures existantes. L'un de ses objectifs est de faire de HAL une porte d'entrée vers la plateforme nationale des données qui ouvrirait un service de transfert des données accompagnant la publication, ce qui soulève la question des métadonnées à adopter, tant sur les plans standards que disciplinaires. RDG prévoit

[94] Site de Nakala. https://www.nakala.fr/
[95] Site de Cat OPIDoR. https://cat.opidor.fr/index.php/Zenodo
[96] Entrepôt national Recherche Data Gouv. https://recherche.data.gouv.fr/fr

également de mettre en place un module catalogue qui permettra le moissonnage d'entrepôts de données externes. L'entrepôt lui-même est développé à l'aide de la solution logicielle Dataverse et est hébergé dans un datacenter labellisé. Le projet est sous la supervision d'un comité de pilotage agissant au nom du Comité de pilotage de la science ouverte. Ce comité est chargé de proposer la feuille de route globale et de proposer des solutions et des moyens, tout en assurant la cohérence globale et la trajectoire du projet. L'INRAE, en collaboration avec le CNRS et les universités de Grenoble, Lorraine, Lille, Strasbourg, Paris Nanterre et Paris Cité, est responsable des modules entrepôt et catalogue[97]. « Actuellement financé par le FNSO, RDG aura pour défi futur d'établir la structure organisationnelle, la gouvernance et le modèle économique après 2023 »[98].

Face aux plus récents enjeux d'ouverture des données de la recherche, les entrepôts qui se développent répondent à de nouvelles méthodes d'exposition. La sélection d'entrepôts appliquée au profil institutionnel de la BnF et à la nature de ses données nous amène à la conclusion que RDG est celui qui offre aujourd'hui le plus de perspectives à nos yeux pour l'exposition des jeux de données de la BnF, Nakala et Zenodo devenant, au fil de l'accumulation de données qu'ils engrangent, des lieux de stockage sans moyens suffisants donnés à la curation, et n'étant pas en mesure d'avoir une certification à ce jour. Les données déposées dans Recherche Data Gouv sont sauvegardées dans plusieurs datacenters opérés par des établissements de l'enseignement supérieur et de la recherche : en Île-de-France, à Toulouse et à Strasbourg. L'accompagnement au dépôt de données mis en place par RDG met en évidence les techniques de curation des données qu'exige la publication de données de la recherche, et qui feront l'objet de notre troisième partie.

[97] Recherche Data Gouv : plateforme nationale fédérée des données de la recherche, 16/07/2021. Site Ouvrir la science. https://www.ouvrirlascience.fr/recherche-data-gouv-plateforme-nationale-federee-des-donnees-de-la-recherche/

[98] P. CORMIER, *Le positionnement des bibliothèques universitaires et de recherche françaises dans les politiques publiques des données de la recherche*. op. cit.

# Troisième partie

# FAIRiser les jeux de données : la curation aujourd'hui

## 3.1. La curation : remettre le geste humain au coeur des données

### 3.1.1. Définition

La curation des données de recherche désigne le fait de les gérer activement à tout moment de leur cycle de vie, à mesure qu'on les crée, les met à jour, les utilise, les archive, les diffuse et les réutilise. Elle recouvre divers aspects, tels que la création de documentation et de métadonnées visant à contextualiser les données, le travail sur la qualité au service de l'analyse des données, les étapes de vérification de fichiers et de code, la préparation d'ensembles de données aux fins de dépôt, la transformation de fichiers dans le but de les optimiser en prévision de leur réutilisation et de leur préservation à long terme, l'augmentation de métadonnées visant à faciliter la découverte de données, et bien plus.

Améliorer les pratiques de gestion des données est un processus complexe qui suppose un travail long et coûteux, des moyens techniques et humains parfois importants, et qui comprend plusieurs étapes avant d'aboutir à la publication et l'archivage de données fiables, de qualité, respectueuses du droit des personnes et de la législation en vigueur. Au cœur de cette gestion, la curation est définie de la sorte par l'INIST (Institut de l'Information Scientifique et Technique) rattaché au CNRS : « *On désigne par curation l'ensemble des activités et opérations nécessaires à une gestion active des données de recherche numérique, tout au long de leur cycle de vie. L'objectif est de les rendre accessibles, partageables et réutilisables de façon pérenne. Trois intervenants peuvent être identifiés dans le cycle de vie de données : les créateurs, le plus souvent les chercheurs, les « curateurs » et les utilisateurs* ».

### 3.1.2. Les bibliothécaires curateurs

Dans ce domaine, les bibliothèques universitaires et de recherche bénéficient de l'avantage de leur expérience collective et partagée dans tout ou partie de ces tâches. Grâce à l'ouverture des catalogues et à la diffusion de leurs métadonnées, ils sont déjà devenus des responsables de la gestion de données potentiellement exploitables par la communauté des chercheurs. Aussi, les bibliothécaires ont-ils un rôle important à jouer dans l'élaboration de métadonnées de qualité pour leur domaine spécifique, qui implique l'utilisation de syntaxes et de vocabulaires

communs, l'adoption de protocoles d'échange partagés, l'utilisation d'identifiants, de normes et de référentiels, ainsi que l'utilisation de technologies de traitement des données. Ce rôle nécessite également une connaissance approfondie du cadre juridique et politique de la science ouverte. En d'autres termes, la curation de données est une tâche qui consiste à identifier dans un catalogue de données celles qui peuvent être valorisées, exploitées et dans un deuxième temps, les mettre à la disposition des utilisateurs susceptibles d'en tirer les meilleurs enseignements. La curation remet le regard et l'action humaine au centre des données : devant la masse toujours croissante de données, le curateur les organise en les enrichissant.

Pour mettre en place une curation efficace et pertinente, il faut donc commencer par s'adosser à une cartographie précise de la donnée disponible, qui constitue le socle d'une gouvernance des données pragmatique et opérationnelle. Une fois les règles de gouvernance établies, c'est vers l'utilisateur des données qu'il faut concentrer l'attention. C'est en effet ce dernier qui est à l'origine du projet de curation des données.

### 3.1.3. Préparer le dépôt : motivations et vigilances

La publication sur un entrepôt requiert la vérification de la conformité du jeu de données aux règles établies par l'entrepôt-catalogue et par l'administrateur de la collection. Le travail de curation visant à assurer une bonne compréhension des données publiées porte sur la complétude des métadonnées et s'appuie dans certains entrepôts sur des guides de saisie des métadonnées. Il assure la présence d'une documentation complémentaire (fichier *lisez-moi*, dictionnaire des données), de formats de fichiers de données ouverts ou largement utilisés par la communauté concernée, la présence d'une licence et si nécessaire des précisions sur les conditions d'utilisation des données. Il veille à offrir la possibilité d'une découverte et d'une navigation la plus simple possible pour l'utilisateur novice.

Pour préparer la publication de jeux de données, les questions qui se posent sont diverses. La première consiste à interroger la pertinence de la ressource pour la communauté scientifique visée. Il faudra s'assurer aussi de l'état technique de la ressource, de la façon dont elle est maintenue, de sa stabilité dans la durée. Il est nécessaire également de s'assurer qu'elle contribue à l'adoption des bonnes pratiques en matière de science ouverte, de penser comment l'enrichir, par le biais des collections virtuelles d'un entrepôt fédérateur notamment.

Autre point de vigilance, il est nécessaire de vérifier si la ressource est déjà publiée. La pratique de double dépôt est fortement déconseillée, en raison tout d'abord des enjeux environnementaux évoqués plus haut,[99] mais également parce que le multiple dépôt entraînera la création de plusieurs identifiants pérennes à gérer, et, partant, un problème de lisibilité de la citation du jeu de données, avec un risque d'éparpillement et de divergence des données elles-mêmes. Dès lors, deux cas se posent : si les données appartiennent exactement au même jeu de données, il ne faut pas dupliquer le jeu de données, mais utiliser la notion de collections virtuelles qui existent sur de nombreuses plateformes pour indiquer les convergences utiles. Si les mêmes données appartiennent à plusieurs jeux de données, les données pourraient se retrouver de facto déposées à plusieurs endroits, ce qui n'est pas une bonne pratique. Enfin, prendre en charge la curation spécifique liée à chaque type de dépôt et mettre à jour ces dépôts mobiliseront évidemment davantage de ressources humaines. Autant de raisons qui invitent à réfléchir impérativement aux critères utilisés pour définir le jeu de données en amont.

Enfin, avant le dépôt, certains points de vigilance sont à respecter sur le plan juridique. Il est nécessaire de vérifier si le jeu de données est concerné par les secrets protégés (secret défense, médical, industriel etc.), ou soumis au droit d'auteur. Également, s'il s'agit de données à caractère personnel, la demande d'un consentement aux personnes concernées ou l'anonymisation des données sera requise. Si le jeu de données est concerné par une recherche partenariale entre acteurs publics et privés, la diffusion sera soumise à l'accord de consortium ou de contrats.

### 3.1.4. Se doter d'un modèle de description

L'enjeu du travail de curation, on l'a vu, réside dans cette question essentielle : comment convertir des montagnes d'informations en précieux indicateurs ? L'article scientifique *Datasheets for Datasets*[100], traduit sur le forum *#TeamOpenData*[101], fait le constat de l'absence de méthode normalisée pour documenter la façon et les raisons pour lesquelles un jeu de données a été créé, quelles informations il contient,

[99] Voir paragraphe 1.1.4.

[100] GEBRU, T., MORGENSTERN, J., VECCHIONE, et al., *Datasheets for datasets. Communications of the ACM, 64*, p. 86 - 92, 2018.

[101] Forum #TeamOpenData. https://teamopendata.org/t/traduction-et-adaptation-du-modele-de-description-des-donnees-datasheet-for-datasets/1400

les tâches pour lesquelles il devrait et ne devrait pas être utilisé, et propose le concept de fiche technique (*datasheets*) pour les jeux de données, avec une liste de questions à documenter. Samuel Goëta a traduit, simplifié et adapté cette liste de questions pour qu'elle réponde plus aux besoins des producteurs de données publiques ouvertes. Dans les grandes lignes, cette liste de questions concerne les motivations pour la création du jeu de données, sa composition, le processus de collecte des données qui le constituent, le pré-traitement des données, la diffusion du jeu de données et sa maintenance.[102]

### 3.1.5. L'identification pérenne, les référentiels et la citabilité des jeux de données

L'une des opérations majeures de la curation réside dans l'identification des données, car des identifiants dont la pérennité est garantie par une gouvernance adaptée facilitent l'interopérabilité des systèmes. Au cœur du foisonnement de la majeure partie de la production scientifique mondiale qu'on trouve désormais signalée, et même disponible en ligne, ils répondent à l'une des exigences découlant du premier principe FAIR, le critère Facile à trouver, qui porte précisément sur l'attribution à la donnée d'un identifiant unique. L'exploitation de cette masse de données nécessite de pouvoir identifier chaque entité, de manière univoque et pérenne, grâce à des systèmes d'identifiants adaptés. La définition d'un identifiant proposée par le CoSO est la suivante : « [..] *un numéro ou une étiquette alphanumérique, opaque ou explicite, qui peut être lu par des machines et des humains. Il permet de désigner et de retrouver de manière univoque et pérenne un objet, un document, une personne, un lieu, un organisme ou toute autre entité dans le monde physique ou numérique*. » [103]

Deux grands types d'identifiants se distinguent : les identifiants contributeurs pour les auteurs et les institutions, et les identifiants objet pour les productions scientifiques (publications, données et logiciels). On parle souvent de PID pour Persistent IDentifier en anglais. Les PID jouent un rôle essentiel dans plusieurs

[102] Traduction et adaptation du modèle de description des données « Datasheet for Datasets ». billet de blog *#TeamOpenData*, 2019. Consulté le 24 juillet 2023. https://teamopendata.org/t/traduction-et-adaptation-du-modele-de-description-des-donnees-datasheet-for-datasets/1400

[103] « Des identifiants ouverts pour la science ouverte : note d'orientation », Comité pour la science ouverte, 2019. https://www.ouvrirlascience.fr/des-identifiants-ouverts-pour-la-science-ouverte-note-dorientation/

aspects, notamment l'identification, la découverte, l'accès, le partage et la diffusion des productions scientifiques, mais ils permettent également de lier auteurs, institutions, publications, données et logiciels.

Les identifiants pérennes contributeurs garantissent en effet une identification fiable des auteurs et des institutions, ce qui facilite la traçabilité et la reconnaissance de leurs contributions, et vise à désambiguïser les problèmes d'homonymies ou les variations de translittération. Très répandu, l'ORCID (*Open Researcher and Contributor ID*, créé en 2010), permet d'identifier les chercheurs. De même, les identifiants pérennes objet permettent une identification fiable des productions scientifiques, contribuant ainsi à leur découverte et à leur référencement.

Les identifiants objet, utilisés pour les publications, les données et les logiciels, établissent des liens entre les articles publiés, les ensembles de données sous-jacents et les logiciels nécessaires à leur compréhension. Parmi ceux qui sont les plus utilisés dans le monde de la recherche, on peut citer l'ISSN (International Standard Serial Number, norme ISO créée en 1975) pour les revues, le DOI (Digital Object Identifier, créé en 2000) pour les documents en général, et plus particulièrement les articles, chapitres de livres et jeux de données. Ces PID facilitent le partage, la réutilisation et la reproductibilité des résultats de la recherche, en permettant aux chercheurs de retrouver facilement les ressources associées à un travail spécifique. Les publications scientifiques sont principalement représentées par des DOI. Les identifiants ARK (Archival Resource Key), utilisés principalement par les institutions patrimoniales telles que les bibliothèques, les archives ou les musées, ont été créés en 2001 par la California Digital Library et permettent d'identifier tout type de ressources, qu'elles soient physiques, numériques, immatérielles. Cela permet par exemple à la BnF d'identifier tant ses notices de catalogue que ses ressources numériques.

En France, plusieurs identifiants sont mis en œuvre pour les personnes, les affiliations et les publications, et ils sont proposés par diverses organisations telles que HAL (Hyper Article en Ligne), la plateforme pluridisciplinaire nationale pour le dépôt et la consultation travaux et résultats de recherches scientifiques, Huma-Num, la BnF et l'Abes (Agence bibliographique de l'enseignement supérieur). Parmi ces identifiants, IdRef (Identifiants et Référentiels pour l'ESR), est considéré comme un identifiant universel pour l'Enseignement supérieur et la Recherche. Il couvre les

personnes, une partie des publications et les structures, notamment lorsque l'entité est française ou liée à d'autres entités utilisées en France. IdRef offre ainsi une couverture étendue et permet une identification complète dans le domaine académique français.[104]

L'ouverture de l'entrepôt national Recherche.data.gouv réinterroge les pratiques de curation des données à l'heure actuelle. À travers l'expérience de la préparation et de la publication du jeu de données de Gallica sur le nouvel entrepôt, nous explorons de nouvelles options d'exposition.

## 3.2. L'exemple de la diffusion du jeu de données de Gallica

### 3.2.1. Les jeux de données BnF issues de projets de recherche

Extraites en vue d'un projet de recherche, soit par les professionnels de la BnF, soit par les équipes de recherche elles-mêmes, certaines données à forte valeur ajoutée deviennent un matériau de recherche intéressant à publier sous la forme de jeux de données réellement orientés vers la réutilisation. En effet, les extractions, le nettoyage et la structuration des données réalisés dans le cadre de recherches via des outils qui requièrent technicité et puissance machine, sont coûteux en temps, aussi l'exposition de leurs métadonnées constitue-t-elle une manne pour de nouveaux projets de recherches. Elle fait valoir par ailleurs les liens de la BnF avec la recherche.

A titre d'exemple d'ensemble intéressant à réintroduire sous la forme d'un jeu de données, figurent les données structurées dans le cadre du projet « *Oupoco* », l'Ouvroir de poésie combinatoire. Ce projet, dirigé par Thierry Poibeau, directeur adjoint du laboratoire Lattice[105] et titulaire de la chaire PRAIRIE (PaRis Artificial Intelligence Research Institute)[106], en collaboration avec Mylène Maignant, Frédérique Mélanie-Becquet et Clément Plancq, du laboratoire LATTICE, a pour but de générer des sonnets à partir de poésies d'auteurs français. La BnF a livré à l'équipe de recherche un important fichier étiqueté poésie aux contenus très hétérogènes. Ces données primaires, travaillées quatre années durant tout au long du projet, ont

[104] V. RICHARD, *Métadonnées pour la science ouverte : rôle et action des bibliothèques et des professionnels de l'information scientifique et technique*, Mémoire, Enssib, 2021.
[105] Site du Laboratoire Lattice. https://www.lattice.cnrs.fr/
[106] Site de l'ANR. https://anr.fr/ProjetIA-19-P3IA-0001

permis d'en élaborer de nouvelles : un important corpus de sonnets du XIX$^{e}$ et du début du XX$^{e}$ siècle.

Il existe de nombreux autres ensembles de données à haute valeur ajoutée, aux profils très divers, émanant des collaborations des différents départements en charge des collections avec des projets de recherche. Elles devront, pour celles qui s'inscrivent dans le cadre d'un projet ANR ou européen dans lequel la BnF est membre du consortium, faire l'objet d'une réflexion au cas par cas avec l'équipe de recherche partenaire, afin définir les rôles de chacun - auteurs, contributeurs - dans la publication des données. Si l'équipe partenaire souhaite garder la main sur la publication de ses données, la BnF pourra contribuer au plan de gestion de données originel afin de spécifier les métadonnées ou autres aspects de la documentation complémentaire qu'il lui semblera nécessaire de fournir. Elle pourra proposer les cibles de diffusion qu'elle a identifiées, et selon les modalités choisies, compléter le dépôt réalisé par l'équipe de recherche par d'autres référencements qui renverront vers ce dépôt.

### 3.2.2. Le jeu de données *Monographies de Gallica : texte océrisé*

Parmi ces jeux, l'un d'eux a été choisi pour faire l'objet d'une étude de la préparation de ses données d'abord, de ce que représentent son dépôt et sa publication ensuite. Il s'agit d'un *dump* qui fournit le texte océrisé des monographies en langue française de la collection numérique de Gallica.

Pour comprendre la valeur ajoutée du jeu de données *Monographies de Gallica : texte océrisé*, il faut en appréhender la nature, le contexte de production et les usages de recherche qu'il permet. L'extraction, réalisée par les deux chercheurs auteurs de l'outil Gallicagram, donne potentiellement matière à la création de deux grands jeux de données : l'un constitué de la totalité des monographies, l'autre de la presse numérisée dans Gallica. Nous nous sommes intéressés ici aux seules données constituées par les monographies.

Ce jeu est issu initialement de l'extraction produite par les créateurs de l'outil de lexicométrie Gallicagram, développé par Benjamin Azoulay (ENS Paris-Saclay) et Benoît de Courson (Max Planck Institute - CSL) à l'aide des API Gallica. Il s'agit de données numérisées en mode image. Les documents textuels conservés par la BnF

sont, depuis 2005, convertis en mode texte par un logiciel OCR[107], offrent aujourd'hui les possibilités d'une recherche sémantique par proximité de termes, répondant aux besoins liés aux nouvelles pratiques de fouille de texte et de données (TDM) et de traitement automatique du langage (TAL). Les résultats ainsi obtenus peuvent être analysés, comparés et extraits grâce au rapport de recherche, une fonctionnalité de la bibliothèque numérique Gallica qui améliore le dépouillement des résultats, dans des formats divers. Grâce aux API (IIIF, API documents, métadonnées), en service depuis 2017, les usagers peuvent extraire les contenus de Gallica à distance et lancer des requêtes sur les corpus. Ces données issues de l'océrisation figurent parmi les plus exploitées, comme en témoigne le projet Gallicagram.

Outil de lexicométrie développé pour la recherche en sciences humaines et sociales, Gallicagram offre aux chercheurs un moyen de tirer profit de très vastes bases de données linguistiques, délimitées, accessibles et structurées. Il s'inspire de l'outil Google Ngram Viewer dont les visées étaient semblables, pour offrir des applications plus performantes, avec une maîtrise des corpus et un accès aux documents exploités supérieurs, un traitement distinct de la presse et des livres et des outils d'analyse intégrés au logiciel de meilleure qualité.[108] Benjamin Azoulay et Benoît de Courson ont particulièrement travaillé sur les corpus de presse, - par nature sensibles aux soubresauts de l'actualité -, en moissonnant trois millions de numéros de presse, numérisés et océrisés de Gallica, là où Ngram Viewer a fait le choix d'exclure les journaux de son corpus. Ils donnent cette explication de l'usage de leur outil en contexte de recherche : « *Concrètement, Gallicagram permet de visualiser l'évolution de l'usage des mots au cours du temps en fouillant les corpus de presse et de livres numérisés par la BnF (Gallica) et par bien d'autres bibliothèques nationales et locales, en cinq langues. Le logiciel permet d'observer non seulement les tendances séculaires et les évolutions de moyen terme, mais surtout - et c'est une nouveauté - de s'approcher au plus près des événements.* »

Le travail d'extraction du corpus a été conséquent en termes de technicité, de puissance machine et de temps passé, c'est pourquoi il a été identifié par les experts du DCP comme le point de départ de deux jeux de données à réintroduire. Une

[107] L'OCR (Optical Character Recognition) est une technologie de reconnaissances de textes imprimés à partir d'images numérisées. Disponible dans Gallica depuis 2005, l'OCR constitue aujourd'hui une chaîne d'entrée interne à la BnF, qui est engagée notamment dans des programmes de rétroconversion.

[108], B. AZOULAY et B. DE COURSON, *Gallicagram : un outil de lexicométrie pour la recherche*. 2021. https://doi.org/10.31235/osf.io/84bf3

première version du jeu de données avait été réalisée à partir de l'extraction produite par l'équipe de Gallicagram en mars 2021. Deux ans après, en avril 2023, l'objectif était de faire une première mise à jour, en complétant le jeu de données initial avec les données des nouvelles monographies mises en ligne sur Gallica (18 000 nouveaux documents) entre le 1er avril 2021 et le 1er avril 2023. À compter de 2024, la mise à jour du jeu de données deviendrait annuelle. En vue de mettre en place une pratique structurée et réfléchie de la mise à jour du jeu de données, les activités de curation effectuées, d'abord par la préparation des données, ensuite leur publication, ont fait l'objet, dans le cadre du stage, d'une documentation et de guides pour chacun des lieux de dépôt, qui sont disponible en annexe.

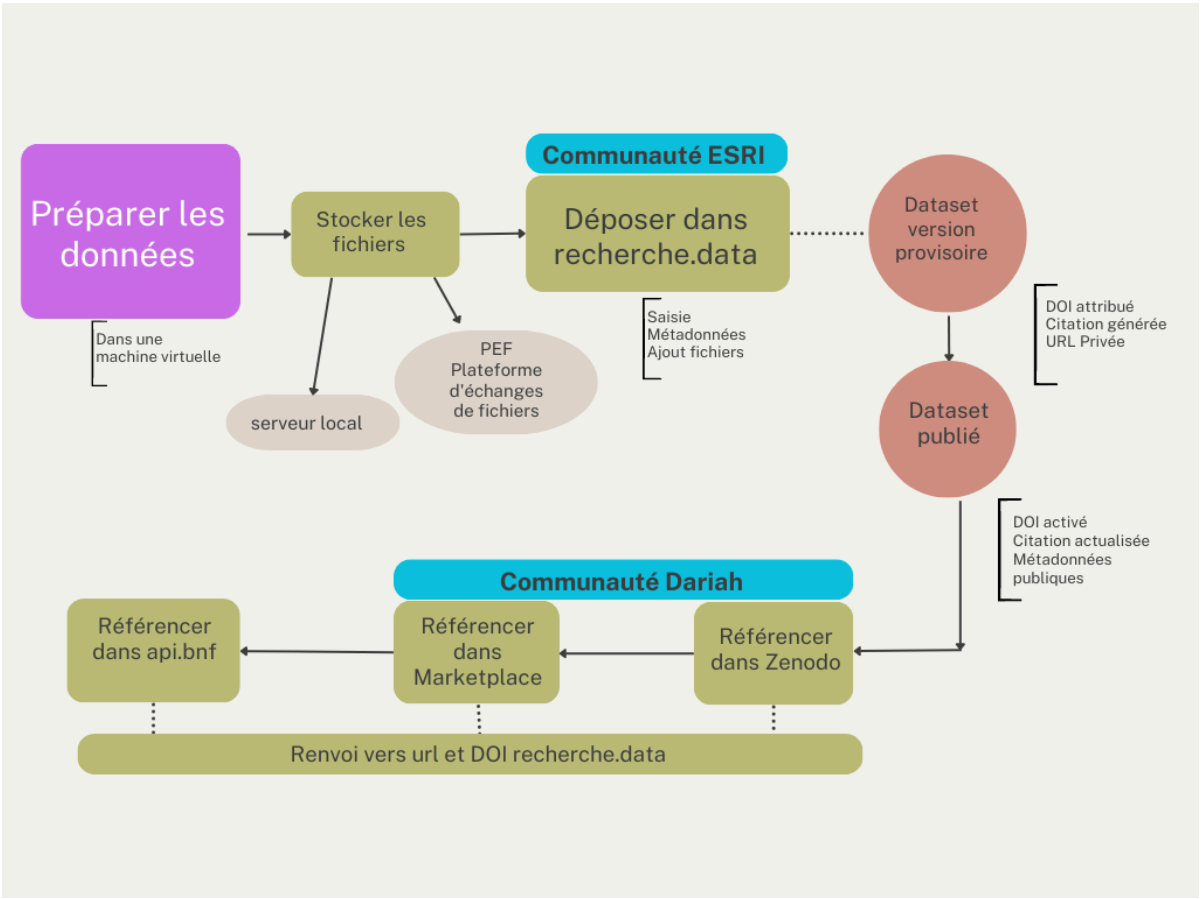
Les différentes étapes de la préparation des données - recherche avancée dans Gallica, export des résultats, nettoyage des métadonnées, compression des fichiers, dépôt et stockage des données produites - ont été restituées, et permettront d'effectuer la mise à jour du jeu de données en 2024. Elles sont documentées en annexe 1.

Dans un deuxième temps, en vue de la publication, la démarche de publication du jeu de données *Monographies de Gallica : texte océrisé* a fait l'objet d'une prise de contact avec les deux chercheurs de Gallicagram, afin d'obtenir leur accord avant de déposer le jeu dans l'entrepôt RDG, en le liant à leur projet de recherche, en référençant les publications liées et leurs identifiants le cas échéant. Une fois l'accord de ses auteurs obtenu, la diffusion du jeu de données a fait l'objet d'une réflexion approfondie pour déterminer les lieux stratégiques de référencement, et l'ordre des activités de curation à réaliser, et trois guides de publication correspondant respectivement aux annexes 2, 3 et 4, ont été élaborés afin d'éclaircir dans les grandes lignes quelques spécificités qui se sont avérées importantes lors des tests de dépôts qui ont été effectués pour le jeu de données Gallica.

### 3.3.1. Propositions pour un référencement élargi des jeux de données

Afin d'offrir une vue claire des étapes de la curation et des référencements proposés, ces dernières ont été représentées dans le diagramme de tâches figurant ci-dessous. La première étape représentée est celle de la préparation des données documentée dans l'annexe 1. Il a été envisagé que cette tâche soit réalisée dans une configuration dédiée, celle d'une machine virtuelle (VM)[109] créée dans une interface *DataLab center* sécurisée. En effet, comme indiqué dans l'annexe 1, paragraphe 1.3, si la préparation des données a vocation à être réalisée dans une VM, c'est que cette dernière permet l'utilisation de python, dont l'installation n'est pas possible sur les postes professionnels, et dont l'usage est nécessaire à la remise à jour annuelle du jeu de données.

Figure 4 : Flow pour la publication d'un jeu de données




[109] Une machine virtuelle est un environnement virtualisé qui fonctionne sur une machine physique. Elle permet d'émuler un OS sans l'installer physiquement sur l'ordinateur.

Vient ensuite le dépôt dans l'espace générique de l'entrepôt RDG. Cette publication se fait en deux grandes étapes, mises en évidence sur ce diagramme. La saisie des métadonnées et le versement des fichiers font l'objet d'une relecture par l'équipe de l'entrepôt. A ce stade, la version du jeu de données générée met à disposition un DOI et une url provisoire qui ne deviendront définitifs qu'après validation par les opérateurs de données de l'entrepôt. Ces derniers peuvent en effet être amenés à faire des recommandations pour augmenter la précision et la qualité des métadonnées, ce qui constitue un support non négligeable aux activités de curation, quand elles s'appuient sur des compétences naissantes. Un guide de publication disponible en annexe 2 décrit les principales étapes à effectuer. Chaque nouveau jeu de données déposé sera d'une nature différente et nécessitera une curation spécifique adaptée. Néanmoins le guide permettra de s'appuyer en partie sur cette première expérimentation de dépôt effectuée dans l'espace Travaux pratiques de RDG et validée par son équipe.

Le jeu de données alors identifié par ce DOI pérenne fera l'objet de référencements complémentaires sur Zenodo, sur le marketplace de DARIAH et sur api.bnf, qui renverront tous vers l'url du dépôt de RDG et son DOI. On l'a vu en effet, il n'est pas conseillé de déposer un même jeu de données dans plusieurs entrepôts, pour éviter la dilution des citations. Un DOI ne doit renvoyer qu'à un unique jeu de données, qui restera identique au fil des mises à jour annuelles. Les métadonnées, elles, peuvent être présentes sur plusieurs sites, mais renverront au dépôt unique réalisé sur RDG.

Le référencement sur l'entrepôt Zenodo, qu'il nous semble important de réaliser aujourd'hui car il entraînera l'indexation automatique du jeu de données sur OpenAire, ne permet pas de signaler le jeu de données sans effectuer de dépôt de fichiers. Aussi a-t-il été envisagé de réaliser un dépôt minimal en déposant uniquement, parmi les trois fichiers qui constituent le jeu de données, le fichier au format .csv contenant les métadonnées des monographies, qui est peu volumineux. Ce versement de fichiers renverra vers le DOI du dépôt disponible sur RDG, et s'assortira également d'une saisie de métadonnées la plus complète possible, avec l'aide du guide de publication disponible en annexe 3. Ce choix de référencement dans Zenodo ne nous semble utile que provisoirement, puisque l'ambition de Recherche Data Gouv est de devenir un service de l'EOSC, qui est soutenu par des politiques d'incitation européennes financières ambitieuses, pour fédérer les acteurs

de la science ouverte au niveau européen. Il est probable qu'alors, les données de RDG seront aspirées par l'EOSC.

Pour concrétiser le partenariat de la BnF avec DARIAH en référençant le jeu de données de Gallica dans le SSHOC Marketplace, on verra en annexe 3 comment renvoyer vers le DOI et l'url liés au dépôt existant sur RDG, comment utiliser les vocabulaires contrôlés spécifiques sur lesquels s'adosse le portail, et associer des publications en lien. Il est à noter que la publication de *worflows* illustrant la fabrique des jeux de données constitue une perspective de valorisation très intéressante, encouragée par Huma-Num, et sur laquelle le DataLab envisage de se pencher.

Enfin, la description du jeu sur le site api.bnf pourra s'appuyer sur la documentation interne de la BnF intitulé « Manuel de contribution Drupal », ainsi que sur l'annexe 4 qui résume les sections de publication de contenu importantes pour valoriser un jeu de données et leur fonctions dans le back office Drupal.

### 3.3.2. Promouvoir les jeux de données

Aujourd'hui la problématique n'est plus seulement celle du dépôt, mais celle de la diffusion et de la promotion des données. En effet, rendre les données largement accessibles ne signifie pas que le public visé, noyé par le foisonnement des publications, y accèdera. Créer une communauté et impliquer le grand public passe par la mise en place de méthodes de communication en ligne telles que les réseaux sociaux et les blogs qui ont parfois eu mauvaise réputation dans les cercles scientifiques et ne sont généralement pas perçues comme savantes. Pourtant, rendre compréhensibles les données passe par des actions de pédagogie et de communication, telle que la programmation d'ateliers de la donnée mise en place par RDG depuis son lancement. Pour atteindre le grand public et créer des communautés, l'animation donne de la valeur aux données. Participer à un podcast de communication scientifique, écrire un article de blog, tweeter la dernière publication d'un jeu de données en apportant du contexte - le dépôt dans RDG offre une fonction de partage du jeu de données sur les réseaux sociaux -, est mobilisateur, et permettra d'être visible dans les cercles publics, à travers les médias de masse traditionnels, plutôt que se restreindre aux cercles académiques.

Cette préoccupation de l'ouverture au plus grand nombre relève des principes directeurs du mouvement de la science ouverte, qui estiment que la durabilité et

l'inclusion lui sont essentielles, favorisent des pratiques, des infrastructures et des modèles de financement partagés, qui garantissent la participation équitable des scientifiques d'institutions moins favorisées à la poursuite du savoir et du progrès.

## 3.4. Vers un portail open data interopérable

Le niveau de curation attendu par le nouvel entrepôt national nécessitera le développement de compétences, de pratiques et d'une organisation collaborative liée. Ces nouvelles méthodes de diffusion, chronophages aujourd'hui, gagneront, une fois rodées, à être automatisées et regroupées dans un espace dédié afin de mieux créer de la valeur à partir des données.

### 3.4.1. Automatiser la publication de jeux de données sur data.gouv

Nous l'avons déjà évoqué, la deuxième version d'api.bnf en 2020 s'est donné entre autres l'objectif d'exposer les données du portail, conformément aux normes portées par les institutions d'*open data* du gouvernement américain, avec le modèle de données DCAT. Ce schéma de métadonnées vise à normaliser la description des catalogues d'informations publiques et est actuellement maintenu par l'organisme de normalisation du web W3C (World Wide Web Consortium). Il peut être utilisé pour décrire à la fois un catalogue (dcat :catalogue) de jeu de données publié et maintenu par une organisation, mais également pour décrire finement chacun des éléments constitutifs de ce catalogue, à savoir les jeux de données (dcat :dataset) décrivant les fichiers (dcat :resource) qu'il rend accessibles ainsi que les différents acteurs (foaf :agent) impliqués dans cette mise à disposition. Il recommande également l'utilisation de vocabulaires contrôlés (skos :ConceptScheme), c'est-à-dire des listes fermées de termes permettant de catégoriser les jeux de données, publiées sur Internet, de manière à faciliter la mise en relation des jeux de données avec d'autres jeux de données décrits au sein d'autres catalogues. Il fournit par son caractère extensible, un cadre d'interopérabilité avec les langages de description utilisés par les communautés du patrimoine culturel.[110]

[110] *Guide mise en œuvre DCAT*. https://docs.google.com/document/d/1qMDqBjrTJVu3t9RH94aLSW7Z3jhH1SjoBrWhW9PZkJ4/edit?usp=embed_facebook

Cette implémentation rend possible aujourd'hui l'automatisation de la publication des jeux de données sur la plateforme de l'État data.gouv.fr, par le moyen d'un moissonnage. Il s'agit d'un processus de récupération des données publiées sur un portail open data local afin de rapatrier les données et informations sur la plateforme data.gouv.fr. Le tableau de bord du compte BnF sur la plateforme data.gouv permettra de configurer un moissonneur qui sera validé par l'équipe de l'entrepôt. De cette manière, il ne sera plus nécessaire d'importer à la main dans data.gouv.fr les jeux de données déjà décrits par api.bnf.

### 3.4.2. Développer un entrepôt-catalogue Dataverse aux contenus élargis

A l'image du portail data.bl.uk développé par la British Library, un portail de données BnF en lien avec l'entrepôt national fédérateur que vise à devenir RDG, consoliderait la fairisation des jeux de données BnF. Ce portail présentant des contenus élargis, constitué d'un catalogue et d'un entrepôt, et dont les données seraient indexées de façon automatique par le moyen d'un moissonnage sur RDG permettrait une couverture bien plus complète des besoins liés à la coopération numérique.

Etant donné la place qu'il est en train de prendre dans l'écosystème des données de la recherche, c'est en effet vers l'entrepôt RDG qu'il est prometteur de tourner les yeux aujourd'hui, afin de suivre sa progression et d'adapter un modèle d'ouverture de données ouvert sur cet entrepôt national, en développant un portail de données BnF basé sur la même solution logicielle libre, Dataverse[111]. Cette solution en effet, développée par l'université de Harvard, présente l'avantage d'être largement adoptée par la communauté française. Un entrepôt Dataverse héberge plusieurs Dataverses, chacun d'eux contient des jeux de données, et chaque jeu de données contient des métadonnées descriptives et des fichiers de données (y compris la documentation et le code qui accompagnent les données). Le logiciel est disponible pour téléchargement sur GitHub avec un guide d'installation détaillé afin d'aider les institutions à faire fonctionner leur entrepôt. Plusieurs entrepôts Dataverse ont déjà

[111] *Gestion et partage des données et des logiciels. Dataverse : un logiciel open source pour créer des entrepôts de données.* Blog de l'Institut Pasteur. https://openscience.pasteur.fr/2018/03/08/dataverse-un-logiciel-open-source-pour-creer-des-entrepots-de-donnees/

été installés par des institutions françaises. Récemment, Science Po, le CIRAD et l'INRA[112] ont ouvert leur entrepôt de données sur la base de cette solution logicielle.

Chaque établissement d'enseignement supérieur ou de recherche qui a son propre entrepôt peut demander la création d'un espace institutionnel dans RDG afin de déposer et publier ses données de recherche. Les jeux de données déposés jusque là dans l'espace générique basculent alors vers l'espace institutionnel. Cet espace est administré par l'institution responsable, et nécessite, pour que son ouverture soit validée par l'équipe de RDG, un engagement à mobiliser des ressources humaines, exprimé en Poste équivalent temps plein (ETP) dédié à la gestion de données. Le service de moissonnage des autres entrepôts Dataverse est en cours de développement par RDG. Quand il sera opérationnel, les institutions ayant des portails en Dataverse contribueront au catalogue de données RDG via le moissonnage des métadonnées de leurs collections institutionnelles.

Recherche.data.gouv présente les atouts d'être une solution souveraine qui s'inscrira dans un paysage international en évolution en devenant un service de l'EOSC dont l'équipe projet est augmentée de quelques membres européens (Norvège, Pays-Bas). Compte tenu de son expérience des dispositifs mis à disposition des communautés scientifiques ainsi que du maillage national constitué, l'IR* Huma-Num a été labellisé « Centre de référence thématique » de RDG pour les sciences humaines et sociales. Aussi l'inscription des collections de la BnF dans RDG prendra-t-il tout son sens, dans une logique partenariale également.

Mettre en place un tel portail permettra d'assurer la curation, la publication et le partage d'une collection de jeux de données coproduits par les experts BnF et leurs partenaires du monde de la recherche en faisant usage des collections « d'unités de recherche » qu'il est possible de créer dans l'espace institutionnel, et qui pourront mentionner la collection BnF et réciproquement.

Ce portail offrira la possibilité de donner accès à l'ensemble des collections numériques par sources, et ainsi de valoriser davantage les archives de l'internet. Donnant accès aux API et autres outils d'extraction de données il permettra également de mettre à disposition le code qui sera prochainement rassemblé sur un dépôt GitHub de la BnF, ainsi que les notebook Jupyter[113] qui simplifient encore l'expérience d'exploration des collections numériques. Une boîte à outils construite

[112] L'Institut national de la recherche agronomique

[113] Application web originale pour la création et le partage de documents informatiques, qui offre une expérience simple, rationalisée et centrée sur le document.

par le DataLab depuis deux ans y trouvera sa place, ainsi que la publication de *workflows* mis en place dans les parcours de recherche.

L'objectif d'un portail *open data* réside également dans la promotion des nombreux services de soutien à la recherche numérique mis en place par le DataLab, ainsi qu'à celle de ses événements - qui en dehors de leur mention dans les listes de diffusion, ne sont disponibles aujourd'hui que sur un carnet Hypothèses[114] -, aux opportunités de formation et aux tutoriels élaborés par la communauté professionnelle de la BnF.

La description des procédés de curation, puis l'expérimentation de leur mise en application à l'échelle d'un jeu de données à haut potentiel, a permis d'envisager le développement des compétences et des moyens que représente la publication de données ouvertes. Pour aller plus loin dans le partage des données, les portails développés aujourd'hui permettent de décloisonner les acteurs de la curation ainsi que les catégories de données au sein d'une organisation, et d'automatiser les processus grâce à des connecteurs, des web services, et du moissonnage. Rendre la publication la plus fluide possible permettra de mieux se concentrer sur l'analyse des données plutôt que sur leur préparation et leur traitement et de contribuer à démocratiser leurs usages.

[114] Carnet Hypothèses de la recherche à la BnF.
https://bnf.hypotheses.org/category/labos-de-la-bnf/bnf-datalab

# Conclusion

Le mouvement de la science ouverte, s'appuyant sur l'opportunité que représente la mutation numérique pour développer d'abord l'accès ouvert aux publications, se préoccupe désormais également de l'ouverture des données, codes sources et méthodes de la recherche. Beaucoup de chemin a été parcouru, et les politiques publiques se sont emparées véritablement de ces questions ces dernières années pour mettre en cohérence les initiatives existantes et les multiplier. Confrontés à une explosion du coût de la documentation scientifique depuis le début du XXI$^{e}$ siècle, les bibliothèques de recherche ont imaginé activement des solutions pour concilier accès à l'information et soutenabilité économique.

Oeuvrant depuis toujours, au fil des mutations technologiques, à la capacité des systèmes d'informatique documentaire à opérer ensemble, elles se sont engagées dans l'accompagnement des pratiques des chercheurs en matière de science ouverte, et font face aujourd'hui à un nouvel enjeu, celui d'assister les projets de recherche dans l'amélioration de la qualité des données et des métadonnées qu'ils produisent à partir de leurs collections. Afin d'y répondre, elles ont donné naissance à de nouvelles infrastructures telles que le BnF DataLab, qui offrent des opportunités pour se positionner à présent auprès des équipes de recherche dès l'amont de la production des résultats, prendre en compte les besoins d'une science ouverte lors du montage des projets, et accompagner la rédaction de plans de gestion de données.

L'exploration des enjeux liés aux données de recherche, des fonctionnalités et services que doit proposer un dépôt de données ouvertes, des principes FAIR et de leur mise en œuvre, ont permis de montrer que la bonne gestion des pratiques d'ouverture inclut la publication dans un entrepôt spécialisé dans la diffusion de ce type de données tel que RDG. En effet, une sélection d'entrepôts ayant fait l'objet d'une étude plus approfondie a montré que parmi les entrepôts véritablement dédiés aux données de la recherche et non aux données publiques, RDG est celui qui se voit attribuer les moyens humains et budgétaires ainsi que la volonté politique qui permettra de le rendre FAIR.

Le travail de curation des données engagé pour le seul jeu de données *Monographies de Gallica : texte océrisé* a montré que les techniques de curation sont une activité essentielle à la dynamique d'exposition, car elles assurent la pérennité des données sur le long terme, leur qualité, et leur réexploitation.

Si la démarche d'ouverture des données de recherche dans laquelle les bibliothèques sont engagées est positive à de nombreux égards (préservation pérenne, reproductibilité, etc.), elle doit néanmoins, et de manière urgente, être considérée aussi du point de vue de son impact environnemental. Une réflexion sur les outils, les infrastructures et les formats à utiliser s'impose autant qu'une gestion FAIR rigoureuse avec une sélection stricte des données utiles, nécessaires, validées et suffisamment bien qualifiées (avec des métadonnées de qualité) pour éviter de sauvegarder et de conserver des données inutilisables. À chaque phase du cycle de vie des données, il sera important d'identifier les possibilités d'amélioration en tenant compte de ces enjeux : réduire les transports physiques en les mettant au plus près de l'usage, développer les stockages hors ligne pour des données faiblement utilisées et donc accessibles à la demande dans des délais de traitement acceptables, gérer intelligemment les flux de données à toutes les étapes, éviter les duplications, la production et le stockage superflus de données, archiver en collaboration avec des centres adaptés (on pensera par exemple, pour organiser la préservation pérenne des données au Centre informatique national de l'enseignement supérieur (CINES) dans le monde de l'ESR). Pour ce qui est du matériel, les enjeux de durée de vie sont la première clé de la diminution des impacts. Quant aux logiciels et aux formats de fichiers, les formats et les logiciels libres assurent une pérennité incontestable et apparaissent comme une réponse incontournable. Il est essentiel d'éviter de refaire localement ce qui existe à d'autres échelles, comme développer ses propres solutions logicielles.

Ces nouvelles missions ont de nombreuses implications pour la communauté professionnelle, et représentent des défis en termes organisationnels. D'abord les compétences déjà mobilisées au sein des services : ingénierie pédagogique, normes et standards de la diffusion numérique pérenne, doivent s'adapter au contexte de la science ouverte. Ensuite, de nouvelles compétences restent à développer dans le domaine de la gestion des données de recherche. Enfin, les organisations et les services doivent s'adapter, avec une emprise plus forte du fonctionnement en mode projet.

Cet écosystème de la science ouverte est en effet un espace dynamique en cours de structuration qui ne peut se développer qu'en y associant étroitement les communautés utilisatrices. Au-delà de la fonction documentaire, il nécessite en effet la coopération de plusieurs acteurs des structures documentaires : directions de la recherche, direction des systèmes d'information et du numérique, archivistes.

La construction de partenariats, voire de guichets uniques, permet de mettre à disposition des chercheurs une palette large de compétences. Elle offre également la possibilité de disposer d'une vue prospective sur les dynamiques en cours, afin de pouvoir anticiper les usages futurs. Le BnF DataLab, construit dans cette logique partenariale, et engagé depuis ses débuts aux côtés de l'IR* Huma-Num, joue activement cette carte : récemment, la mobilisation pour adhérer à DARIAH vient ouvrir la BnF à une nouvelle dynamique internationale, et la pousse à revoir la diffusion de ses jeux de données dans de nouvelles modalités. Ouvrir les données du chercheur, et pour cela le placer au centre d'un réseau de services, offre l'opportunité enthousiasmante d'une coopération numérique et d'une rencontre entre cultures professionnelles toujours renouvelée.[115]

[115] M. GÉROUDET, « La science ouverte, nouvelles pratiques numériques, nouvelles compétences.» *Le numérique universitaire des BU*. Avril 2022. N° 20, pp. 8-9.

# Annexes

# Annexe 1

# Constitution du Jeu de données

# *Monographies de Gallica : texte océrisé*

Documentation technique

# 1. Contexte de production

## 1.1. Objectif de mise à jour du jeu de données

L'équipe de Gallica a déposé pour la 1ere fois sur api.bnf en mars 2021 le jeu de données *Dumps* Gallica : OCR des monographies | Api. Cette première version du jeu de données avait été réalisée à partir d'une extraction produite par l'équipe de chercheurs de l'outil Gallicagram en mars 2021 à des fins de développement de leur outil de lexicométrie.

En avril 2023, l'objectif était de faire une première mise à jour du jeu, en complétant le jeu de données initial avec les données des nouvelles monographies mises en ligne sur Gallica entre le 1er avril 2021 (+18000 documents) et le 1er avril 2023.

À compter de 2024, la mise à jour du jeu de données deviendra annuelle et la préparation des données devra s'appuyer sur cette documentation.

## 1.2. La génération initiale

La génération initiale a été réalisée par les auteurs de Gallicagram via l'API SRU.
Il avait été envisagé d'utiliser le protocole d'échange OAI-PMH, pratique pour récupérer une catégorie de documents existante dans l'entrepôt OAI-Gallica mais ici non pertinent car il ne permet pas de faire des requêtes en amont sur les catégories. Le requêtage OAI dans le cas présent nécessiterait de télécharger d'abord les 630 000 monographies avant de pouvoir filtrer sur les critères OCR et langue française.

Si le résultat dépasse 60 000 documents (volumétrie maximale du rapport de recherche), l'utilisation de l'API SRU (= service web fournissant les métadonnées des documents renvoyés par une requête Gallica, selon un format XML) est nécessaire :

1. Dans Gallica, création d'une requête avec le formulaire de recherche avancée.
2. Dans l'url de la page de résultats de Gallica, copier le champ *query*.
3. Coller la requête dans l'API SRU Gallica. Appeler l'API.
4. L'API renvoie par défaut des tranches de 15 documents, sauf à utiliser le paramètre maxRecord (lui-même limité à 50 max)
5. Parser le flux XML renvoyé et extraire les informations requises, en particulier les identifiants ARK des documents de la liste de résultat dans un fichier txt ou csv (voir la section suivante).

6. Pour les étapes ultérieures voir ci-après

## 1.3. Configuration de la VM dédiée à la préparation du dataset

La préparation des données a vocation à être réalisée dans une machine virtuelle. La VM permet en effet l'utilisation de python dont l'installation n'est pas possible sur les postes professionnels.

Dans le contexte de la mise à jour 2023, la VM Datalab dédiée à la préparation des fichiers n'a pas été utilisée du fait des problèmes techniques rencontrés. Les fichiers du *dump* initial ont été remis à jour et découpés en utilisant le serveur Snoop.

- **Habilitation** : nécessite de pouvoir s'authentifier sur un compte Datalab center
- **Accès** : adresse ip et mot de passe se retrouvent sur l'interface Datalab center
- **Dimensionnement :** 6 CPU, 16 Go, disque 500 Go. La définition du nombre de CPU (processeurs), de la taille mémoire, et de celle du disque dur a été faite selon le poids du corpus et les types de traitement appliqués. Un échantillonnage aide à évaluer le poids du corpus cible pour envisager sa volumétrie. Dans le cas du *dump* OCR, la volumétrie est importante mais les traitements restent simples.
- **Logiciels pré-installés systématiquement dans les VM** :
  - R et R Studio, un environnement de développement pour R, un langage de programmation utilisé pour le traitement de données et l'analyse statistique.
  - PyCharm, un environnement de développement intégré utilisé pour programmer en langage Python,
  - Visual Studio Code, un éditeur de code,
  - OCRFeeder et tesseract, des outils d'OCR.
- **VM ouverte sur le web** : ouverte pour publier le jeu en ligne
- **Structuration des données** : un dossier contenant cette documentation technique et les scripts python nécessaires à l'extraction des données chaque année.
- **Durée de vie de la VM :** cette VM est basique et n'a pas besoin d'être clonée donc il n'est pas nécessaire d'en garder une image et elle sera supprimée après la mise à jour du jeu de données.

## 1.4. Contexte technique de la préparation du jeu de données en 2023 :

Le fichier .txt initial des OCR, d'un volume de 105 Go, n'était pas récupérable depuis api.bnf. Les options suivantes ont échoué :

- Le téléchargement depuis le portail API n'étant plus fonctionnel depuis un an car le protocole ftp de la plateforme d'échange de fichiers (PEF) n'est plus supporté par les navigateurs. Internet explorer qui est plus ancien peut utiliser le ftp mais ne pouvait pas être installé sur Linux.
- Le téléchargement depuis la PEF sur la VM via Filezilla n'était pas possible car le réseau Datalab est étanche au réseau BnF Pro.
- Clé USB non reconnue par la VM du Datalab center
- Alternative : les fichiers du dump initial ont été remis à jour et découpés sur le serveur Snoop puis intégrés avec GoDrive, qui permet de créer des liens de téléchargement (upload via protocole HTTPS). Une licence a été créée pour le DCP (compte GoDrive DCP : https://pef.bnf.fr/webclient/WebClient.xhtml).
  **Cette solution a réglé le problème pour les mises à jour ultérieures.**

# 2. Mettre à jour le jeu de données

Cette partie vise à décrire le processus de mise à jour qui peut s'appliquer à n'importe quelle période. L'étape de requêtage dans Gallica devra être adaptée en fonction de la volumétrie (rapport de recherche en deçà de 60 000 documents, et API SRU au-delà)

Pour chaque mise à jour, extraire les données à partir du 1er avril de l'année *n* en utilisant le rapport de recherche de Gallica si le nombre d'océrisations annuelles ne dépasse pas 60 000 documents, pour obtenir les fichiers suivants :

- la liste des ARK dans un fichier texte (les url avec les ARK de toutes les monographies océrisées) dans l'interface Gallica en date du 25 avril de l'année *n* ;
- le fichier de métadonnées CSV correspondant;
- les fichiers texte (compressés).

## 2. 1. Faire une recherche avancée dans Gallica :

→ Par type de document / *Livres*
→ Notice et texte intégral / langue /*français*
→ Par date de mise en ligne / *Après le 1er avril de l'année n* (date de génération du précédent jeu de données)

→ Par format : Choisir *En Mode texte* (pour avoir les documents océrisés) et *Libres* (pour avoir seulement les libres de droit)
→ Lancer la recherche avec affichage au volume / fascicule (exemple pour la mise à jour 2021-2023 = 18 293 résultats)

Requête obtenue:
https://gallica.bnf.fr/services/engine/search/sru?operation=searchRetrieve&exactSearch=true&collapsing=false&version=1.2&query=(dc.language%20all%20%22fre%22)%20and%20(dc.type%20all%20%22monographie%22)%20and%20(gallicapublication_date%3E=%221380%22)%20and%20(ocr.quality%20all%20%22Texte%20disponible%22)%20and%20(indexationdate%3E%222021/04/01%22

## 2.2. Exporter les résultats de la recherche Gallica avec le rapport de recherche

→ Bouton *Export des résultats* puis *Exporter tous* au format csv
→ Vérifier les métadonnées dans le tableur : Par défaut, les ARK seront en colonne A. Le séparateur est le point virgule par défaut dans le tableur or il peut aussi apparaître dans les notices, ce qui peut générer ponctuellement des problèmes de colonnes dans le fichier CSV. C'est un problème sans solution, mais à identifier avant de publier.

## 2.3. Avec l'API document Gallica .texteBrut, extraire les fichiers texte de l'OCR correspondants aux ARK identifiés

Lancer le script telechargement.py qui prend en entrée le résultat du rapport de recherche (= le csv obtenu après requêtage avec la recherche avancée Gallica) et qui pour chaque ARK utilise l'API .texteBrut afin de récupérer le texte brut de chaque document.
Cela génère un fichier .txt par document.

**Usage** : >python3 telechargement.py
**Notes** :

- Le nom du fichier csv doit être paramétré dans le script (par défaut liste-arks.csv)
- Le nom du dossier où seront générés les fichiers texte doit être indiqué en paramètre dans le script
- Le script crée un dossier par document, nommé d'après l'ARK du document, dans lequel est stocké le fichier texte.

- Vérifier le nombre de documents obtenus avec :
  > ls dossier_textes | wc -l
- Certains documents texte n'y figurent pas car soumis à une licence d'usage avec restriction (il s'agit de documents au format PDF)
- Générer la liste des ARK :
  > ls dossier_textes | wc -l > liste.txt

## 2.4. Supprimer la section de métadonnées

Supprimer la section de métadonnées en tête de chaque fichier (données inutiles, car déjà présentes dans le fichier csv) avec un script.

**Usage** : >python3 cleanTextFromTXT.py -dir répertoire_des_textes -out dossier_des_textes_nettoyés

**Notes** :

- Le script génère tous les fichiers nettoyés dans le dossier de sortie
- Vérifier le nombre de fichiers obtenus avec :
  > ls dossier_des_textes_nettoyés | wc -l

## 2.5. Compresser le jeu de données

La commande tar permet de compresser tous les fichiers en une seule archive.

**Usage** : >tar -cvf dump.tar dossier_des_textes_nettoyés

Si le .tar obtenu est supérieur à 5 Go, il est préférable de compresser en plusieurs fichiers compressés (ZIP ou TAR) de moindre taille. Pour cela, utiliser le script suivant, en choisissant par exemple une taille de 5 Go, qui correspond à la limite de transfert.bnf.fr et peut correspondre au seuil d'upload de certaines plateformes.

**Usage** : >python3 zip_files.py -d dossier_des_textes_nettoyés -o ZIP_annee_n -s 5000

## 2.6. Stockage/dépôt des données produites

Les fichiers générés doivent être sauvegardés sur l'espace serveur $NumDatasets et intégrés dans la PEF avec GoDrive :

- le fichier de métadonnées au format csv
- la liste des ARKS des documents ayant un texte au format csv
- le ou les archives texte compressées au format .zip ou .tar

### 2.6.1. Sauvegarde sur l'espace serveur $NumDatasets

Dossier :
\\num_datasets.bnf.fr\Num_Datasets$\api.bnf.fr\Gallica\Dumps\OCR\monographies
Créer un dossier dont le nom sera l'année de la mise à jour et y intégrer les 3 fichiers.

### 2.6.2. Intégration dans la PEF avec GoDrive

- Se connecter à l'interface Web Go Drive sur l'url https://pef.bnf.fr
- Rentrer son login et son mdp Windows BnF en production.
- Créer un nouveau dossier nommé année *n*
- Une fois positionné dans l'arborescence au niveau du bon dossier, uploader un ou plusieurs fichier(s) à la fois, en cliquant sur le bouton « importer » puis en sélectionnant via la fenêtre modale Windows les fichiers sur un disque dur, une clef USB ou le réseau.
- Générer le lien public :

cliquer sur l'icône rouage à gauche du document ou dossier sur lequel on veut générer
cliquer sur « Modifier les paramètres » pour s'assurer que l'expiration du lien est désactivée
cliquer sur « Copier le lien »

# Annexe 2

# Guide pour la publication sur l'entrepôt recherche.data.gouv dans l'espace générique

## Table des matières



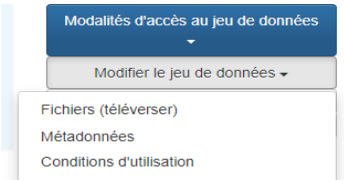
## 1. S'entraîner dans l'espace Travaux pratiques

- Créer un compte pour déposer dans le Bac à sable de l'entrepôt.
- Créer un compte pour déposer dans l'espace générique de l'entrepôt doit faire l'objet d'une demande officielle préalable de la BnF à l'équipe de gestionnaires de l'entrepôt.
- S'authentifier pour afficher l'espace de publication.
- Créer une collection BnF et des sous-collections (par exemple BnF Gallica) qui permettront de regrouper et ainsi rendre plus accessibles les données déposées.

## 2. Ajouter/Modifier des données.

- Se placer dans la collection identifiée, cliquer sur *Nouveau Dataset* dans *Ajouter des données*.


- Un bouton *modifier le jeu de données* permet de revenir au dépôt pour le compléter. Trois aspects du dépôt correspondent aux trois sous-rubriques Fichiers / Métadonnées / Conditions d'utilisation à renseigner dans *Modalités d'accès aux jeux de données*.




## 3. Renseigner les métadonnées

***Titre*** : Doit pouvoir renseigner n'importe quel réutilisateur sur le contenu du fichier.

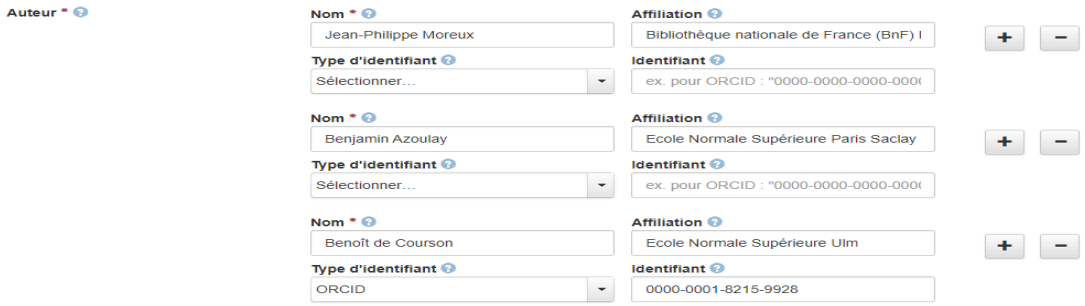
***Sous-titre*** : possibilité d'un renseignement alternatif, d'un éventuel titre plus parlant.

***Autre titre*** : il est recommandé d'y indiquer le titre en anglais

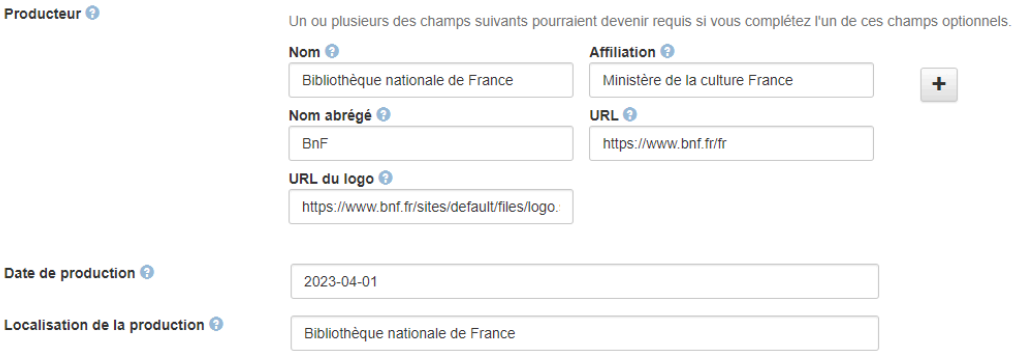
***Points de contact*** : adresse mail du déposant ou adresse générique du Département contributeur

***Lien vers les données*** : en cas de référencement sans dépôt indiquer le lien vers les données.

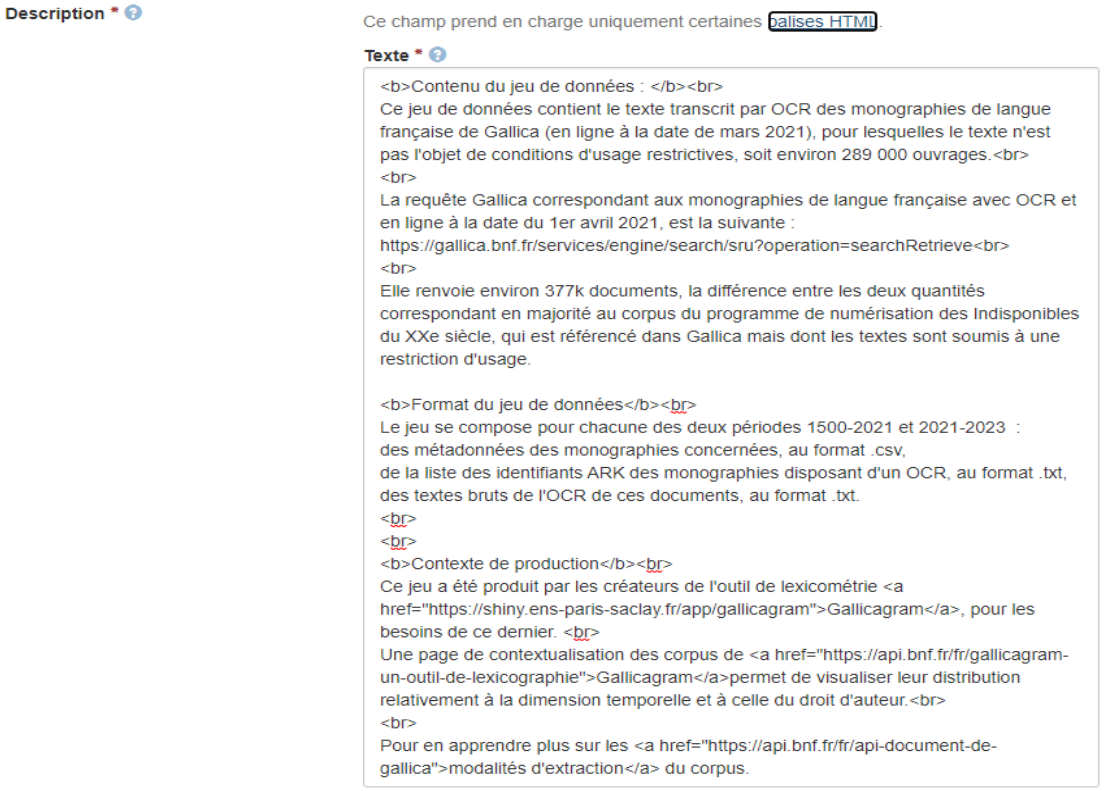
***Auteurs*** : Indiquer les types d'identifiants auteurs, les identifiants eux-mêmes et l'affiliation en toutes lettres des auteurs




***Producteur*** : Indiquer BnF Affiliation Ministre de la culture et url du logo
***Date de production*** : la date choisie a été la date d'extraction des données
***Localisation de la production*** : Indiquer BnF



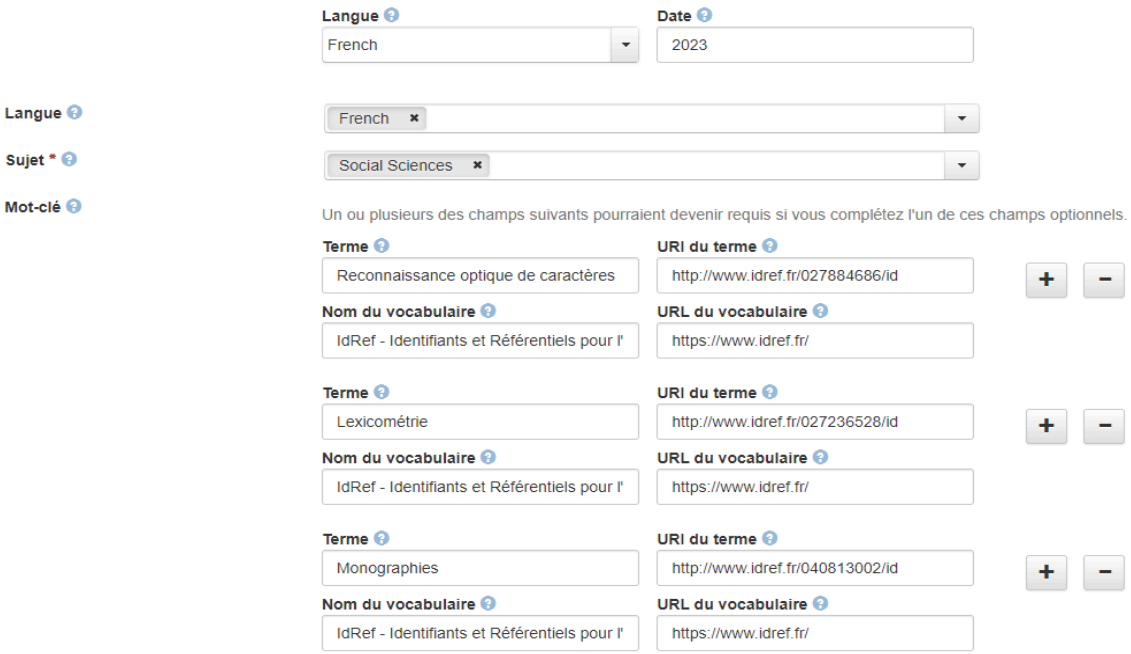
***Description*** : Utiliser pour la mise en forme les balises html indiquées en haut, notamment  pour le caractère gras, <br>pour le passage à la ligne et <br><br> pour le saut de ligne.



***Langue*** **:** indiquer impérativement la langue et la date de mise à jour des données

***Sujets*** : Si la couverture thématique des données est vaste, choisir un terme générique.

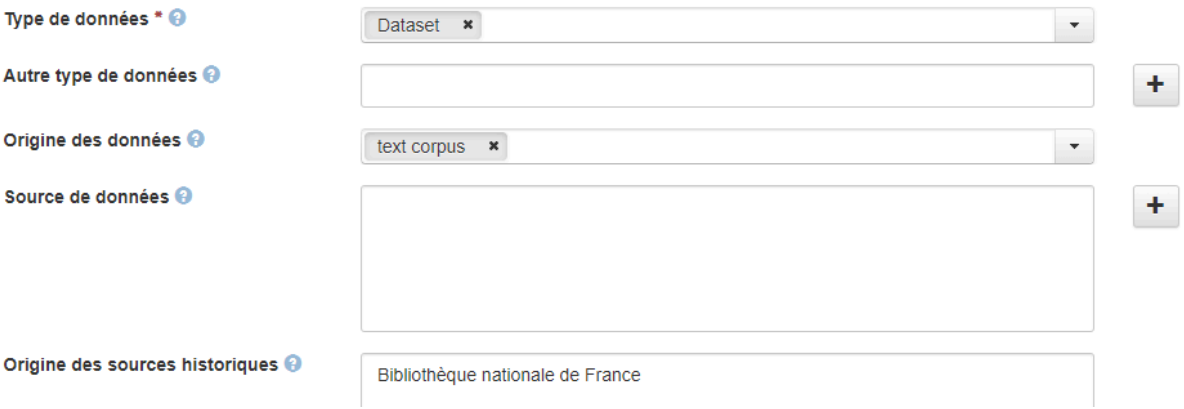
***Mots-clé*** : Se référer à un vocabulaire contrôlé. Par exemple le référentiel Idref pour l'enseignement supérieur.



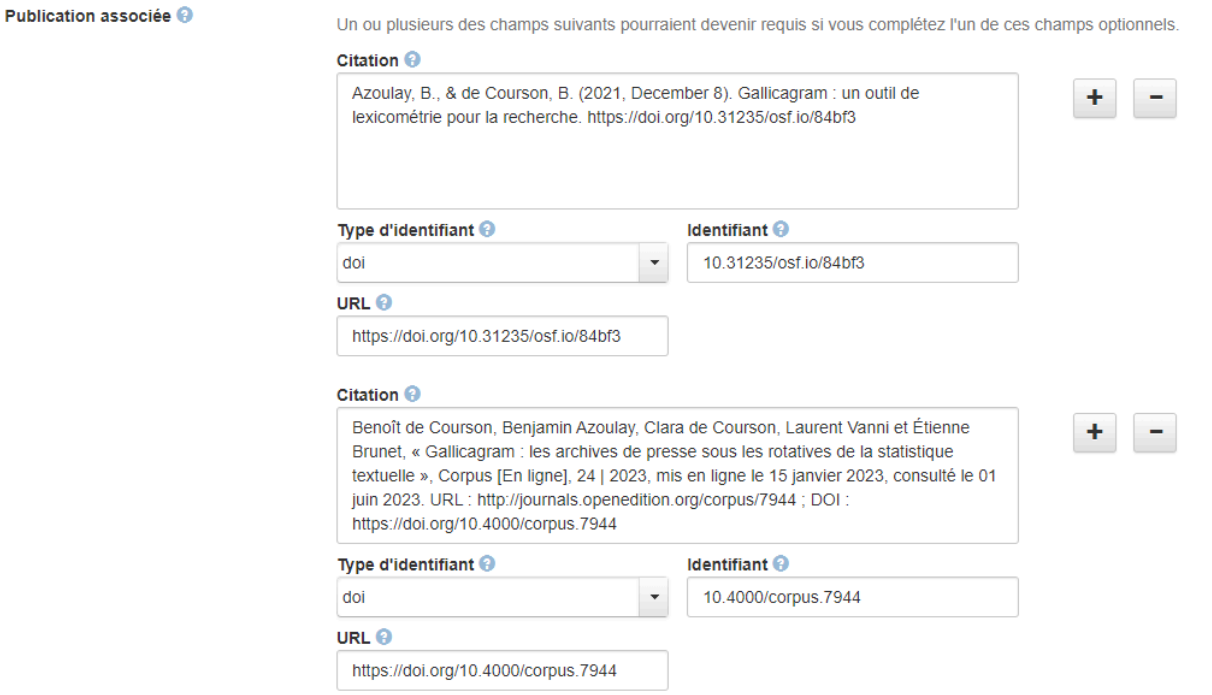

***Type de données*** **:** *Dataset*

***Origine des données*** **:** Text Corpus

***Etape du cycle de vie*** **:** Indiquer Original Release ou bien le numéro de version



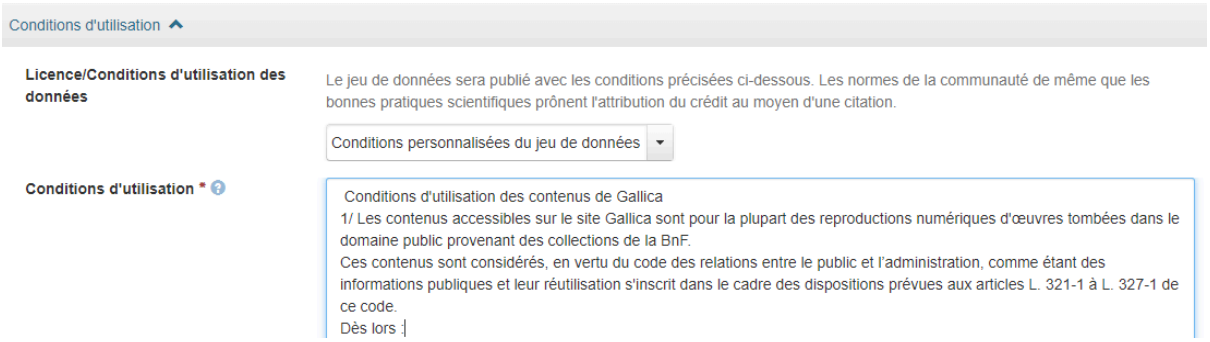
***Publication associée*** : Pour indiquer les publications, utiliser un style de citation (règles de formatage d'une source pour les documents académiques). Le style APA définie par l'American Psychological Association est aujourd'hui préféré par les universités françaises. Indiquer également leurs identifiants et les url liées.




***Période couverte*** : indiquer les dates des documents au format ISO 8601 année (4 chiffres), mois, jour.
***Date de collecte*** : indiquer la date de collecte au format ISO 8601 année (4 chiffres), mois, jour. Il peut s'agir en *Début* de la date de la 1ere extraction et en *Fin* de la date de la dernière extraction.

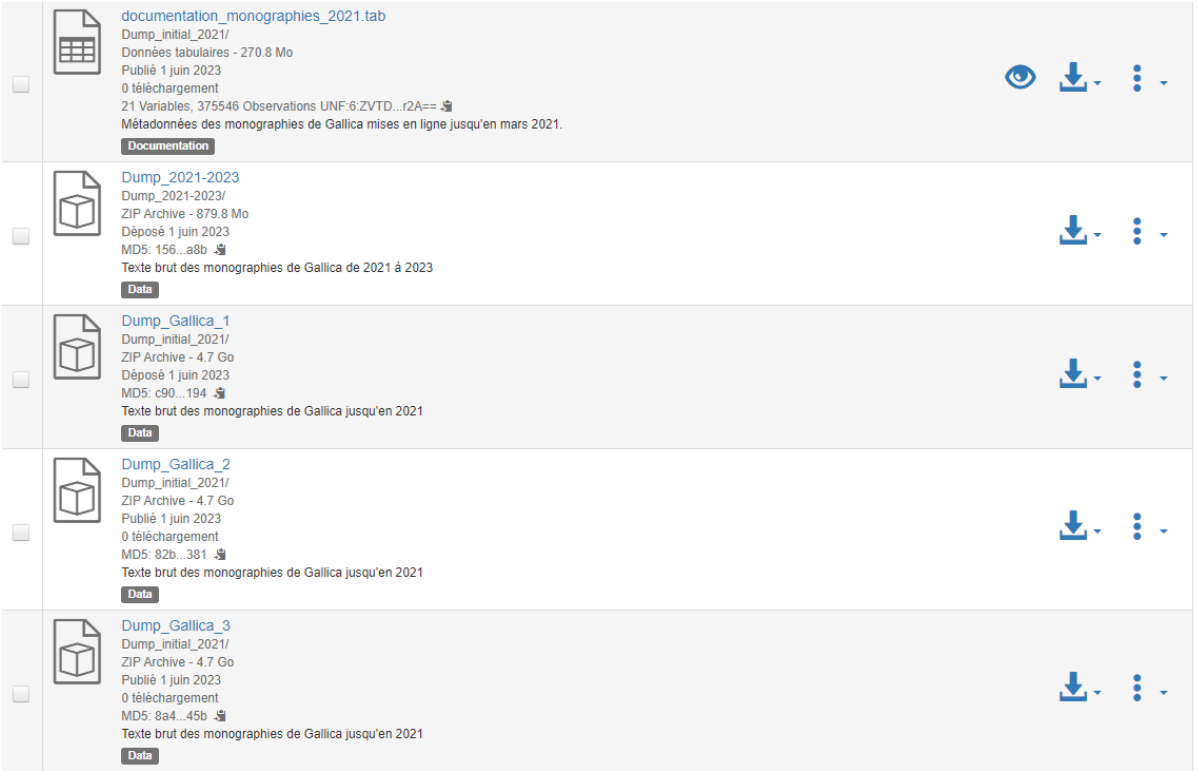
## 4. Renseigner les conditions d'utilisations Gallica

***Conditions d'utilisations*** : Renvoyer vers les conditions d'utilisations particulières des contenus de Gallica




## 5. Verser les fichiers

Choisir un nommage clair explicite. Ici le nom documentation_monographies_2021 renvoie au fichier tabulaire affichant les métadonnées de chaque monographie océrisée. Le mot métadonnées n'aurait pas convenu, car le fichier lui-même devient une métadonnée du jeu de données dans ce dépôt.



Créer une arborescence structurée par date de mise à jour et l'afficher dans le menu suivant.


## 6. Versions du jeu de données

Afficher les versions pour s'assurer de la logique de publication et du statut du jeu de données.

# Annexe 3

# Guide pour la publication sur le SSHOC Marketplace et Zenodo

**Table des matières**



## Publication sur le SSHOC Marketplace

## Créer un compte et un jeu de données

Créer un compte à cette adresse en signant avec un fournisseur d'identité préexistant ou à partir d'une authentification ORCID le cas échéant, sinon avec un compte mail individuel, puis créer un jeu de données avec *Create Dataset*.

### Edit Dataset

***Label :*** Indiquer un titre explicite du jeu de données
***Version :*** Indiquer Original Release si aucune modification du jeu original n'est envisagée, mais seulement l'ajout des fichiers de monographies mis en ligne chaque année. Sinon versionner au fil des mises à jour en choisissant par exemple le nommage V1, V2, pur des versions majeures et V1.1, V1.

***Description*** *:* utiliser le markdown pour la mise en forme
En particulier les balises  et  pour la mise en gras, la balise <br> pour le passage à la ligne et les balises <br></br> pour le saut de ligne.
***Accessible at*** : indiquer l'url du dépôt sur recherche.data ou sur api.bnf sinon.
**ID Service :** indiquer le type d'identifiant pérenne du dépôt **DOI ,** puis dans Identifier indiquer l'identifiant lui-même

## Dates

***Date last updated*** : ajouter la date de dernière mise à jour du jeu de données

## Actors

***Role*** : Choisir Contributeur et indiquer les noms des personnes responsables du jeu de données et auparavant référencées dans le SSHOMP.
Si un auteur n'est pas référencé, le créer en utilisant le bouton *Create Actor* en haut à droite
***Provider*** : Indiquer Bibliothèque nationale de France

## Properties :

***Property type*** **:** Choisir en priorité **Activity** pour utiliser TaDiRAH, la Taxonomy of Digital Research Activities in the Humanities. Si le terme lexical cherché n'y figure pas utiliser alors le *Property type* **Keywords** et créer chacun des mots clés nécessaires.
***Property type*** **:** choisir **Language** pour indiquer la langue des données
***Property type*** **:** choisir **Terms of use url** pour indiquer une licence particulière telle que la licence Gallica
***Property type*** **:** choisir **See also** pour indiquer les liens vers les outils à utiliser pur l'extraction des données (API, worflows ou autre)

## Related items :

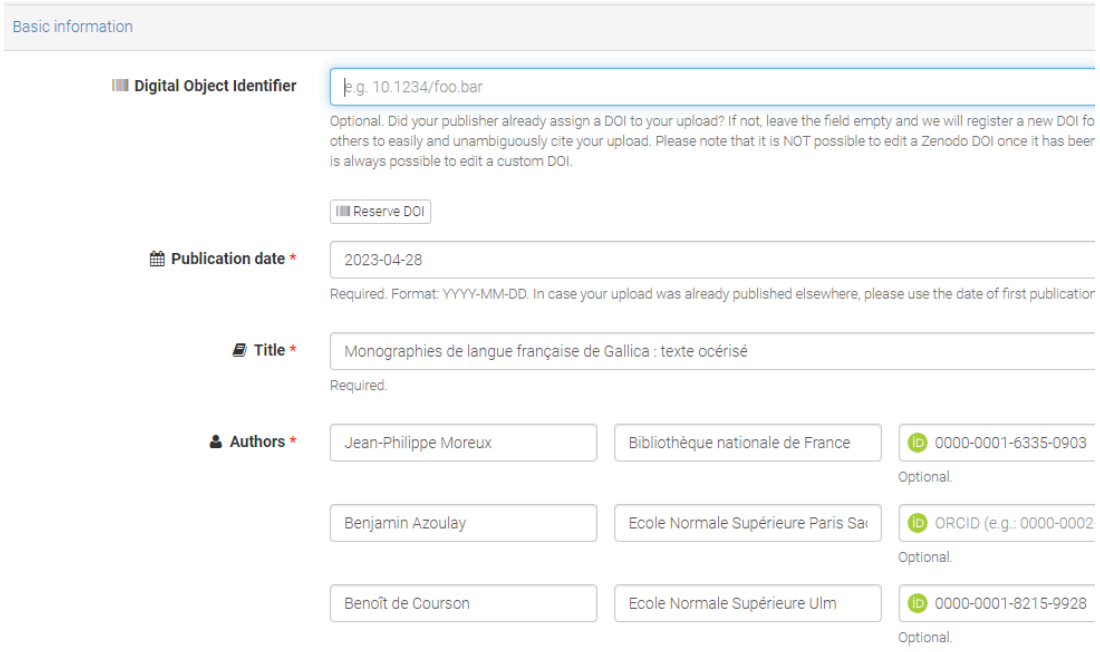
***Relation type*** : Choisir **is related to** pour indiquer une publication liée

Si une publication n'est pas référencée, choisir dans le menu de choix des éléments *Create Publication* et indiquer les métadonnées essentielles (activity/

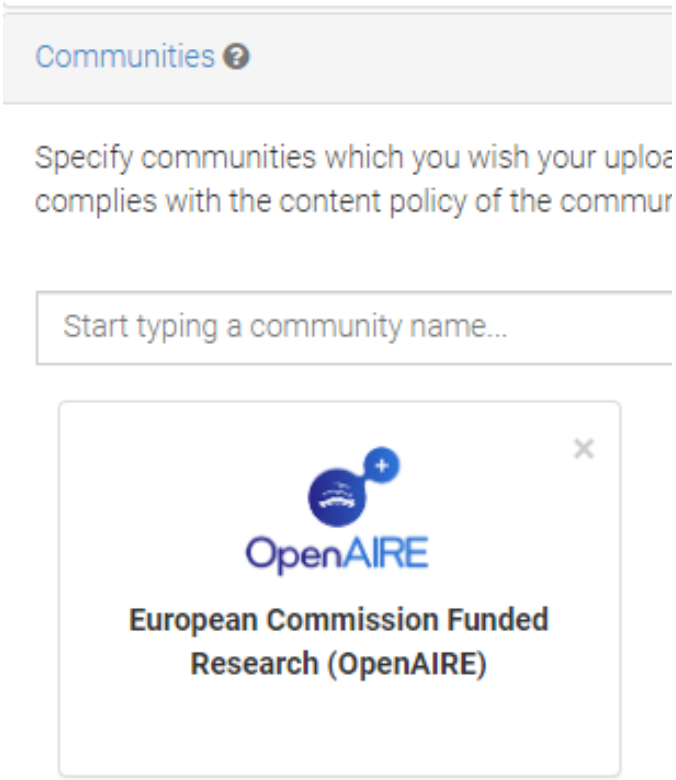
keyword, language, publisher, place, année, nom de la publication, volume/issues, pages).

## Publication sur Zenodo

**Dépôt de fichier :** Sur Zenodo, le dépôt d'au moins un fichier est obligatoire pour référencer le jeu de données. Le choix de déposer un fichier léger du jeu de données, en évitant le double dépôt sur Recherche.data et sur Zenodo, peut permettre de référencer le jeu complet. Dans le cas du *dataset* des OCR Galllica, le versement du fichier csv indiquant la liste des ARKS est une bonne option.
**Renvoi vers le DOI :** Indiquer le DOI attribué par le dépôt sur recherche.data
**Autres métadonnées :** Remplir les champs Auteurs (du jeu de données), Date de publication, Titre, Version, Language, Keywords (sans vocabulaire contrôlé), Subjects.

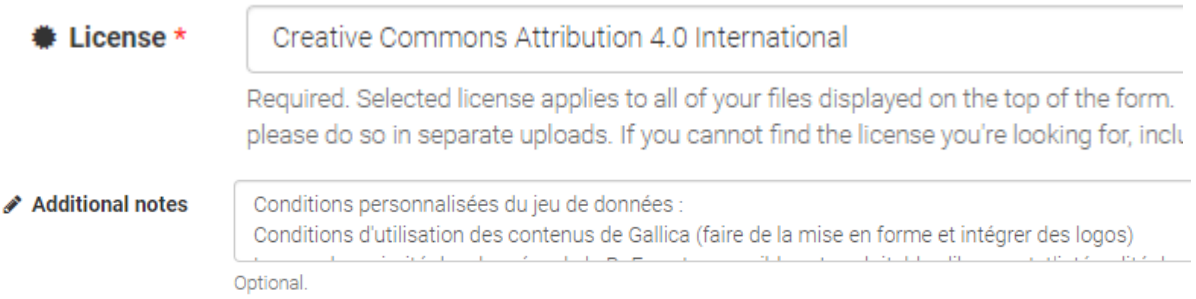



**Related/alternate identifiers** : pour indiquer les publications liées et leurs identifiants




**Communities** : Indiquer la communauté OpenAIRE, portail de découverte soutenu par l'ERIC DARIAH

**License :** Champ obligatoire. Choisir une licence proche comme la Creative Commons 4.0 International




**Additional notes** : permet d'indiquer la licence personnalisée Gallica

# Annexe 4

# Guide pour la publication sur le portail api.bnf.fr

## Table des matières

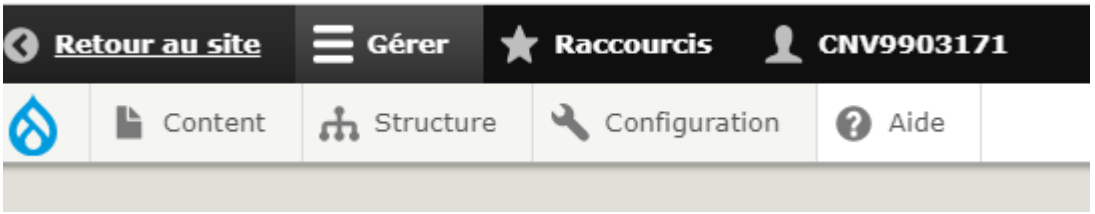


## Accès  au contenu dans le back-office Drupal

**Se connecter** au back office Drupal de api.bnf sur une session pro sur un poste pro :

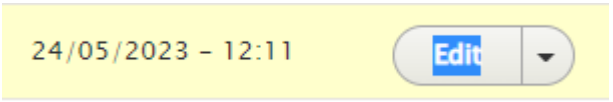
http://cms-api-adm.bnf.fr / Identifiants de connexion = identifiants Windows

**Cliquer sur Content** en haut à gauche pour afficher la liste des pages



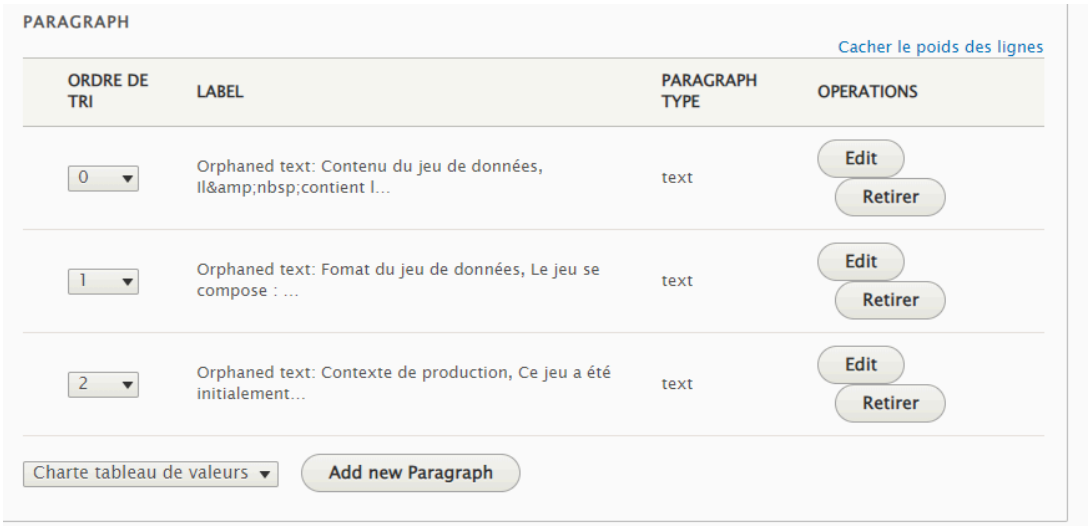

**Cocher:** Dump Gallica : ocr des monographies


**Cliquer sur Edit pour éditer  le contenu**



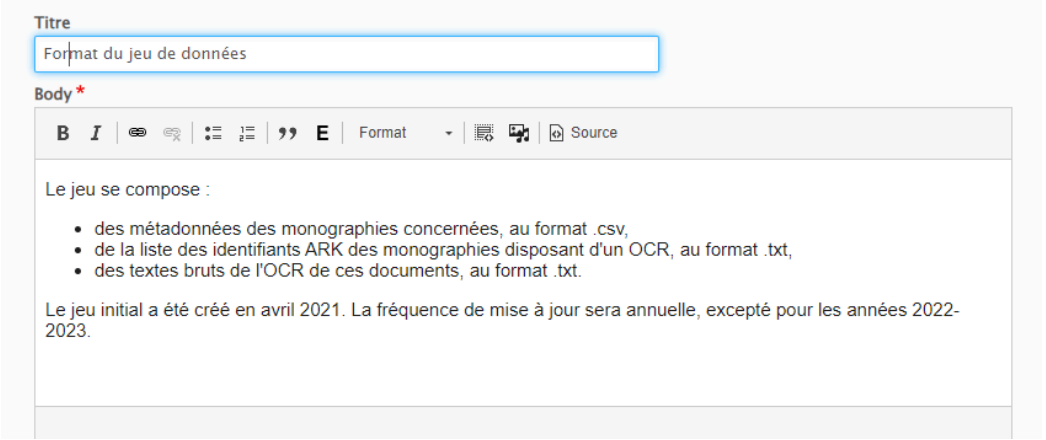
# Front

## Cliquer  sur Edit pour éditer chacune des trois sections

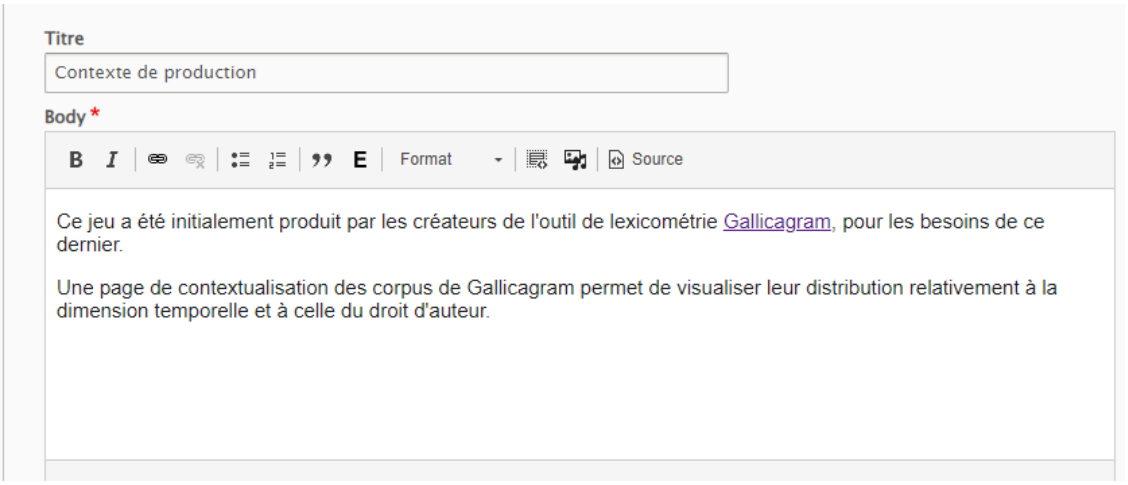



## Contenu du jeu de données :


## Format du jeu de données :

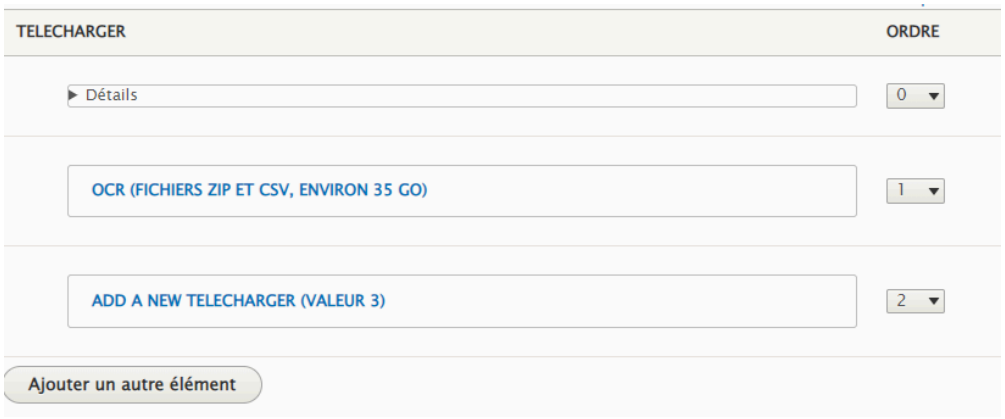


**Contexte de production :**




**Télécharger :**

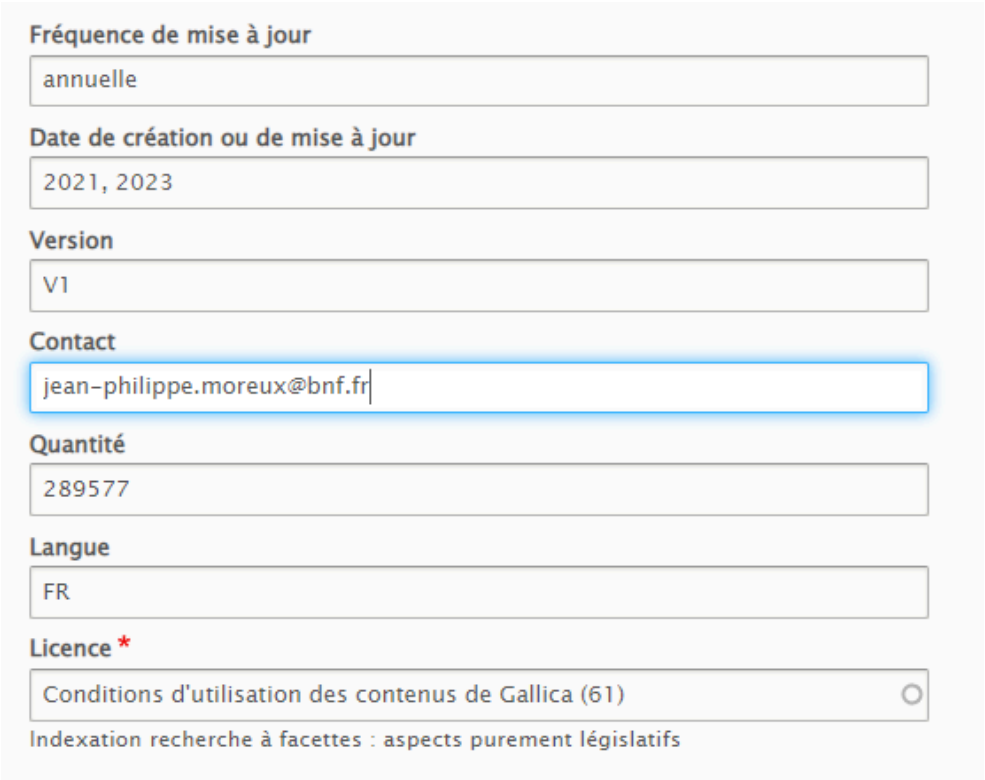
Insérer le lien GoDrive vers l'arborescence des fichiers. Cf. Annexe 1 paragraphe 2.6.2




**Ressources :**

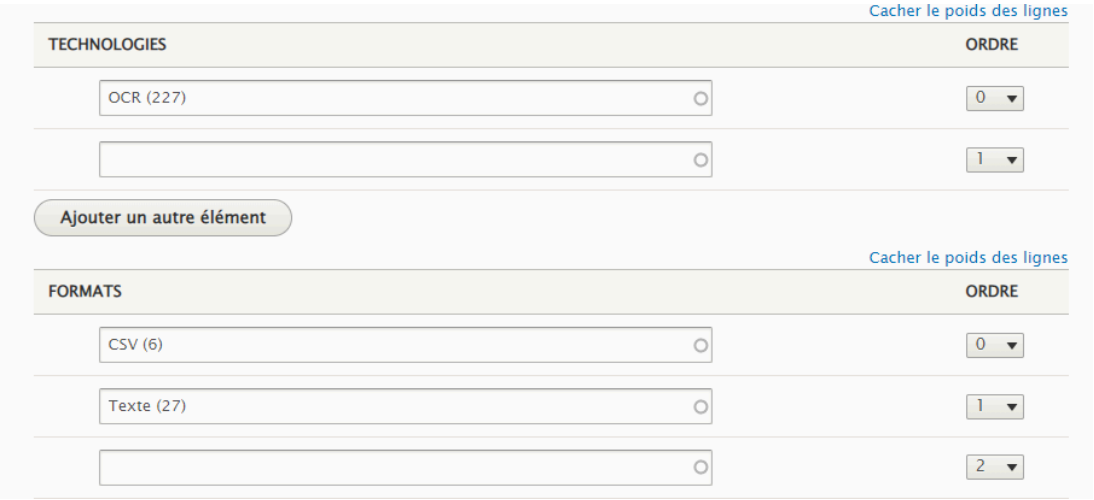
Pour le renvoi vers des pages éditoriales en lien si besoin.

# Fiche technique :



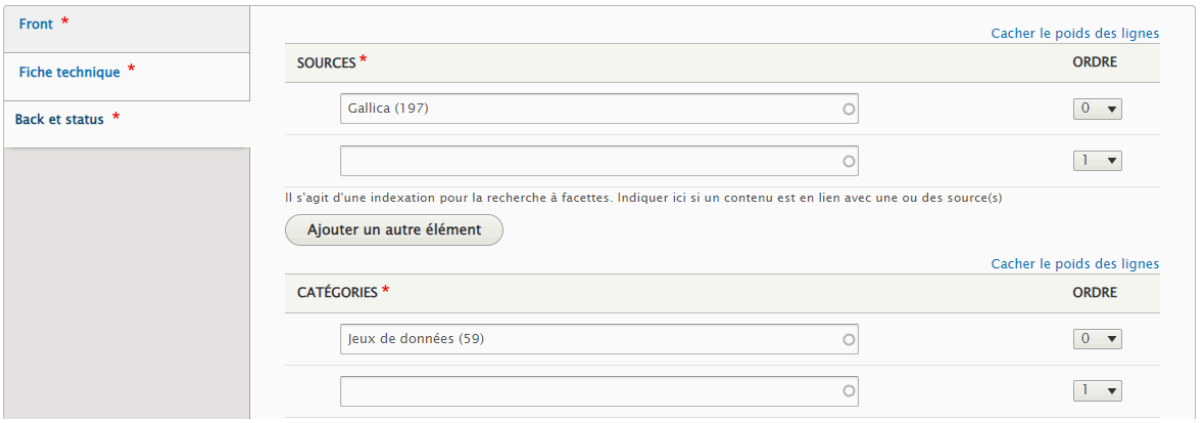

**Date de création ou de mise à jour :** ajouter derrière la précédente la dernière date de mise à jour

**Technologies et formats :**




**Sujets :**

# Back et Status

# Annexe 5 : Comparaison de plateformes de données publiques

| - Nom Entrepôt<br>- Gouvernance | - Usages et publics cibles | - Accès aux données<br>- Licences | Volumétrie | Identification pérenne | Standard de métadonnées | Statistiques | Services associés<br>API, Moissonnage | - Visu données<br>- Perspectives |
|---|---|---|---|---|---|---|---|---|
| **data.gouv (2011)**<br>- Etat<br>- Etalab, département de la direction interministérielle du numérique (DINUM) | - Données publiques administratives France<br>- Utilisé par les administrations territoriales | - Restriction d'accès possible<br>- Licence Ouverte version 2.0 ou licence ODbL uniquement | - Sans limite formelle de volume, mais upload difficile au delà de qq gigas.<br>Pas adapté > de qq dizaines de Go | - Attribution d'un id propre à data.gouv non pérenne | - Schéma DCAT (normalisation catalogues d'informations publiques) | - Stats de réutilisations uniquement déclaratives de la part des réutilisateurs. | - Moissonnage des portails extérieurs en schéma DCAT, moissonne les jdd de dataculture<br>- Un portail lié dédié aux API<br>- automatisation de la publication possible | - Préviz avec explore.data.gouv de données fichier au format txt/csv non compressés et < à 100 Mo<br>- Préservation données sans limite |
| **data.culture (2016)**<br>- Ministère de la Culture<br>- Données de l'administration, (services à compétence nationale, services déconcentrés) | - Données publiques des institutions culturelles France<br>- recherches patrimoniales | Licence Ouverte version 2.0 / ODbL uniquement<br>- pas de dépôt pour les ét. sous tutelle, référencement seul | - Visualisation pour 240 mo max par fichier csv, ods ou txt (json, rdf xml non) / volume jdd non limité | - Attribution d'un id propre à data.culture non pérenne | -Métadonnées basiques | - Pas de statistiques pour ét. sous tutelle dont les jdd sont uniquement référencés. | - Moissonne les ét. publics qui se sont signalés et déposent sur data.gouv<br>- Dataviz automatique à certaines conditions | - Réunion envisagée pour que la publication des ét. sous tutelle se fasse sur data.culture (avec mutualisation des financements)<br>- Projet de dvt d'un portail data.culture sur data.recherche |

# Annexe 6 : Comparaison d'entrepôts de données de la recherche

| Nom Entrepôt<br>Gouvernance<br>Préservation | Usages et publics cibles | Accès aux données<br>Licences | Volumétrie | Identification pérenne | Standard de métadonnées | Stats | Services associés<br>Automatisation<br>Moissonnage | Remarques<br>Perspectives |
|---|---|---|---|---|---|---|---|---|
| **recherche.data.gouv (2022)**<br>- MESRI, CNRS, INRAE, CEA et Universités.<br>- Logiciel Dataverse<br>- Préservation 5 ans | - Généraliste<br>- Données de la recherche plus large que SHS au sens données produites dans contexte de recherche | - Restriction d'accès<br>- Licences Ouverte, ODbL, CC-BY ou licence personnalisée | - 50 Go par fichier / jdd sans limite de volume | -DOI attribué si dépôt<br>- Possible de signaler et renvoyer vers un ID préexistant<br>- un DOI par version | - schéma DDI ((Data Documentation Initiative) | - Stats consultation et téléchargement | - jeux de données soumis à curation<br>- espace institutionnel 5 To en auto-dépôt avec engagement RH de la BnF<br>- moissonne seulement les solutions Dataverse | - Transforme le csv en .tab + présente de la preview<br>- Automatisation publication : outil de dépôt local vers Recherche Data Gouv |
| **Zenodo (2013)**<br>- CERN, Suisse (commission européenne)<br>- inclus dans EOSC<br>- Préservation 20 ans | - Généraliste<br>- Données de la recherche / Europe<br>- anglophone | - Restriction d'accès<br>- licences cc-by ou licence personnalisée | - 50 Go par jdd | - dépôt fichier obligatoire<br>- DOI attribué si pas d'ID préexistant.<br>- un DOI par version | - schéma DDI | - Stats consult et téléchargement | - Synchronisation GitHub / ORCID<br>- Moissonné par OpenAire (Dariah)<br>- FAQ | - Présence d'une communauté Dariah<br>- Simple d'utilisation |
| **Nakala (2014)**<br>Huma-Num / CNRS<br>Dépôt niveau avancé en partenariat CINES (archivage pérenne) | - Généraliste<br>- Données de la recherche /France | - Restriction d'accès<br>- licences prédéfinies ou personnalisée | Sans limite de volume | - DOI attribué systématiquement<br>- Pas de DOI versionné | - schéma DC | - pas de statistiques | - Moissonnable via Isidore<br>- pas de données sensibles / sous droits | - Moissonnage OpenAire KO<br>- Simple d'utilisation |

# Annexe 7 : Liste des sigles et abréviations

— ANR : Agence nationale de la recherche

— ARK : Archival Resource Key

— API : Application Programming Interface

— BDLI : Bibliothèques de dépôt légal imprimeur

— CCFr : Catalogue collectif de France

— CoSO : Comité pour la science ouverte

— CINES : Centre informatique national de l'enseignement supérieur

— CIRAD : Centre de coopération internationale en recherche agronomique pour le développement

— DARIAH : Digital Research Infrastructure for the Arts and Humanities

— DOI : Digital Object Identifier

— DCAT : Application Profile for data portals in Europe

— DDI : Data Documentation Initiative

— DINUM : Département de la direction interministérielle du numérique

— EOSC : The European Open Science Cloud

— ERIC : European Research Infrastructure Consortium

— FTP : Protocole de transfert de fichiers

— FAIR : Facile à trouver, accessible, interoperable, réutilisable

— HAL : Hyper article en ligne

— HTML : HyperText Markup Language

— HTTPS : HyperText Transfer Protocol Secure

— HTR : Handwritten Text Recognition

— INIST : Institut de l'Information Scientifique et Technique

— IIIF : International Image Interoperability Framework

— IRD : Institut de recherche pour le développement

— IST : information scientifique et technique

— LRN : Loi pour une République numérique

— LIBER : Ligue des Bibliothèques Européennes de Recherche

— OAI-PMH : Open Archives Initiative Protocol for Metadata Harvesting

— OCR : Optical Character Recognition

— ORCID : Open Researcher and Contributor ID

— PDF : Portable Document Format
— PGD : Plan de Gestion des Données
— PNSO : Plan national pour la science ouverte
— RDF : Resource Data Framework
— REEN : Réduire l'empreinte environnementale du numérique
— RDG : Recherche Data Gouv
— SPAR : Système de Préservation et d'Archivage Réparti
— SPARQL : SPARQL Protocol and RDF Query Language
— SSHOC : Social Sciences and Humanities Open Cloud
— TAL : Traitement Automatique des Langues
— TEI : Text Encoding Initiative
— TDM : Text Data Mining
— RGPD : Règlement général sur la protection des données
— URI : Uniform Resource Identifier
— XML : eXtensible Markup Language
— WARC : Web ARChive
— W3C : World Wide Web Consortium

## Table des matières