Didak'TIC

Convergence des cultures de l'information

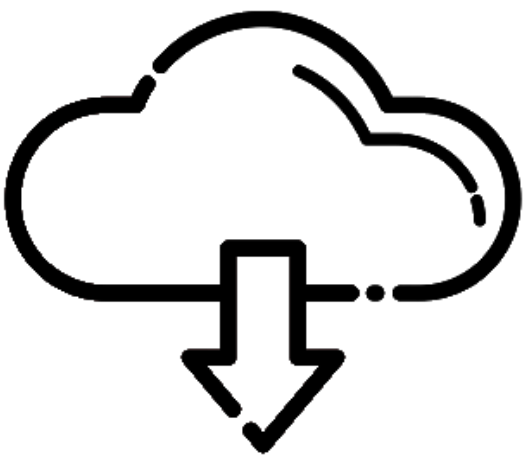
n° 5

2021

FAIR OPEN DATA

Une revue proposée par le Master 2
Management et Valorisation de
l'Information Numérique (MAVINUM)
Promotion 2020-2021

71ème section CNU : sciences de l'information et de la communication (SIC)



LOADING DIDAK'TIC

« FAIR » de l'open data un enjeu

Chères lectrices, chers lecteurs,

Cette année, nous avons décidé de vous faire voyager au cœur de la donnée, de cette information qui circule et se partage.
Si l'*open science* favorise et encourage ce mouvement, ce dernier n'est pas anarchique et doit se « FAIR » dans le respect de certaines règles et grâce à des moyens mis à la disposition des chercheurs.

Comprendre pourquoi l'*open data*

Les données ouvertes sont, comme l'indique leur nom, des données numériques, d'origine privée ou publique, laissées libres d'accès et d'usage. Elles sont généralement produites par un établissement public ou une collectivité. Les notions de **bien commun** et de **droit d'accès à l'information** s'inscrivent au travers de ce mouvement d'*open data.*

Au-delà de cet aspect philosophique, l'intérêt de l'ouverture et du partage des données de la recherche est multiple. La **visibilité des travaux** des chercheurs est augmentée, favorisant ainsi la notoriété scientifique. L'ouverture des données favorise également l'évaluation par les pairs et apporte de **l'intégrité** aux résultats des recherches. De plus, dans ce domaine qu'est la recherche

où l'aspect financier est primordial, la réutilisation des données permet une **rentabilité** accrue et un gain de temps dans l'étude menée. Pour finir, l'autre raison poussant vers l'adoption de l'open data est que ce libre accès et la diffusion des données de la recherche peut amener des chercheurs à développer d'autres théories et ainsi aboutir à découvrir autre chose que ce qui était recherché. En d'autres termes, l'*open data* favorise la **sérendipité**.

Favoriser l'*open data*, oui mais pas n'importe comment

De la publication scientifique, à l'entrepôt de données, en passant par les TGIR, de nombreux moyens sont mis à la disposition des chercheurs pour partager leurs données. Cependant, des **règles** sont à respecter et de bonnes pratiques à adopter.

Que ce soit le RGPD pour protéger les données à caractère personnel, ou bien encore la mise en œuvre d'un plan de gestion des données, en passant par le stockage dans des entrepôts de « confiance », des **lignes directrices** existent afin de « FAIR » de l'*open data* un principe de la science ouverte à adopter par tous. Au-delà de leur partage, ces principes ont également pour but de pérenniser l'exploitation, et la ré-exploitation, de l'information qu'elles véhiculent.

Une ouverture illusoire

Mais ne nous leurrons pas chères lectrices et chers lecteurs, toutes les données ne peuvent (ou ne doivent ?) pas être partagées et encore moins réutilisées. La crise sanitaire du SARS-COV 2, et le débat autour du statut du brevet comme bien commun de l'humanité, en est un parfait exemple. Si la question ne se pose pas pour les données de la recherche issues de fonds publics, celle résultant de financement privés, ou de recherches "sensibles", reste soulevée.

Mais finalement, la vraie question n'est-elle pas plutôt de se demander s'il est nécessaire de tout partager ? Et surtout de tout réutiliser…

Nous vous invitons à nous rejoindre dans ce flot de données que sont les articles de ce nouveau numéro de la revue des étudiants du Master MAVINUM et que nous partageons avec vous avec plaisir.

Bonne lecture.

Nadia NEMER-BRUN

LES ACTUALITÉS DE L'OPEN RESEARCH DATA

ASSOCIATION

L'IFLA ('International Federation of Library Associations and Institutions) s'associe à la Wikimédia Fondation, pour « développer un réseau gratuit de connaissances sur tous les sujets ».La fédération internationale des associations et institutions de bibliothèques (en français), compte, entre autres, fournir des données structurées basées sur les principes des données FAIR afin de pouvoir transformer de manière transparente les données en information pour créer des connaissances FAIR et fournir des structures de gouvernance ouvertes et les intégrer dans les systèmes existants.

CHIFFRES

En 2014, l'INIST estimait que "90% des données de la recherche sont stockées sur les disques durs locaux et potentiellement non réutilisables par d'autres". Selon l'édition 2020 du Baromètre de la Science Ouverte (BSO), 56 % des 156 000 publications scientifiques françaises publiées en 2019 sont en accès ouvert en décembre 2020. Le taux observé en décembre 2019, relatif aux publications réalisées en 2018, n'était que de 49%...

ÉVÈNEMENT

La Semaine internationale de l'Open Access 2021 aura lieu du 25 au 31 octobre, sur le thème « L'importance de notre façon d'ouvrir la connaissance: construisons l'équité structurelle », aligné sur la recommandation de l'UNESCO sur la science ouverte et la question du libre accès pour un développement équitable de la recherche.

La Semaine internationale de l'Open Access est l'occasion pour l'ensemble de la communauté des chercheurs de coordonner les actions visant à faire de l'ouverture la configuration par défaut de la recherche, et de veiller à ce que l'équité soit au centre de ce travail.

http://www.openaccessweek.org/

JUSTICE

Le calendrier d'application de l'Open Data des décisions de justice se précise et s'étale de septembre 2021 à décembre 2025 grâce à l'arrêté du 28 avril 2021 pris en application du décret n°2020-797 du 29 juin 2020. L'ouverture se veut progressive, à la fois par degré de juridiction et par type de contentieux. Ces étapes permettent entre autres d'assurer "la protection de la vie privée et de la sécurité des personnes". Les décisions d'ordre judiciaire (en opposition aux décisions administratives) seront soumises au régime de l'occultation (anonymisation et pseudonymisation).

Du point de vue de la mise à disposition, elle pourra se faire sous forme électronique par le biais d'un portail internet sous la responsabilité du ministre de la Justice et via les sites du Conseil d'Etat et de la Cour de cassation.

SOMMAIRE

ILS ONT PARTICIPÉ !

Gérard BODO :
Douala, CAMEROUN

Julien COLIN :
Clermont-Ferrand, FRANCE

Pierre ERNENWEIN :
Montpellier,FRANCE

Selin GÜDER :
Orléans, FRANCE

Chacha MAHFOUD :
Paris, FRANCE

Pauline MOLLALIOGLU :
Paris, FRANCE

Nadia NEMER-BRUN :
Montélimar, FRANCE

Daphné NOVIANT :

Zaïna NZOGU-LOZANO :
Reims, FRANCE

Pierre POUGET :
Bordeaux, FRANCE

Guénola TARDY-JOUBERT :
Paris, FRANCE

Le magazine Didak'TIC
ITIC Bâtiment Marc Bloch (Bât. E) Université Paul-Valéry Montpellier - Route de Mende 34199 Montpellier Cedex 5 - www.didaktik.fr

DÉPÔT LÉGAL de la revue Didak'TIC : décembre 2017, ISSN 2605-8812,
DATE DE PARUTION de Didak'TIC n°5 : octobre 2021.

Didak'TIC

DIRECTRICE DE LA PUBLICATION : Lise VERLAET.
PILOTES DU MAGAZINE : Nadia NEMER-BRUN, Gérard BODO.
RÉDACTION
Directeurs de la rédaction : Hans DILLAERTS, Lise VERLAET.
Rédacteurs et Rédactrices : Gérard BODO, Flora CHONG, Julien COLIN, Cécile DESCAMPS, Selin GUDER, Laura HOUCK, Nadia NEMER, Guillaine POON, Pierre POUGET, Guénola TARDY-JOUBERT, Chloé TRAVERS, Laurent TRIPOLI, Zaïna NZOGU-LOZANO.

CONCEPTION TECHNIQUE
Pour la partie PRINT :
Responsable : Nadia NEMER-BRUN
Maquettiste : Chacha MAHFOUD, Zaïna NZOGU-LOZANO, Guénola TARDY-JOUBERT, Julien COLIN, Nadia NEMER-BRUN
Pour la partie WEB :
Responsable : Gérard BODO
Webistes : Chacha MAHFOUD, Zaïna NZOGU-LOZANO, Guénola TARDY-JOUBERT, Pierre ERNENWEIN, Selin GUDER, Daphné NOVIANT, Pierre POUGET, Gérard BODO

COMITÉ DE VALIDATION : Hans DILLAERTS, Lise VERLAET, Mireille BACHELOT.
DIFFUSION Imprimé en France

ILLUSTRATIONS COUVERTURES
1ère de couverture : Open data database integration api internet technology concept. *WrightStudio*
4ème de couverture : Doors opening to reveal beautiful sky against lines of blue blurred letters falling. *Vectorfusionart*

DOSSIER : SCIENCE OUVERTE ET PUBLICATION SCIENTIFIQUE

Data paper, un type de publication de données scientifiques en accord avec les *principes FAIR*.

Guillaine POON

Le data paper permet de mettre au premier plan les données qui servent habituellement à réfuter ou confirmer des hypothèses de recherche dans les publications scientifiques. Ces données sont mises en lumière en détaillant à la fois leur obtention et leur exploitation. Elles font partie intégrante d'une publication suivant les normes classiques de rédaction avec des métadonnées associées. Dans une logique d'ouverture des données, le data paper est stocké dans des entrepôts, favorisant le partage et la consultation des données décrites.

Les **données** font **partie intégrante** d'une **publication scientifique,** pourtant le travail de fond conséquent effectué en amont pour leur obtention et leur exploitation n'est pas systématiquement valorisé - selon les disciplines et les politiques éditoriales des revues - dans un article classique. Un nouveau type de publication, le ***data paper***, est en plein essor depuis ces dernières années et répond à cet enjeu. Il bouleverse les codes de l'article standard. Bien que **formalisé** dans la forme, son objectif premier n'est pas de tirer des conclusions basées sur l'analyse de données mais bien de décrire de façon exhaustive ces données, à savoir les processus de collecte, de traitement et de production des données. À savoir les processus de collecte, de traitement et de production des données. L'intérêt de publier dans un data paper est multiple. Le premier objectif est d'être un vecteur de reproductibilité et de transparence pour la science. De là découle naturellement une valorisation des données et de son, ou ses, créateur(s) en raison de l'augmentation de leur visibilité respective. En effet le ***data paper*** est une **publication citable** et **repérable** facilement car intégrée directement dans des revues classiques ou dans des ***data journals,*** revues spécialisées dans ce type de publication. La **réutilisation** facilitée des données présentées dans l'article concourt à l'objectif des **principes FAIR**, à savoir une **ouverture des données** de la recherche.

Qu'est-ce qu'un data paper ?

Il s'agit d'une publication **évaluée par les pairs** [1]. Quant à son contenu, on trouvera des **descriptions faites via des métadonnées sur des jeux de données issus de la recherche scientifique**. Ayant été soumis à un processus de révision par un comité de lecture, le *data paper* est aussi un moyen d'assurer la **qualité des données** de recherche. De plus on communique l'existence de ces données à la communauté scientifique et par ce biais une reconnaissance **des auteurs** [2] via lien de paternité des données est faite. Ces données sont structurées et lisibles par un humain.

Où est-il publié ?

Dans des ***data journals***, revues qui dédient la totalité de leurs publications à ces *data papers* ou bien dans une revue classique les publiant également. Ceux-ci sont évalués par le **comité de lecture de la revue les hébergeant**.

D'après DoRANum (Données de la Recherche : Apprentissage Numérique), les principaux critères d'évaluation [3] sont : l'importance et l'originalité des données, leur potentielle valeur de réutilisation, la qualité et la fiabilité des données, l'accès aux données, la qualité et rigueur de la méthode de collecte des données, le choix des métadonnées descriptives et formats et le respect des normes classiques de rédaction.

Quelles sont les similarités et différences observées avec un article classique ?

Il adopte une **structure standardisée** [4] totalement ou partiellement, avec des sections obligatoires dans ce dernier cas de figure, imposée par la revue. Le *data paper* contient des métadonnées sur les jeux de données ainsi que sur le *data paper* en lui-même. Des identifiants normalisés sont à joindre dans la publication, comme le DOI (*Digital Object Identifier*) ou moins fréquemment, ORCID (*Open Researcher and Contributor ID*).

Trois sections sont systématiquement présentes : une introduction, une description des données et la méthodologie avec l'équipement nécessaire pour la production des données.

La section description des données peut contenir les informations suivantes : le type de données, la couverture géographique et temporelle, les conditions d'accès et de **réutilisation**, la taxonomie, l'intérêt du jeu de données ainsi que les méthodes utilisées pour garantir un contrôle qualité.

D'autres sections optionnelles et communes avec un article classique sont possibles comme les résultats/ discussion, la conclusion et les annexes.

D'autres sont spécifiques au *data paper* et facultatives comme la réutilisation potentielle des données par d'autres chercheurs ou d'autres publics, leur accessibilité et leur disponibilité (lien vers l'entrepôt de données, les droits et restrictions) et leur valeur et validation (au sens de la qualité de contrôle selon le type d'évaluation utilisé).

Son but diffère des articles standards, il n'a pas une fonction de compte-rendu et ne sert pas non plus à prouver ou à réfuter des hypothèses formulées. Les articles classiques se focalisent sur les résultats obtenus et dans ce contexte, les données servent d'appui de preuve de ces recherches. Par opposition, les *data papers* informent et redirigent vers les données décrites, stockées dans des entrepôts prévus à cet effet.

Le *data paper* a donc un **rôle descriptif** et prend en compte l'environnement de production des jeux de données à

savoir les conditions et le contexte de leur collecte ainsi que leur utilité potentielle. Il **améliore la visibilité ainsi que la réutilisation et donc l'accessibilité des jeux de données** avec l'ajout de métadonnées supplémentaires.

En effet, il faudra distinguer les métadonnées du *data paper* en lui-même qui décrivent des informations générales (auteur, organisation…) de celles décrivant les jeux de données (contenu des données).

Il faudra cependant nuancer cette différence de traitement des données avec le fait que la structure d'un data paper pouvant varier énormément d'une revue à une autre, la frontière distinctive en matière de contenu qui différencie un article classique d'un *data paper* peut s'avérer ambiguë. Les *data papers* deviennent eux-mêmes des sources de connaissances et de données.

LE RÔLE DU DATA PAPER DANS L'OUVERTURE DES DONNÉES

Le but premier des principes FAIR [5] est d'améliorer la réutilisation des données de la recherche. Le *data paper* y répond en partie en permettant une meilleure interprétation des données de recherche par machine ainsi qu'un transfert plus performant de ces données.

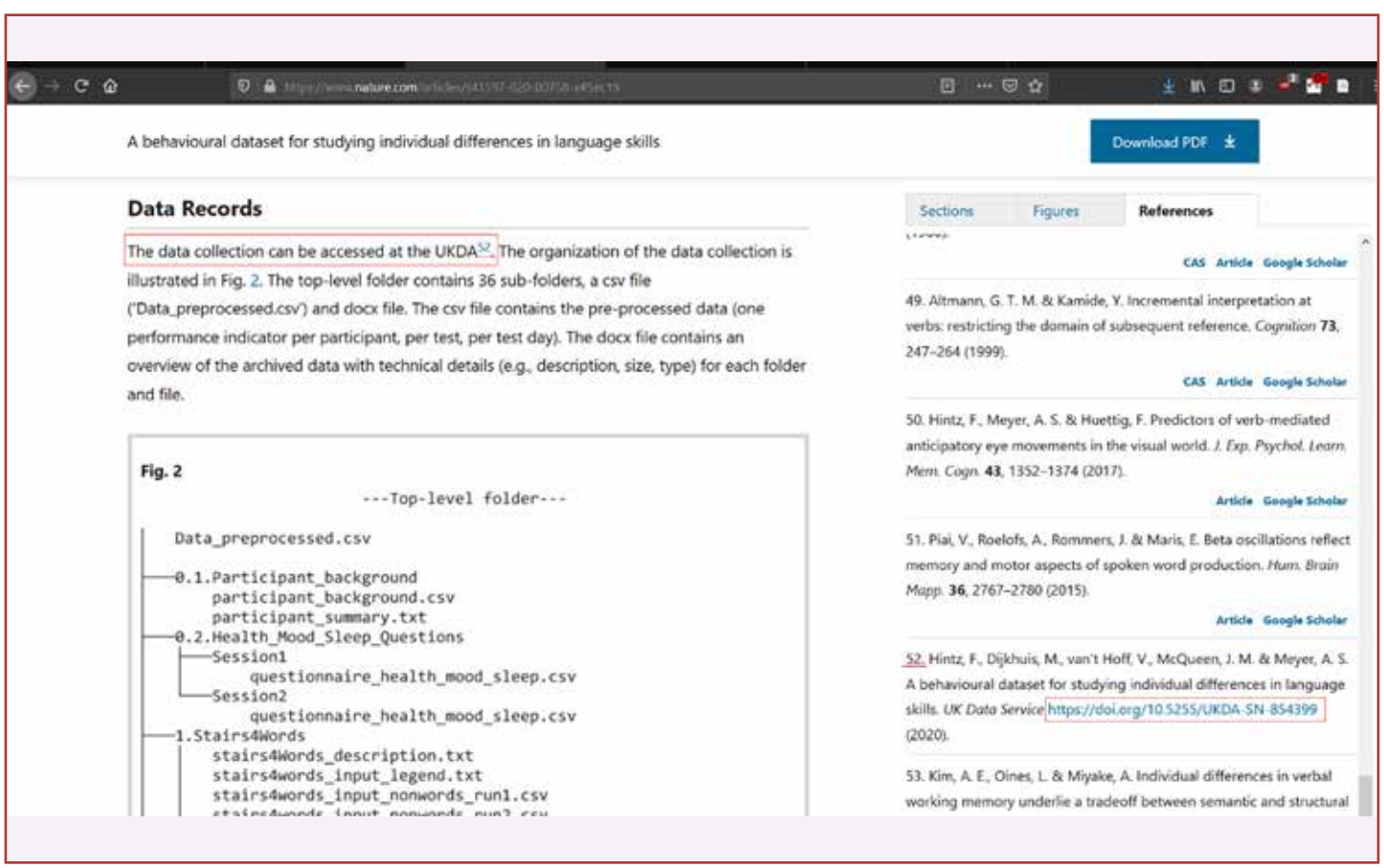
1 - Faciles à trouver :

- Les **métadonnées déjà présentes dans les jeux de données** (exemple de Nakala où il est possible d'ajouter des métadonnées suivant le standard Dublin Core) sont **enrichies** avec les métadonnées de description du data paper (suivant éventuellement des formats standards, selon le cadre imposé par la revue de la publication).
- **Les métadonnées et données du *data paper* sont répertoriées, indexées et conservées** sur la plateforme du *data journal.*

2 - Accessibles :

- Les métadonnées sont accessibles même si les données associées ne le sont plus. Bien qu'un lien soit établi entre métadonnées et jeux de données, les métadonnées du *data paper* n'ont pas besoin des jeux de données pour exister à elles seules.

Cette **conservation pérenne des métadonnées** est important car celles-ci continuent d'avoir une valeur informative par exemple pour montrer l'existence d'une production de données déjà faite, pour contacter les


A behavioural dataset for studying individual differences in language skills

Download PDF

Data Records

The data collection can be accessed at the UKDA[52]. The organization of the data collection is illustrated in Fig. 2. The top-level folder contains 36 sub-folders, a csv file ('Data_preprocessed.csv') and docx file. The csv file contains the pre-processed data (one performance indicator per participant, per test, per test day). The docx file contains an overview of the archived data with technical details (e.g., description, size, type) for each folder and file.

Fig. 2

```
                    ---Top-level folder---
  Data_preprocessed.csv
├──0.1.Participant_background
      participant_background.csv
      participant_summary.txt
├──0.2.Health_Mood_Sleep_Questions
   ├──Session1
      questionnaire_health_mood_sleep.csv
   └──Session2
      questionnaire_health_mood_sleep.csv
├──1.Stairs4Words
   stairs4Words_description.txt
   stairs4words_input_legend.txt
   stairs4words_input_nonwords_run1.csv
```

Sections Figures References

CAS Article Google Scholar

49. Altmann, G. T. M. & Kamide, Y. Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* **73**, 247–264 (1999).

CAS Article Google Scholar

50. Hintz, F., Meyer, A. S. & Huettig, F. Predictors of verb-mediated anticipatory eye movements in the visual world. *J. Exp. Psychol. Learn. Mem. Cogn.* **43**, 1352–1374 (2017).

Article Google Scholar

51. Piai, V., Roelofs, A., Rommers, J. & Maris, E. Beta oscillations reflect memory and motor aspects of spoken word production. *Hum. Brain Mapp.* **36**, 2767–2780 (2015).

Article Google Scholar

52. Hintz, F., Dijkhuis, M., van't Hoff, V., McQueen, J. M. & Meyer, A. S. A behavioural dataset for studying individual differences in language skills. *UK Data Service* https://doi.org/10.5255/UKDA-SN-854399 (2020).

53. Kim, A. E., Oines, L. & Miyake, A. Individual differences in verbal working memory underlie a tradeoff between semantic and structural

Exemple d'un data paper paru dans la revue en libre accès Scientific Data [6]

personnes ayant participé à la création des données si celles-ci sont indiquées dans ces métadonnées.

3 - Interopérable :

- **Les métadonnées et les données** utilisent un **langage formel, accessible, partagé** et largement applicable pour une représentation des connaissances compréhensible par la communauté.
- -les métadonnées et données font référence à d'autres métadonnées et données comme des liens vers d'autres ressources associées comme des publications de recherche, d'autres jeux de données…

4 - Réutilisable :

- Les **métadonnées et données** sont mises à disposition avec une **licence explicitant les droits de réutilisation**.

La réutilisation est généralement facilitée par les *data journals* qui autorisent pour la plupart du temps la publication de ces articles avec une **licence libre** (type CC-Zero et CC-By, c'est-à-dire respectivement sans et avec attribution de l'auteur avec comme unique contrainte de réutilisation).

- **Les métadonnées et données** sont fournies avec **leur source de production détaillée** : d'où elles viennent, qui citer, qui a généré ou collecté les données et comment celles-ci ont été exploitées.

L'URL de l'entrepôt où sont stockées les données présentées peut être placée dans une section dédiée à cet effet, dans une annexe, dans l'abstract ou bien encore dans les métadonnées. Certains *data papers* contiennent les données jointes en téléchargement directement dans le document.

Dans la section *Data Records*, une référence contenant un lien pérenne DOI vers l'entrepôt de données contenant les jeux de données présentés est joint.

Les **métadonnées en tant qu'informations précises servant de descripteurs des jeux de données** sont le noyau des data papers, autant pour la publication en elle-même que pour les données qu'elle explicite. C'est pour cette raison que celles-ci doivent adopter une **standardisation avec l'attribution d'un PID** (Persistent Identifier). Le DOI est un type de PID utilisé par des entrepôts de données (ex : Zenodo) pour faciliter la localisation, l'identification et la citation des données par la création d'un lien hypertexte unique et pérenne associé à un jeu de données.

Les *data journals* imposent l'utilisation d'identifiants standards pour les jeux de données (d'où l'importance du choix de l'entrepôt de données qui se charge de l'attribution de cet identifiant). Les métadonnées du *data paper* qui sont associées aux données doivent ainsi contenir le PID des données qu'elles complètent. L'attribution de ces identifiants aux *data papers* permet d'une part de créer un lien entre *data papers* et jeux de données et d'autre part de garantir l'accessibilité aux jeux de données.

En complément à sa valorisation des données par l'enrichissement de celles-ci via des métadonnées les explicitant, le *data paper* doit les rendre accessibles soit directement annexées dans la publication soit via un lien pérenne vers un entrepôt de données. Ce document permet soit de montrer la disponibilité du jeu de données ou de conserver la trace de la production des données décrites dans le cas où elles n'existeraient plus. L'auteur du *data paper* a un rôle d'argumentaire justifiant l'originalité et la portée du jeu de données pour inviter ses confrères à la réutilisation de ces données. Le *data paper* incluant les processus de collecte, de traitement et de production des données sert à valider la qualité des données, confortant la réutilisation potentielle par d'autres scientifiques. Ceci s'inscrit donc dans la démarche entreprise par la science ouverte, d'étendre les données de la recherche à la communauté scientifique et plus globalement au public.

▪ **Guillaine POON**

Références :

[1] *Pasteur, C.-I. (2019, novembre 25). Data papers : An emerging form of publication which contributes to the compliance with FAIR principles. Open science : évolutions, enjeux et pratiques. https://openscience.pasteur.fr/2019/11/25/data-papers-an-emerging-form-of-publication-which-contributes-to-the-compliance-with-fair-principles/*

[2] *Data papers et data journals – DoRANum. (s. d.). Consulté 17 mars 2021, à l'adresse https://doranum.fr/data-paper-data-journal/*

[3] *Les critères d'évaluation des data papers – DoRANum. (s. d.). Consulté 19 mai 2021, à l'adresse https://doranum.fr/data-paper-data-journal/criteres-evaluation-data-papers/*

[4] *Schöpfel et al. - 2019—Data Papers as a New Form of Knowledge Organizatio.pdf. (s. d.).Schöpfel, J., Farace, D., Prost, H., & Zane, A. (2019). Data Papers as a New Form of Knowledge Organization in the Field of Research Data. KNOWLEDGE ORGANIZATION, 46(8), 622-638. https://doi.org/10.5771/0943-7444-2019-8-622*

[5] *Les principes FAIR – DoRANum. (s. d.). Consulté 16 mars 2021, à l'adresse https://doranum.fr/enjeux-benefices/principes-fair/*

[6] *Hintz, F., Dijkhuis, M., van 't Hoff, V., McQueen, J. M., & Meyer, A. S. (2020). A behavioural dataset for studying individual differences in language skills. Scientific Data, 7(1), 429. https://doi.org/10.1038/s41597-020-00758-x*

(Source : Freepick.com)

Comment "FAIR" de la réutilisation des données un des piliers de la recherche scientifique : publier, déposer, partager.

Nadia NEMER-BRUN

Data paper, data journal, matériel supplémentaire, entrepôt de données… Si les années 90 et l'essor d'Internet ont vu naître et se développer différentes initiatives et programmes politiques autour du libre accès et des principes FAIRs, les solutions d'ouverture des données de la recherche se sont mises en place petit à petit pour aboutir à une liste de possibilités offertes aux chercheurs.

Mais comment ne pas se perdre dans ce labyrinthe du partage des données de la recherche ? Et pourquoi, finalement, "prendre le temps" de les rendre publiques ? Nous tenterons de nous frayer un chemin dans le dédale de la publication et du partage des données de la recherche en prenant pour modèle la Data Publication Pyramid, et plus précisément les deuxième, troisième et quatrième niveaux...

OPEN RESEARCH DATA : UN OBJECTIF, CINQ ENJEUX !

"L'ouverture des données de la recherche (open research data) a pour ***objectif*** *la diffusion libre, gratuite et universelle, via internet, des données d'origine publique ou privée. Le terme ouvert est défini comme la liberté d'utiliser, de modifier et de redistribuer librement les données"* [1]. L'***open data*** considère la science comme un bien commun dont la diffusion est d'intérêt public et général. Ce mouvement s'inscrit donc dans l'Open science [2] et l'Open knowledge [3].

Selon Laurence Dedieu et Marie-Françoise Fily, l'ouverture des données de la recherche répond à **cinq enjeux majeurs** [4] :

1. accélérer les découvertes scientifiques, les innovations et le retour sur investissement en recherche et développement ;
2. encourager la collaboration scientifique et les possibilités de recherche interdisciplinaire ;
3. éviter la duplication des expériences, favoriser la réutilisation des données et minimiser le risque de perte des données ;
4. assurer l'intégrité et la reproductibilité de la recherche (meilleure qualité des résultats, transparence des méthodologies);
5. accéder librement à une masse de données ouvrant de nouveaux champs d'analyse non envisagés par le producteur des données (gain de temps et de ressources).

Nous comprenons bien là l'importance de rendre publiques les données de la recherche.

VOUS AVEZ DIS RESEARCH DATA, DATA SET, DATA BASE ? AVANT TOUTE CHOSE, DE QUOI PARLONS-NOUS ?

Afin de mieux appréhender notre propos, arrêtons-nous quelques instants sur quelques définitions... Selon l'OCDE [5], les **données scientifiques** sont ***« des enregistrements factuels (chiffres, recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche. Un ensemble de données de recherche constitue une représentation systématique et partielle du sujet faisant l'objet de la recherche. Ce terme ne s'applique pas aux éléments suivants : carnets de laboratoire, analyses préliminaires et projets de documents scientifiques, programmes de travaux futurs, examens par les pairs, communications personnelles avec des collègues et objets matériels (par exemple, les échantillons de laboratoire, les souches bactériennes et les animaux de laboratoire tels que les souris). »***

Le ***Digital Curation Center*** et l'***Australian National Data Service*** apportent des définitions complémentaires. Pour le premier, une donnée est "une représentation réinterprétable de l'information dans une manière formalisée et adaptée à la communication, l'interprétation, ou le traitement" [6]. Pour le second, "fournir une définition faisant autorité

Les données de la recherche peuvent se regrouper de la façon suivante [8] :

- Les **données d'observation** : il s'agit là de données "capturées" en temps réel, habituellement uniques et donc impossibles à reproduire ;
- Les **données expérimentales** : autrement dit les données obtenues à partir d'équipements de laboratoire, qui sont souvent reproductibles mais parfois coûteuses ;
- Les **données computationnelles ou de simulation** : ce sont les données générées par des modèles informatiques ou de simulation, souvent reproductibles si le modèle est correctement documenté ;
- Les **données dérivées ou compilées** : par définition les données issues du traitement ou de la combinaison de données "brute"», elles sont souvent reproductibles mais coûteuses ;
- Les **données de référence** : elles se présentent sous forme de collections ou d'accumulations de petits jeux de données qui ont été revus par les pairs, annotés et mis à disposition.

des données de recherche est difficile, car toute définition est susceptible de dépendre du contexte dans lequel la question est posée". [7]
Mais tous définissent les données de la recherche comme l'ensemble des informations collectées, observées ou créées sous une forme numérique dans le cadre d'un projet de recherche.

Le **jeu de données**, "dataset", rassemble les données brutes ou dérivées en un ensemble cohérent. Ces informations sont généralement numériques, textuelles, sonores et/ou picturales. L'action de les rassembler permet leur recherche, leur récupération et leur réorganisation. Il peut être également défini comme une collection d'éléments connexes de données associées entre elles et accessibles individuellement ou de façon combinée, ou gérées comme une entité. Les jeux de données numériques sont formatés. Ils sont alors communicables, interprétables et adaptés à un traitement informatisé. Le jeu de données vient étayer les résultats d'une recherche publiés dans une revue. Il sera alors **soit cité, soit déposé, soit cité et déposé**.

Une **base de données** numérique (*database*) est un ensemble structuré et organisé permettant le stockage de grandes quantités d'informations afin d'en faciliter l'exploitation (ajout, mise à jour, recherche de données). D'où l'importance d'accompagner le jeu de données de métadonnées descriptives.

RENDRE PUBLIC UN JEU DE DONNÉES : PUBLIER, DÉPOSER, PUBLIER ET DÉPOSER

En tant qu'organisme de financement de la recherche, l'Union européenne détermine, à ce titre, les règles d'accès et de diffusion de l'information scientifique issue de ses fonds. Le Programme-cadre pour le recherche et le développement technologique (PCRD), appelé Horizon 2020, se caractérise par un *Open Research Data Pilot* des données générées au cours de recherches financées par Horizon 2020. L'objectif étant de promouvoir l'ouverture et la **réutilisation des données** et des métadonnées associées. Le 11 décembre 2020, le budget d'un montant de 95,5 milliards d'euros a été voté par l'Union européenne, lançant ainsi le programme **Horizon Europe** [9] consacré au financement de la recherche et de l'innovation jusqu'en 2027. Ce dernier s'inscrit dans la continuité du programme Horizon 2020 en mettant l'accent sur la science ouverte et l'**accès aux résultats et aux données des recherches.** Il se structure autour de trois piliers : Pilier I, Excellence science, Pilier II, Défis mondiaux et compétitivité industrielle européenne et Pilier III, Europe innovante. Ces piliers s'organisent autour de politiques et stratégies diverses.

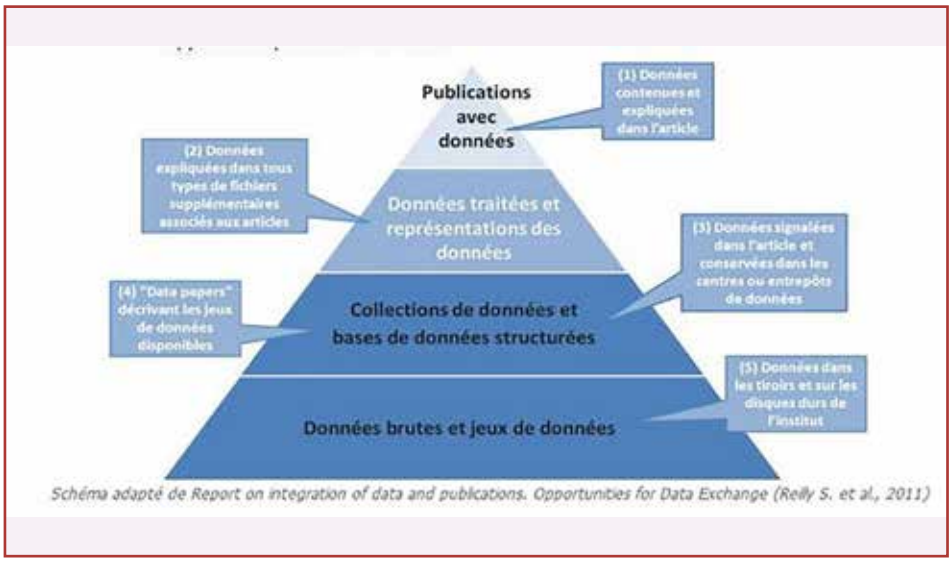
Focus sur le Pilier I d'Horizon Europe : Excellence science

- **Conseil Européen de la Recherche** : première organisation européenne de financement pour la recherche exploratoire.
- **Actions Sklodowska-Curie** : financent des programmes de formation doctorale et postdoctorale et des projets de recherche collaborative.
- **Infrastructures de recherche** : fournissent des ressources et des services aux communautés de recherche pour mener des recherches et favoriser l'innovation dans leurs domaines. Les infrastructures de recherche contribueront également à la réalisation des 4 orientations stratégiques clés du plan stratégique Horizon Europe [10].

Comme nous l'avons évoqué en préambule, notre propos s'articule autour du critère de "**réutilisation**" des principes FAIR, nous ignorons donc délibérément le niveau concernant les données brutes stockées sur des disques durs ou "à l'abri" dans des tiroirs de laboratoire. En effet, elles ne répondent pas aux Principes FAIR et ne sont réutilisables que pour les chercheurs appartenant au laboratoire propriétaire de la recherche et, surtout, elles présentent un risque de perte d'exploitation au vu de l'éventuelle obsolescence du support et de l'outil de lecture de celui-ci. Il nous semble cependant opportun de souligner l'importance, en termes de volume, de ces données. En 2014, l'INIST [11] estime que *"**90% des données de la recherche** sont **stockées sur les disques durs locaux** et potentiellement **non réutilisables** par d'autres"*.

(Source : Publicdomainepictures.net)

Avançons dans notre propos et arrêtons-nous au deuxième niveau, à savoir celui des collections de données et des bases de données structurées.



Pyramide de publication des données

(Source : INIST)

*** Données publiées dans des *Data Papers* :**

Forme récente de publication scientifique centrée sur les données de la recherche, ce type d'article se rencontre sous différentes appellations: *data article, dataset paper, data notes* et sous son intitulé le plus courant, le *data paper.* Le principe appliqué dans ce genre de production scientifique est que le contenu n'a pas pour objectif de mettre en avant un résultat, une analyse ou une conclusion.

Le but du data paper est de fournir une voie formalisée au partage des données plutôt que de tester des hypothèses ou présenter de nouvelles analyses. Ils ont pour but de rendre les données accessibles, interprétables et **réutilisables**.

Pour les auteurs de l'enquête *"Data Journal : A Survey"*, le *data paper* est un *"article qui décrit un set de données et donne des informations sur le quoi, où, pourquoi, comment et qui de la production des données. Celui-ci contient un lien (DOI) vers le repository sur lequel est déposé le set de données et le journal qui publie l'article ne les héberge pas, ce qui garantit que même en cas d'accès restreint à l'article, le set de données reste accessible librement"* [13].

Les *data papers* peuvent être publiés dans des revues qui leur sont dédiées, les *data journals*, sous la forme d'un article examiné par les pairs.[14] Leurs jeux de données sont déposés dans des entrepôts de données, ces derniers leur attribuant un identifiant pérenne (DOI ou PID, par exemple).

Un "*Data Journal*" est un journal (toujours en libre accès) qui publie des articles de données. Il fournit habituellement des modèles (*templates*) de description des données et guide les chercheurs sur les lieux de dépôt et sur la façon de décrire et de présenter leurs données. Les *Data Journals* valorisent la liaison réciproque entre un article de données et le(s) jeu(x) de donnée(s) correspondant(s). Ils garantissent la qualité des données (processus d'examen par les pairs – *Peer review*) et peuvent également fournir des lignes directrices éditoriales pour l'évaluation de la qualité des données [15].

Les avantages de la publication dans un *data paper* sont multiples. Les données, en étant normalisées et standardisées, sont visibles, repérables et citables par d'autres études. Ce mode de publication assure au chercheur une reconnaissance en tant qu'auteur du jeu de données en informant la communauté scientifique de l'existence et la disponibilité de ce même jeu. La méthodologie et les procédures décrites dans le *data paper* assurent la rigueur scientifique de l'étude.

En assurant une visibilité aux données, le *data paper* assure ainsi l'un des critères majeurs du principe FAIR, à savoir la **réutilisation des données pour des études futures** et valorise les données.

Cependant, avant de choisir la revue de publication, il convient au chercheur de vérifier l'éditeur scientifique, ses conditions d'accès (le *data journal* favorise l'*open access*, les revues hybrides non) mais également le montant des éventuels frais de

Recette et ingrédients d'un data paper [12] :

-Les métadonnées de l'article ;
-Le contexte de l'étude, sa conception,les procédures de production et/ou collecte des données, les méthodes d'interprétation et de reproduction ;
-La description des données ;
-Les notes d'usage et instructions pour la **réutilisation** des données.

publication. La visibilité du *data paper* dépendra également des bases de données dans lesquelles sont indexées les revues.

*** Données déposées dans des entrepôts de données reconnus :**

Un entrepôt de données (*Data repository, digital repository*) est un réservoir constitué majoritairement de données de recherche, brutes ou élaborées, qui sont décrites par des métadonnées de façon à pouvoir être retrouvées et **réutilisées**. Le programme cadre "Horizon 2020" exige, sauf restrictions juridiques et éthiques, que les données issues de projets relevant de l'initiative pilote sur le libre accès aux données soient déposées dans un entrepôt accessible. Une fois déposés, les jeux de données se voient attribuer un DOI et, la plupart du temps, une licence de réutilisation. Les jeux de données sont accessibles sur l'entrepôt, éventuellement après une période d'embargo, et peuvent être réutilisés.

Mais pourquoi, me direz-vous, déposer ses jeux de données dans un entrepôt ? Outre le fait que cela soit imposé par le programme "Horizon 2020", l'entrepôt de données permet de conserver les données dans un environnement sécurisé, de rendre les jeux de données plus visibles et citables via un identifiant pérenne. Ce dépôt permet aussi la réutilisation des données et leur reproductibilité par la communauté scientifique, ce dernier critère favorisant la validation scientifique et l'intégrité en recherche.

Il existe différents types d'entrepôts : thématique ou disciplinaire (Réseau Quetelet, par exemple, pour les sciences sociales), multidisciplinaire (Zenodo), institutionnel (Datapartage pour l'INRAe) et spécifique d'un projet de recherche. Il est conseillé de privilégier un entrepôt de confiance, répondant aux critères de qualité exigés pour obtenir une certification [16] (format des données, qualité des métadonnées, conditions d'accès et de réutilisation, identifiant pérenne, archivage à long terme...). Le chercheur, en tant que producteur de données et rédacteur du data paper, peut choisir lui-même l'entrepôt où déposer ses données, tout en tenant compte des recommandations de son institution ou du bailleur de fonds. Mais, si le jeu de données fait l'objet d'une publication, que ce soit dans une revue ou un data paper, le choix de l'entrepôt peut être imposé par la revue de publication, certaines d'entre elles possèdant leur propre entrepôt de données (exemple de GigaDB qui est lié à Oxford Univ. Press).

Passons maintenant au niveau supérieur et voyons ce qu'il en est des données intégrées dans les *Supplementary files* [17] (Matériel supplémentaire) associés à un article. Au vu de l'augmentation significative du nombre de documents supplémentaires dans les revues scientifiques, il est possible de publier les données sous forme de fichiers supplémentaires associés à un article (article de recherche, étude de cas, etc.). Le contenu de ce matériel supplémentaire est varié. Il peut s'agir de fichiers audio, vidéo, images à haute résolution, analyses statistiques, explications méthodologiques approfondies, etc … Ce mode de publication des données apportent plusieurs avantages. Outre le fait que les données en matériel supplémentaire soient libérées des contraintes éditoriales, les *supplementary files* assurent la paternité des données et le crédit aux chercheurs. Cependant, quelques pratiques restent à revoir. En effet, le signalement des fichiers supplémentaires restent encore à standardiser et, même si l'identification des données indépendamment de l'article est possible, via notamment un DOI, elle reste rare. Et surtout, les données restent difficiles à trouver **indépendamment de l'article** et dans **une forme peu ou pas réutilisable**.

*** Données intégrées dans des revues scientifiques :**

Et nous voici au **sommet de notre pyramide** : les données intégrées dans les **revues scientifiques**. Il s'agit du modèle de publication traditionnel dans lequel le chercheur traite, analyse l'ensemble des données et expose les conclusions qu'il en tire. Les données ainsi exposées sont citables, repérables et réutilisables. Cependant, les données sont difficilement repérables en dehors de la publication de l'article.

Aux côtés des données intégrées, certains éditeurs de revues demandent aux chercheurs de lier aux articles les données qui en sont la base. Elles peuvent reposer sur une politique des données [18] et dans ce cas l'auteur devra s'y conformer. Elles peuvent également imposer des entrepôts de données, en suggérer ou laisser libre choix à l'auteur. Pour exemple, là où *Cahiers Agricultures* n'impose pas de politiques de données, *PLOS Neglected Tropical Diseases* impose le choix d'un entrepôt de données public [19].

A l'issue d'un groupe de travail, le Collège Données de la Recherche concluent cependant que si "*ces politiques de données sont en passe de se généraliser pour les revues en sciences, technologies et médecine*", elles "*sont encore peu courantes pour les revues en sciences humaines et sociales"*.

Conclusion

Notre propos s'achève ici. Force est de constater que le partage des données de la recherche n'est pas un long fleuve tranquille et que le chercheur doit prendre en considération plusieurs critères avant de choisir son mode de diffusion et de mise à disposition. Sans oublier que le principe de la science ouverte est de permettre la réutilisation de ces données en les rendant "aussi ouvertes que possible"...

▪ Nadia NEMER-BRUN

Références :

[1] *Dedieu, L., & Fily, M.-F. (2015). Rendre publics ses jeux de données. https://coop-ist.cirad.fr/content/download/5705/42112/version/5/file/CoopIST-rendre-publics-jeux-donnees-avril-2015.pdf*

[2] *https://www.ouvrirlascience.fr/*

[3] *https://okfn.org/*

[4] *Dedieu, L., & Fily, M.-F. (2015). Rendre publics ses jeux de données. https://coop-ist.cirad.fr/content/download/5705/42112/version/5/file/CoopIST-rendre-publics-jeux-donnees-avril-2015.pdf*

[5] *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics. (2007). http://www.oecd.org/fr/science/sci-tech/38500823.pdf*

[6] *"A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing." - Representation Information: what is it and why is it important? | DCC*

[7] *"Providing an authoritative definition of research data is challenging, as any definition is likely to depend on the context in which the question is asked" ands.org.au*

[8] *Gaillard, R. (2014). De l'Open data à l'Open research data : Quelle(s) politique(s) pour les données de recherche ? [ENSSIB]. https://www.enssib.fr/bibliotheque-numerique/notices/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche*

[9] *https://ec.europa.eu/info/research-and-innovation/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe_en*

[10] *https://op.europa.eu/en/web/eu-law-and-publications/publication-detail/-/publication/3c6ffd74-8ac3-11eb-b85c-01aa75ed71a1*

[11] *Une introduction à la gestion et au partage des données de la recherche—Données et publications (2014). https://www.inist.fr/wp-content/uploads/donnees/co/module_Donnees_recherche_20.html*

[12] *Le Deuff, O. (2018). Une nouvelle rubrique pour la RFSIC : Le Data Paper. Revue française des sciences de l'information et de la communication, 15, Article 15. http://journals.openedition.org/rfsic/5275*

[13] *Candela, L., Castelli, D., Manghi, P., & Tani, A. (2015). Data journals : A survey. Journal of the Association for Information Science and Technology, 66(9), 1747-1762. https://doi.org/10.1002/asi.23358*

[14] *Dedieu, L. (2016). Publier un data paper [Vidéo]. https://video.cirad.fr/videos/2016_03_data_paper/Laurence_Dedieu.mp4*

[15] *Key components of data publishing : Using current best practices to develop a reference model for data publishing - RDA/ WDS Working group on Publishing data Workflows Paper. (2015, décembre 4). RDA. https://www.rd-alliance.org/key-components-data-publishing-using-current-best-practices-develop-reference-model-data-publishin-0*

[16] *Application des principes TRUST transparence (Transparency), responsabilité (Responsibility), orientation vers l'utilisateur (User focus), durabilité (Sustainability) et technologie (Technology).*

[17] *aussi appelés : Supplemental material, Supplemental data, Auxiliary information, Supporting information, Supplementary content, Additional content …*

[18] *Une politique de données précise ce que la revue attend de ses auteurs en matière de gestion, d'archivage et de diffusion des données de recherche liées aux publications qu'elles éditent.*

[19] *https://coop-ist.cirad.fr/publier-et-diffuser/publier-dans-une-revue-en-libre-acces/6-l-acces-aux-donnees-de-recherche-associees-aux-articles*

Open Research Europe : La Commission européenne lance sa plateforme d'édition scientifique pour les publications financées par l'Europe

Chloé TRAVERS

Les politiques de libre accès aux publications scientifiques se sont rapidement développées au cours des dix dernières années. Il est primordial, pour que les connaissances puissent circuler librement, de veiller à ce que les chercheurs ou leurs institutions aient le droit de partager sans restriction les résultats de la recherche évalués par les pairs et financée par des fonds publics. L'accès libre et immédiat à des publications financées par des fonds publics a l'avantage de permettre le partage, dans la mesure du possible, des résultats de la recherche, favorisant ainsi l'efficacité de la recherche et l'excellence scientifique. Dans le cadre d'Horizon Europe, la Commission vient de lancer une plateforme de publication appelée Open Research Europe afin d'améliorer la qualité, l'efficacité et la réactivité de la recherche.

QU'EST-CE QUE L'ORE ?

La nouvelle plateforme en *Open Access* a été lancée en janvier 2021 et est dirigée aux bénéficiaires de subventions de l'initiative Horizon Europe dans laquelle au moins un des auteurs est impliqué. Elle présente les **travaux de recherche originaux** dans les domaines de la science, de la technologie, de l'ingénierie et des mathématiques, ainsi que des sciences sociales, des arts et des sciences humaines et propose à tous, citoyens et chercheurs, un **accès gratuit aux dernières découvertes scientifiques**. Les chercheurs peuvent publier pendant ou après la fin d'une subvention, en respectant les politiques de libre accès. La plateforme fournit des lignes directrices détaillées ainsi que des indications sur la manière de publier et de relier les données.

Elle fournit aux bénéficiaires de projets européens une façon simple de publier en *Open Access*, sans coût supplémentaire, et en conformité totale avec la politique *Open Access* de la Commission Européenne. La plateforme emploie un modèle de publication immédiate des articles soumis accompagné d'un examen par les pairs transparent et ouvert. Elle fournit une solution directe aux principales difficultés généralement liées à la publication des résultats scientifiques, et plus particulièrement les retards, les obstacles à la réutilisation des résultats et les coûts élevés. L'**ORE dispose d'un conseil scientifique international solide** qui conseille la Commission sur les questions scientifiques stratégiques et accepte une grande variété de publications : articles de recherche, *reviews*, *case report*… mais également des *Data Notes* (une autre façon de nommer les *data papers*) ou encore des *Software Tool Articles*.

Environ quarante articles scientifiques portant sur des domaines de recherche très diversifiés ont déjà été soumis et peuvent être consultés et examinés par la communauté scientifique. Les premiers articles ont été publiés fin mars 2021.

UNE HISTOIRE DU LIBRE ACCÈS AUX PUBLICATIONS ET AUX DONNÉES

Menée à l'origine par des membres de la communauté scientifique, la question du libre accès s'est peu à peu répandue auprès des instances de décision et institutions. L'***Open Research Europe*** fait suite à l'***Open Research Publishing Platform***.

Début des années 90 - Les premiers pas : des chercheurs lancent leurs premières revues en accès libre.

(Source : Pixabay)

2002 – Un appel capital : lancement de l'Initiative de Budapest pour l'Accès Ouvert. Les signataires demandent le libre accès aux publications, et font remarquer les bienfaits pour les chercheurs eux-mêmes : plus de visibilité, plus d'impact et l'accès à de nouveaux publics.

2003 – Des institutions prennent part au libre accès : la Déclaration de Berlin sur le libre accès à la connaissance en sciences exactes, sciences de la vie, sciences humaines et sociales, signée par des institutions, élargit le concept de libre accès à l'ensemble des résultats de la recherche.

2010 - L'Europe favorise le libre accès avec la mise en place d'**OpenAIRE** (***Open Access Infrastructure for Research in Europe***). Cette structure a pour objectif de participer au développement du libre accès au sein de l'Union Européenne. OpenAIRE va mettre en place **Zenodo** (2013), plateforme de dépôt d'archives en libre accès. En 2015, OpenAIRE est composée de plus de 11 millions de publications.

2013 – Prise de position de la Commission européenne : la Commission européenne, dans l'article 29 du modèle de convention de subvention du programme Horizon 2020, rend obligatoire la transmission des résultats de recherche pour tout bénéficiaire de financement public. Celui-ci doit assurer un accès libre et gratuit à toutes les publications scientifiques évaluées par les pairs et à leurs métadonnées, en les déposant immédiatement dans une archive ouverte.

2016 – Plus d'ouverture dans l'Union européenne : dans un communiqué, la Commission européenne élargit le principe d'accès ouvert aux données de la recherche et à l'intégrité scientifique. Dès 2017, le libre accès aux données est devenu obligatoire pour tous les nouveaux projets financés dans le cadre d'**Horizon 2020**.

2018 – Le principe d'ouverture devient une réalité : dans le cadre d'Horizon 2020, la Commission européenne propose de financer une plateforme d'édition ouverte, *Open Research Publishing Platform,* permettant la publication gratuite des articles des bénéficiaires d'Horizon 2020. En juin 2018, la Commission européenne décrète que le principe de « science ouverte » devient le mode opératoire d'Horizon Europe en exigeant un accès ouvert aux publications et aux données.

Comment juger objectivement un article scientifique ? L'open peer review

La plateforme fonctionne sur la base d'**évaluation ouverte par les pairs** (***Open review***) et d'une **assurance qualité par un conseil consultatif scientifique**. Il s'agit d'un examen transparent par les pairs : les auteurs suggèrent des réviseurs appropriés. Les noms des évaluateurs sont ouverts, ainsi que leurs avis, qui peuvent également être cités. Les auteurs et *reviewers* sont clairement et ouvertement identifiés et leurs échanges sont publics. Les tenants de l'*open peer review* soulignent la responsabilité endossée par les *reviewers*. Les auteurs doivent assumer pleinement leurs critiques. Les vertus du dialogue entre auteurs et *reviewers* permettent de faire émerger peu à peu un consensus.
La publication suit quatre étapes. Premièrement, le chercheur soumet son article qui subit des contrôles de prépublication par l'équipe éditoriale interne afin d'évaluer son éligibilité comme auteur, de vérifier le champ d'application de l'article, de vérifier l'absence de plagiat, de s'assurer que sont respectées les recommandations aux auteurs et les politiques en matière de rapports, d'édition et d'éthique ainsi que la disponibilité des données. Les pairs peuvent également aider les auteurs à rendre leurs données et logiciels FAIR. Puis, dès que l'article a passé les contrôles, la prépublication est publiée sous dix jours, permettant ainsi une consultation et une citation immédiates. Ensuite, il est évalué par des experts qui ont été sélectionnés et invités ; leurs évaluations sont nominatives et publiques, ainsi que les réponses des auteurs et les commentaires des utilisateurs enregistrés. Les auteurs sont incités à publier les versions révisées de leur article. Toutes les versions d'un article sont liées et peuvent être citées indépendamment. Enfin, les articles qui passent l'évaluation par les pairs sont signalés aux principales bases de données bibliographiques et aux archives. Finalement, au niveau de la production, les articles sont mis à disposition dans des formats de texte et d'extraction de données (PDF, HTML, XML).

Les principes FAIR au coeur des publications

À ce titre, la Commission exige des bénéficiaires de tous les financements de la recherche et de l'innovation qu'ils mettent leurs publications à disposition en libre accès et qu'ils rendent leurs données aussi ouvertes que possible et aussi fermées que nécessaire. Au cours des dernières années, plusieurs intervenants, des financeurs de la recherche aux éditeurs, ont donné leur approbation sur un

ensemble concis de principes, connus sous le nom de principes des données FAIR pour améliorer la **réutilisation** des données. Ainsi comme l'exige **Horizon Europe**, l'ORE doit respecter les conditions d'accès ouvert et immédiat. Les données de recherche associées aux publications devront être **disponibles en permanence, publiquement et gratuitement** pour une réutilisation éventuelle et donc être déposées dans un entrepôt et rendues disponibles sous licence CC-BY, CC0 ou équivalent, avant la soumission de l'article. Les articles ainsi que leurs métadonnées (une fois que les articles ont passé l'examen par les pairs) sont archivés de façon permanente dans Portico ou dans Zenodo ou selon leur statut dans une archive thématique comme Europe PMC. Tous les articles reçoivent également un DOI.

La politique de publication de la plateforme précise notamment que **les données de recherche associées aux publications devront être déposées dans un entrepôt** (tel que cités plus haut) et rendues disponibles sous licence CC-BY, CC0 ou équivalent. Le non-respect de cette obligation sans justification entraînera le rejet de l'article. Des exceptions sont permises dans certains cas : protection des résultats, obligation de sécurité ou de confidentialité, protection des données personnelles, données trop lourdes… Par conséquent, un "*Data Availability Statement*" sera exigé pour chaque article, dans lequel les auteurs doivent préciser dans quel entrepôt les données sont accessibles, ou les raisons pour lesquelles les données ne peuvent pas être partagées. Les articles doivent comporter toutes les informations nécessaires pour reproduire, valider et réutiliser les résultats et les analyses sur les données, en particulier les informations sur les logiciels nécessaires pour visualiser et analyser les données.

"Open Research Europe est un grand pas en avant pour les bénéficiaires des programmes de R&I de l'UE et les communautés de recherche de tous les domaines scientifiques, des sciences sociales et des sciences humaines. La nouvelle plateforme de publication leur permettra d'adopter pleinement la science ouverte en répondant à leurs besoins de publication et en partageant, utilisant et trouvant ouvertement des publications et des données liées."

Mariya Gabriel. Commissaire, Commission européenne.

▪ **Chloé TRAVERS**

Bibliographie

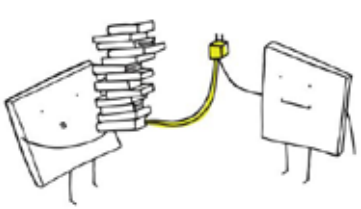
Le libre accès aux publications et aux données de recherche. (s. d.). Horizon 2020. Consulté 19 mai 2021, à l'adresse https://www.horizon2020.gouv.fr/cid82025/le-libre-acces-aux-publications-aux-donnees-recherche.html

Open Research Europe : La Commission européenne lance sa plateforme de publication en libre accès. (s. d.). Consulté 17 mai 2021, à l'adresse https://www.nextinpact.com/lebrief/46561/open-research-europe-commission-europeenne-lance-sa-plateforme-publication-en-libre-acces

Open Research Europe—Observatoire des Sciences de l'Univers—Institut Pythéas. (s. d.). Consulté 18 mai 2021, à l'adresse https://www.osupytheas.fr/?Open-Research-Europe

Pasteur, C.-I. (2021, mai 6). Vous souhaitez publier un article dans Open Research Europe ? N'oubliez pas de partager les données associées. Open science : évolutions, enjeux et pratiques. Consulté 18 mai 2021, à l'adresse https://openscience.pasteur.fr/2021/05/06/vous-souhaitez-publier-un-article-dans-open-research-europe-noubliez-pas-de-partager-les-donnees-associees/

Un historique du libre accès aux publications et aux données. (s. d.). Consulté 19 mai 2021, à l'adresse https://www.ouvrirlascience.fr/un-historique-du-libre-acces-aux-publications-scientifiques-et-aux-donnees



Science ouverte : Les différentes initiatives mises en place en France

Laura HOUCK

Selon le baromètre de la science ouverte, en 2019, 56% de la production scientifique française était en accès ouvert [1], un chiffre en augmentation par rapport à 2018 (49%). Cette progression peut s'expliquer par la mise en place de nouvelles initiatives - et la continuité de celles déjà existantes - visant à démocratiser et encourager les publications ouvertes.

(Source : Pixabay)

À PROPOS DE LA SCIENCE OUVERTE

Revenons tout d'abord sur les principes généraux de la **science ouverte**.

De manière globale, la finalité principale de la **science ouverte** est de généraliser la diffusion en accès ouvert des publications et des **données de la recherche** et de promouvoir la transition numérique. Cela permet d'accéder à une science plus cumulative et à une recherche plus efficace. Ce partage des connaissances scientifiques entraîne ainsi une démocratisation de l'accès aux savoirs ainsi qu'une interdisciplinarité et transversalité de la recherche.
De nouveaux outils et de nouvelles méthodes sont mis en place permettant d'ouvrir les données, les processus, les codes, les protocoles…

LES INITIATIVES GOUVERNEMENTALES (LOIS, APPELS, PLAN DU MESRI - Ministère de l'Enseignement Supérieur, de la Recherche er de l'Innovation)

La loi pour la République numérique - 2016
Cette loi favorise le libre accès aux publications scientifiques, promeut l'ouverture et la circulation des données et l'accompagnement des citoyens dans le passage au numérique. Elle vise à garantir aux internautes la possibilité de disposer d'un environnement numérique ouvert et respectueux de leur vie privée. [2]

Le plan National pour la science ouverte (PNSO) - 2018
Annoncé par Frédéric Vidal le 4 juillet 2018, ce plan national est mis en place dans le cadre de plusieurs partenariats et engagements internationaux pris par la France concernant la science ouverte. Le premier est le partenariat OGP - *Open Government Partnership*, réunissant aujourd'hui 78 pays, dont la France, et dont l'objectif principal est d'encourager les pays membres à rendre leur gouvernement plus ouvert et plus transparent. Le deuxième est le projet appelé "*Amsterdam Call for Action on Open Science*" qui promeut l'accès libre aux publications scientifiques ainsi que le partage et la réutilisation des données pour la Recherche.

"La science est un bien commun que nous devons partager le plus largement possible. Le rôle des pouvoirs publics est de rétablir la fonction initiale de la science, comme facteur d'enrichissement collectif".

Frédérique VIDAL, ministre de l'Enseignement supérieur, de la Recherche et de l'Innovation (MESRI)

Le Plan National pour la science ouverte est une initiative servant à **fixer les modalités pour le développement de la science ouverte** sur le territoire national.
Il est constitué de trois axes comprenant chacun les objectifs et les mesures (9) à mettre en place :

- **Généraliser l'accès ouvert aux publications**. Cela passe par exemple par l'obligation de publier en accès libre (sur des revues en accès ouvert ou sur des archives ouvertes) les publications financées par les fonds publics ;
- **Structurer et ouvrir les données de la recherche**. Nous pouvons ici donner l'exemple du respect des principes FAIR (Faciles à trouver, Accessibles, Interopérables et Réutilisables) ;
- **S'inscrire dans une dynamique durable, européenne et internationale.** Les nouvelles pratiques impliquent un besoin de nouvelles compétences et, nécessairement, de nouvelles formations. Ces besoins s'inscrivent dans une logique nationale et internationale.

« La France s'engage pour que les résultats de la recherche scientifique soient ouverts à tous, chercheurs, entreprises et citoyens, sans entrave, sans délai, sans paiement. »

Frédérique VIDAL, ministre de l'Enseignement supérieur, de la Recherche et de l'Innovation (MESRI)

Le Comité pour la science ouverte (CoSO) - 2018

Dans le cadre du Plan National pour la Science Ouverte, en 2018 a été constitué le Comité pour la Science Ouverte (en remplacement de la BSN - Bibliothèque Scientifique Numérique) un dispositif national avec pour mission principale de **définir une politique de soutien à l'ouverture de la science ouverte, de déterminer la stratégie sur laquelle se baser ainsi que de s'assurer de son (bon) développement dans les établissements et communautés scientifiques** [3]. Cela passe par exemple par un accompagnement efficace des communautés scientifiques ainsi que la sensibilisation des acteurs aux tenants et aboutissants de l'accès ouvert. Les objectifs ont également ici une portée nationale et internationale. Le comité de pilotage se divise en **4 collèges** [4] :

- Publication
- Données de recherche
- Compétences et formation
- Europe et international

L'**Appel à Manifestation d'Intérêt**, ayant pour objectif de recueillir les candidatures des acteurs aspirant à s'impliquer dans la science ouverte et à prendre part intégrante au Comité et à ses objectifs a été publié en mars 2018. Il a également été impulsé par le MESRI. Il est disponible sur *https://www.enseignementsup-recherche.gouv.fr*.

Les appels à projets du Fonds national pour la science ouverte (FNSO) - 2019

Une fois de plus dans le cadre du Plan National pour la Science Ouverte, le Fonds national pour la science ouverte a été créé en 2019. Il sert à **soutenir la science ouverte en mettant en place divers appels à projets** [5]. Pour le premier appel à projet, ce fonds était constitué de 2 675 000€ servant à soutenir les infrastructures, plateformes et projets éditoriaux en faveur de la science ouverte. En termes de processus, les appels à projet sont d'abord publiés, les candidatures reçues, puis le comité de pilotage se charge de sélectionner les projets.

Le Baromètre de la science ouverte - 2019

À l'occasion du Plan national pour la science ouverte, le MESRI a publié en 2019 un baromètre de la science ouverte pour les publications scientifiques. Il permet de **mesurer le taux d'accès ouvert**, c'est-à-dire d'analyser la part des publications en accès libre tout en distinguant les publications ayant été publiées sur des archives ouvertes (HAL par exemple) et la part publiées chez les éditeurs. Il se base sur des **données "fiables, ouvertes et maîtrisées"** [6]. Le site du baromètre est accessible à cette adresse *https://ministeresuprecherche.github.io/bso/.*

LES INITIATIVES MISES EN PLACE PAR LES INSTITUTIONS ET LES ORGANISMES SPÉCIALISÉS

L'APPEL DE JUSSIEU - 2017

L'Appel de Jussieu a surgit du regroupement d'un collectif de chercheurs et professionnels de l'édition scientifique pour **encourager la science ouverte et la bibliodiversité** et "***promouvoir un accès ouvert aux publications scientifiques***" [7] au sein des établissements de recherche et des communautés scientifiques.

Il regroupe 8 apartés indiquant des **méthodes pour répondre aux objectifs de la science ouverte** et contient également les noms des rédacteurs et les institutions signataires.

Il est disponible à l'adresse *https://jussieucall.org.*

L'ANR

En plus d'être signataire de la déclaration précédemment citée, l'ANR - Agence Nationale de la Recherche - a également développé des engagements en faveur de la science ouverte, et ce, depuis 2007. Elle est **signataire de la convention en faveur des archives ouvertes** et de la plateforme HAL (2013) et de la Déclaration de San Francisco (2018), elle **sensibilise à l'importance du libre accès** des publications, elle a rejoint en 2018 la cOAlition S (initiative européenne pour le libre accès aux publications scientifique). Aussi, depuis 2019, l'ANR demande que les **publications issues des projets financés soient déposées dans une archive ouverte**, et demande l'**élaboration d'un plan de gestion des données** respectant les principes FAIR. Elle promeut l'accès ouvert aux publications scientifiques.

L'ANR met en place ces initiatives dans le cadre du Plan National pour la science ouverte et fixe 3 objectifs:

- Le développement de l'***Open Access***
- La contribution à l'***Open Research Data***
- La **concertation** au niveau national, européen et international.

Le flyer de présentation des engagements est disponible sur *https://anr.fr/.*

(Source : Pixabay)

Les initiatives du CNRS

Le CNRS est un acteur national et européen majeur dans la production des données de la recherche. Les initiatives mises en place par cet organisme sont donc significatives pour le développement de l'accès ouvert aux publications scientifiques.

La feuille de route science ouverte - 2019

La feuille de route du CNRS vise à "accélérer le processus vers la science ouverte en s'appuyant sur des actions concrètes". Elle a été publiée après la fixation par l'Europe et le MESRI des objectifs en matière de science ouverte. Elle se base sur 4 objectifs ayant comme finalité d'atteindre une **ouverture à 100% des publications du CNRS**, de développer une culture de partage de données chez les acteurs concernés, de **promouvoir les principes FAIR**, de **promouvoir le développement d'outils et infrastructures** pour l'analyse des contenus ou encore de **transformer les modes d'évaluation** des chercheurs et chercheuses pour les rendre conciliables avec les principes de la sciences ouvertes [8]. Ces objectifs sont expliqués et détaillés sous forme d'actions, en 7 axes.

Pour mettre en place cette feuille de route, le CNRS a mobilisé 3 unités de services:

- L'**Inist** (Institut de l'information scientifique et technique)
- Le **CCSD** (Centre pour la communication scientifique directe)
- **Persée**

Ces dernières sont étroitement liées avec le développement de la science ouverte depuis des années dans la mesure où elles promeuvent l'accès à l'information scientifique et aux archives ouvertes, la valorisation du patrimoine scientifique et des données de la recherche, ou encore la transition numérique.

La feuille de route proposée par le CNRS est disponible à cette adresse *www.science-ouverte.cnrs.fr*.

Le plan "Données de la recherche" du CNRS - 2020

Il s'agit ici d'un plan du CNRS dont la finalité principale est d'**accélérer le développement vers la science ouverte**. Il vise à développer une stratégie pour l'ouverture des données tout en prenant en compte les principes FAIR. Ce plan s'inscrit dans le cadre de l'initiative européenne **EOSC** - ***European Open Science Cloud,*** lancée en 2018 par la commission européenne pour ***"rassembler des acteurs institutionnels, nationaux et européens, des initiatives et des infrastructures de données afin de développer un écosystème de science ouverte inclusif en Europe"*** [9] et ce, sous la forme d'un *cloud* réservé aux pratiques de science ouverte.

L'idée est donc de promouvoir un "EOSC français", et c'est pourquoi le plan d'actions du Plan "Données de la recherche" s'adapte aux **besoins des chercheurs** tout en tenant compte des **différentes disciplines,** en accompagnant les acteurs dans l'**évolution des pratiques** (nouveaux outils, nouvelles pratiques, nouveaux services…) et en prenant en compte les **besoins en ressources humaines** et les formations nécessaires pour mener à bien le projet.
Le Plan est disponible sur la page internet du CNRS : *www.science-ouverte.cnrs.fr.*

L'ANSES

L'Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail s'est récemment engagée en faveur de la science ouverte en signant, aux côtés de ses partenaires "***une déclaration conjointe en faveur de la science ouverte par un réseau d'agences françaises de financement de la recherche***". L'objectif de cette déclaration est d'**établir une approche concertée** permettant la **diffusion des publications et des données de la recherche**, promouvoir une **généralisation des accès** aux savoirs, une **harmonisation** des pratiques au sein des communautés de chercheurs et chercheuses.
Les différents signataires sont : l'Agence nationale de la recherche (ANR), l'Agence de la transition écologique (ADEME), l'Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (ANSES), l'Agence nationale de recherches sur le sida et les hépatites virales (Inserm / ANRS), l'Institut national du cancer (INCa) [10].
Découvrez la déclaration sur le site internet de ANSES *www.anses.fr/.*
En dehors de cette déclaration, l'Agence met également en place des actions concrètes dans le cadre de son Programme National de Recherche Environnement-Santé-Travail [11]. En effet, l'Agence a invité les auteurs des publications bénéficiant de son financement à les **déposer dans une archive ouverte.**

Les universités

Les universités prennent également des initiatives en faveur de la science ouverte. Ci-dessous nous avons répertorié, de manière non exhaustive certaines des universités engagées dans le développement de l'accès ouvert

- L'**Université de Lorraine** qui, depuis 2016, prend des engagements en faveur de la science ouverte: **colloques** thématiques pour accompagner et promouvoir l'ouverture des recherches, le désabonnement de certains journaux d'**éditeurs commerciaux**, ratification de l'Appel de Jussieu, identification d'un chargé de mission pour la science ouverte et d'un **comité de pilotage** pour la science ouverte [12]. L'Université de Lorraine a également mis en place le "**Baromètre lorrain de la science ouverte**" reprenant le code du Baromètre français de la Science Ouverte, afin de l'adapter aux publications de l'Université de Lorraine et ainsi pouvoir **analyser la progression de la Science Ouverte** dans l'établissement et générer des graphiques pertinents. Le code de ce baromètre a été partagé afin de permettre à d'autres établissements de le décliner [13].
- L'**Université de Strasbourg**, en partenariat avec d'autres acteurs de la région, a mis en place, en 2016, la **plateforme d'archive ouverte univOAK** [14]. La plateforme accueille les publications scientifiques de chercheurs de disciplines diverses et les diffuse en libre accès. Découvrez la plateforme à cette adresse : *https://univoak.eu.*

L'Université de Strasbourg a également repris le code ci-dessus mentionné pour établir son propre baromètre de science ouverte.

"L'université fait le pari d'offrir ses données au public, de les diffuser, mais pas sans discernement".

Michel DRENKEN, président de l'Université de Strasbourg

- Les initiatives en faveur de la science ouverte sont également présentes à l'**Université Paris Saclay** qui a également mis en place son baromètre de la science ouverte. L'Université possède une collection spécifique sur la plateforme HAL et promeut également les autres plateformes de diffusion de publications en libre accès [15].
- L'**Université de Montpellier**, quant à elle, a ouvert son portail HAL [16], est signataire de la "Charte de signature des publications scientifiques des structures de recherche de l'Université de Montpellier", de la "Charte relative à l'intégrité scientifique" [17] et a mis en œuvre le protocole de Nagoya. De plus, L'Université P. Valéry Montpellier 3 soutient le projet **NumeRev**, un incubateur de revues *open access*, portail scientifique interdisciplinaire et laboratoire expérimental [18].

Nous pouvons ainsi conclure cet article en affirmant que ces dernières années ont été propices à la création d'initiatives dans le domaine de la science ouverte. Elles proviennent d'acteurs divers et présentent des objectifs communs : la promotion du libre accès et le développement des publications ouvertes afin de faciliter la transition numérique et de généraliser l'accès aux savoirs. La science ouverte constitue une avancée considérable, non seulement pour les auteurs mais également pour les internautes, en permettant un accès libre, ouvert et partagé à l'information.

▪ **Laura HOUCK**

(Source : Unplash - Susan Q Yin)

Références :

[1] Baromètre français de la Science Ouverte 2020. (s. d.). Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. Consulté 20 mars 2021, à l'adresse //www.enseignementsup-recherche.gouv.fr/cid156502/barometre-francais-de-la-science-ouverte-2020.html

[2] Pour une République numérique. (s. d.). Gouvernement.fr. Consulté 21 mars 2021, à l'adresse https://www.gouvernement.fr/action/pour-une-republique-numerique

[3] Appel à Manifestation d'Intérêt (AMI) pour la constitution du Comité pour la Science Ouverte (CoSO). (s. d.). Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. Consulté 21 mars 2021, à l'adresse www.enseignementsup-recherche.gouv.fr/cid128239/appel-a-manifestation-d-interet-ami-pour-la-constitution-du-omite-pour-la-science-ouverte-coso.html

4] Numerev. (s. d.). Consulté 21 mars 2021, à l'adresse https://numerev.com/presentation

Ouvrir la Science—Le comité pour la science ouverte. (s. d.). Consulté 21 mars 2021, à l'adresse https://www.ouvrirlascience.fr/presentation-du-comite/

[5] Appel à projets N°2 « Publications » – Fonds national pour la science ouverte. (s. d.). Consulté 21 mars 2021, à l'adresse https://www.ouvrirlascience.fr/appel-a-projets-n2-publications-fonds-national-pour-la-science-ouverte

[6] l'Innovation, M. de l'Enseignement supérieur, de la Recherche et de. (s. d.). Baromètre français de la science ouverte. Baromètre français de la science ouverte. Consulté 21 mars 2021, à l'adresse https://bso.enseignementsup-recherche.gouv.fr

[7] Plaquette_Science-Ouverte_18112019.pdf. (s. d.). Consulté 21 mars 2021, à l'adresse https://www.science-ouverte.cnrs.fr/wp-content/uploads/2019/11/Plaquette_Science-Ouverte_18112019.pdf

[8] European Open Science Cloud (EOSC). (s. d.). [Text]. European Commission - European Commission. Consulté 21 mars 2021, à l'adresse https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/european-open-science-cloud-eosc_en

[9] Appel de Jussieu. (s. d.). Consulté 21 mars 2021, à l'adresse https://jussieucall.org/

[10] L'Anses s'engage en faveur de la science ouverte | Anses—Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail. (s. d.). Consulté 21 mars 2021, à l'adresse https://www.anses.fr/fr/content/l%E2%80%99anses-s%E2%80%99engage-en-faveur-de-la-science-ouverte

[12] Les engagements de l'Université de Lorraine en faveur de l'ouverture de la science. (2020, mai 4). Factuel | le site d'actu de l'Université de Lorraine. https://factuel.univ-lorraine.fr/node/14275

[13] Baromètre lorrain de la Science Ouverte. (s. d.). Science Ouverte. Consulté 21 mars 2021, à l'adresse http://scienceouverte.univ-lorraine.fr/barometre-lorrain-de-la-science-ouverte/

[14] Science ouverte—Université de Strasbourg—Accueil. (s. d.). Consulté 21 mars 2021, à l'adresse https://scienceouverte.unistra.fr/

[15] Science ouverte. (2020, janvier 7). Université Paris-Saclay. https://www.universite-paris-saclay.fr/recherche/science-ouverte

[16] Science ouverte. Les Bibliothèques Universitaires. Université de Montpellier. Consulté 12 juillet 2021 à l'adresse https://bibliotheques.edu.umontpellier.fr/science-ouverte/

[17] Charte relative à l'intégrité scientifique de l'Université de Montpellier. (s. d.). Université de Montpellier. Consulté 21 mars 2021, à l'adresse https://www.umontpellier.fr/recherche/charte-relative-a-lintegrite-scientifique

[18] Présentation. NumeRev. Consulté 21 mars 2021, à l'adresse https://numerev.com/presentation

Amsterdam Call for Action on Open Science. (s. d.). Consulté 21 mars 2021, à l'adresse https://www.ouvrirlascience.fr/amsterdam-call-for-action-on-open-science

Le mouvement pour la science ouverte. (s. d.). Science Ouverte. Consulté 21 mars 2021, à l'adresse https://www.science-ouverte.cnrs.fr/le-mouvement-pour-la-science-ouverte/

Le Plan national pour la science ouverte : Les résultats de la recherche scientifique ouverts à tous, sans entrave, sans délai, sans paiement. (s. d.). Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. Consulté 21 mars 2021, à l'adresse //www.enseignementsup-recherche.gouv.fr/cid132529/le-plan-national-pour-la-science-ouverte-les-resultats-de-la-recherche-scientifique-ouverts-a-tous-sans-entrave-sans-delai-sans-paiement.html

Ouvrir la Science—La science ouverte. (s. d.). Consulté 21 mars 2021, à l'adresse https://www.ouvrirlascience.fr/category/science_ouverte/

DOSSIER : SCIENCE OUVERTE ET BONNES PRATIQUES

Comment concilier open data et anonymisation des données ?

Selin GUDER

Le terme d'Open data (ou données ouvertes en français) désigne, comme son nom l'indique, des données auxquelles l'accès est totalement public et libre de droit, au même titre que l'exploitation et la réutilisation. Si l'open data ne concerne pas initialement la protection des données à caractère personnel, le nouveau contexte numérique implique de la prendre en considération, au niveau de la mise à disposition et de la réutilisation des données ainsi que la protection de la vie privée. Le nouveau cadre juridique relatif à l'open data permet cette conciliation. Entre ces deux termes, il n'y a pas lieu de choisir, car la libre réutilisation des données ouvertes, tout comme le droit à la vie privée, constituent des principes d'égale valeur qu'il convient de concilier, et non de sacrifier l'un à l'autre. L'enjeu est ici de trouver un équilibre entre ces deux principes qui sont, en théorie, opposés.

PRINCIPES DE BASE DE LA PROTECTION DES DONNÉES

Qu'est-ce qu'une donnée à caractère personnel ?

La **CNIL** la définit comme étant une information se rapportant à une personne physique identifiée ou identifiable. Une personne physique peut être identifiée directement par son nom et prénom ou indirectement par une information qui la concerne telle que son numéro de sécurité sociale, son numéro de téléphone, son adresse postale ou courriel mais aussi par sa voix ou une image. L'identification peut se réaliser à partir d'une seule donnée ou alors par un croisement de plusieurs données.

Le **traitement de ces données**, c'est-à-dire l'ensemble d'opérations portant sur des données personnelles,, quel que soit le procédé utilisé (collecte, enregistrement, organisation, conservation, adaptation, modification, extraction, consultation, utilisation, communication par transmission diffusion ou toute autre forme de mise à disposition, rapprochement), doit être justifié par un objectif légal et légitime.

La protection des données personnelles repose sur **5 grands principes** :

1 **Le principe de finalité** : le responsable d'un fichier ne peut enregistrer et utiliser des informations sur des personnes physiques que dans un but bien précis, légal et légitime;

2 **Le principe de proportionnalité et de pertinence** : les informations enregistrées doivent être pertinentes et strictement nécessaires au regard de la finalité du fichier;

3 **Le principe d'une durée de conservation limitée** : il n'est pas possible de conserver des informations sur des personnes physiques dans un fichier pour une durée indéfinie. Une durée de conservation précise doit être fixée, en fonction du type d'information enregistrée et de la finalité du fichier ;

4 **Le principe de sécurité et de confidentialité** : le responsable du fichier doit garantir la sécurité et la confidentialité des informations qu'il détient. Il doit en particulier veiller à ce que seules les personnes autorisées aient accès à ces informations ;

5 **Les droits des personnes**.

Que dit la Loi ?

Le **RGPD** - Règlement Général sur la Protection des Données - est le texte législatif encadrant le traitement des données personnelles sur le territoire de l'Union européenne. Celui-ci s'inscrit dans la continuité de la **Loi française Informatique et Libertés de 1978** et de la **loi CADA** (Commission d'Accès aux Documents Administratifs).

Ce cadre juridique garantit la protection des **données personnelles** dans le contexte de l'***Open Data*** en France et plus largement sur le territoire européen. D'une part, il tient à l'interdiction de principe qu'une donnée personnelle fasse l'objet d'une mise en ligne par l'administration et d'une **réutilisation** par un tiers. Ce principe connaît trois exceptions : le **consentement** de l'intéressé à cette diffusion, l'**existence** d'une obligation légale de publication, ou l'**anonymisation des données publiées**.

D'autre part, il contribue à la soumission de toute réutilisation de données personnelles aux exigences de la **loi "Informatique et libertés"** et à la sanction lors de non-respect des dispositions précédentes : engagement de la responsabilité de l'État, voire **condamnation pénale pour diffusion de données personnelles par négligence**.

Les données personnelles sont donc des données dites "**sensibles**" et ont l'obligation d'être **protégées**.

Pourquoi protéger les données de la recherche ?

Les projets de recherche intégrant des données à caractère personnel et particulièrement ceux en SHS (Sciences Humaines et Sociales) sont quotidiens et il est important de les protéger.

Au-delà des contraintes juridiques et du cadre législatif posé par le RGPD, les données de recherche ont le mérite d'être protégées pour des raisons éthiques. D'une part, dans le but de **protéger l'identité** des personnes dans les publications et données de recherche contenant des informations sensibles, illégales, confidentielles mais aussi pour **dissimuler le lieu de la recherche**.

Pour finir, les chercheurs souhaitant collecter ou accéder à ces données doivent s'engager à ne pas en faire un usage commercial ou à le redistribuer. Il est important de les sensibiliser sur l'importance de leur collecte mais surtout de leur **sécurisation de stockage** sachant que les données personnelles sont considérées comme « le pétrole du XXème siècle » pour les entreprises du web.

LES TECHNIQUES À METTRE EN OEUVRE POUR LA PROTECTION DES DONNÉES

Comment anonymiser les données ?

L'**anonymisation** est un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible toute identification de la personne par quelque moyen que ce soit et de **manière irréversible** et permet, ainsi, de respecter sa vie privée. Il s'agit d'une solution, parmi d'autres, pour pouvoir exploiter des données

personnelles dans le respect des droits et libertés des personnes. En effet, l'anonymisation ouvre des potentiels de réutilisation des données initialement interdites du fait du caractère personnel des données exploitées, et permet ainsi aux acteurs d'exploiter et de partager leur « gisement » de données sans porter atteinte à la vie privée des personnes. Avec cette méthode, la législation relative à la protection des données ne s'applique plus, car la diffusion ou la **réutilisation des données anonymisées** n'a pas d'impact sur la vie privée des personnes concernées.

Pour construire un processus d'anonymisation pertinent, il est ainsi conseillé :

- d'**identifier les informations** à conserver selon leur pertinence ;
- de **supprimer les éléments d'identification directe** ainsi que les valeurs rares qui pourraient permettre une ré-identification aisée des personnes (par exemple, la présence de l'âge des individus peut permettre de ré-identifier très facilement les personnes centenaires) ;
- **de distinguer les informations importantes** des informations secondaires ou inutiles (c'est-à-dire supprimables) ;
- **de définir la finesse idéale** et acceptable pour chaque information conservée.

Une fois cette étape réalisée, un procédé d'anonymisation doit être défini et appliqué. Il faut donc choisir entre deux techniques:

- La randomisation, qui consiste à modifier les attributs dans un jeu de données de telle sorte qu'elles soient moins précises, tout en conservant la répartition globale. Un jeu de données, ou *dataset*-est un ensemble cohérent de données produites dans le cadre d'un même projet et/ou recueillies sur un même lieu. Toutes les données d'un dataset peuvent donc être décrites avec une majorité de métadonnées communes. Cette technique permet de protéger le jeu de données du risque d'inférence - via une théorie déductive. L'objectif de cette méthode est de réduire le biais de confusion c'est-à-dire un ensemble d'erreurs pouvant survenir dans l'interprétation des liens entre les variables dépendantes et indépendantes, de réduire les biais de sélection c'est-à-dire des erreurs systématiques faites lors de la sélection des sujets à étudier. Cette technique permet, par ailleurs, de respecter les contraintes éthiques et d'interpréter correctement les tests d'hypothèse. Le principal inconvénient de la randomisation est le risque de déséquilibre des traitements pour les facteurs.

- La généralisation consiste à modifier l'échelle des attributs des jeux de données, ou leur ordre de grandeur, afin de s'assurer qu'ils soient communs à un ensemble de personnes. Cette technique permet d'éviter l'individualisation d'un jeu de données. Elle limite également les possibles corrélations du jeu de données avec d'autres.

Quelles pratiques pour quels types de données à anonymiser ?

Anonymisation des données quantitatives

- **Supprimer les identifiants directs** ou les remplacer par des pseudonymes, par exemple les noms, l'adresse, l'institution, la photo ;
- Réduire la précision/le détail par l'agrégation : par exemple l'année de naissance par rapport à la date de naissance, les catégories professionnelles, la région plutôt que le village ;
- Restreindre les fourchettes supérieures et inférieures pour cacher les valeurs.

Anonymisation des données qualitatives

- **Éviter de supprimer les informations ;** utiliser des pseudonymes ou des remplacements
- **Identifier les remplacements**, par exemple avec des [crochets] ;
- **Éviter de trop anonymiser** car la suppression des informations dans le texte peut déformer les données, les rendre inutilisables, peu fiables ou trompeuses : il faut donc trouver un équilibre entre l'anonymisation et la nécessité de préserver le contexte ;
- **Conservez un journal** d'anonymisation des remplacements ou des suppressions effectués.

Anonymisation des données audiovisuelles

La **manipulation numérique** de fichiers audio et d'images peut supprimer des identifiants personnels, par exemple l'altération de la voix ou le flou des images (par exemple des visages). Ce type de données est plus difficile à anonymiser car cela demande un travail intensif avec un coût élevé et risque de nuire au potentiel de recherche des données.

Que faire si l'anonymisation est impossible ?

Différentes pistes s'offrent aux chercheurs. La plus simple reste l'obtention du consentement pour le partage de données non anonymisées auprès des personnes. A défaut de cette solution, il est possible de restreindre l'accès des utilisateurs, réservé aux chercheurs agréés, par exemple, le ***UK Data Archive*** met à disposition des données archivées qui ne sont pas diffusées dans le domaine

public ou encore, de réglementer l'**utilisation de ces données** : les utilisateurs de données signent une licence d'utilisateur final juridiquement contraignante, par exemple, ne pas identifier toute personne identifiée ou identifiable. Les chercheurs doivent dans tous les cas envisager l'**accès aux données** et leur **stockage en toute sécurité**.

(Source : donnees-rgpd.fr)

(Source : legibase.fr)

Les critères et risques liés à l'anonymisation

Pour vérifier l'efficacité de l'anonymisation des données, il faut s'assurer que le jeu de données réponde aux trois critères du RGPD :

- **l'individualisation** : il ne doit pas être possible d'isoler un individu dans le jeu de données. Par exemple, une base de données de CV où seuls les nom et prénoms d'une personne auront été remplacés par un numéro (qui ne correspond qu'à elle) permet d'individualiser cette personne. Dans ce cas, cette base de données est considérée comme **pseudonymisée** et non comme **anonymisée** ;
- **la corrélation** : il ne doit pas être possible de relier entre eux des ensembles de données distincts concernant un même individu. Par exemple : une base de données cartographique renseignant les adresses de domiciles de particuliers ne peut être considérée comme anonyme si d'autres bases de données, existantes par ailleurs, contiennent ces mêmes adresses avec d'autres données permettant d'identifier les individus ;
- **l'inférence** : il ne doit pas être possible de déduire, de façon quasi-certaine, de nouvelles informations sur un individu.

À défaut de remplir parfaitement ces trois critères, le responsable de traitement qui souhaite anonymiser un jeu de données doit démontrer que le risque de ré-identification avec des moyens raisonnables est nul.
Les techniques d'anonymisation et de ré-identification étant amenées à évoluer régulièrement, il est indispensable pour tout responsable de traitement concerné, d'effectuer une veille régulière pour préserver, dans le temps, le caractère anonyme des données produites. Cette veille doit prendre en compte les moyens techniques disponibles ainsi que les autres sources de données qui peuvent permettre de lever l'anonymat des informations.

Si un jeu de données publié en ligne comme « anonyme » contient en réalité des données personnelles et qu'aucune des exceptions mentionnées à l'**article L.312-1-2 du Code** des relations entre le public et l'administration **(CRPA)** n'est applicable, cela peut être considéré comme une violation de données. Il est alors nécessaire de :
- **procéder au retrait du jeu de données** en question dans les plus brefs délais ;
- en informer la **CNIL** et les personnes concernées si cette violation est susceptible d'engendrer un risque pour les droits et libertés des personnes.

Pour conclure, l'anonymisation est donc une technique fondamentale favorisant la réutilisation des données initialement interdite du fait du caractère personnel des données exploitées, et permet ainsi aux chercheurs d'exploiter et de partager leur "gisement" de données sans porter atteinte à la vie privée des personnes. Elle permet également de conserver des données au-delà de leur durée de conservation. Le RGPD ne s'applique pas aux données anonymisées dans la mesure où l'utilisation de ces données n'a pas d'impact sur les droits et libertés des personnes concernées.

En revanche, lorsque l'anonymisation n'est pas possible notamment lorsque les objectifs de la recherche nécessitent de mentionner l'identité de l'interviewé (personnalité, expert…), il convient de leur préciser que des données identifiantes seront publiées et de leur garantir l'accès à la retranscription.

Enfin, une dernière possibilité s'offre au chercheur quant à la diffusion des données et la publication : la transmission de données non-anonymisées à d'autres chercheurs possible sous autorisation du responsable de traitement en vertu du décret du 1er août 2018 (article 100-1).

▪ **Selin GUDER**

(Source : europe.groupebgfibank)

Bibliographie

Calimaq. (2018, juillet 18). Données personnelles et recherche scientifique : Quelle articulation dans le RGPD ? - S.I.Lex -. https://scinfolex.com/2018/07/18/donnees-personnelles-et-recherche-scientifique-quelle-articulation-dans-le-rgpd/

Coulibaly, I. (s. d.). La protection des données à caractère personnel dans le domaine de la recherche scientifique. 1123.

*L'anonymisation des données, un traitement clé pour l'open data | CNIL. (s. d.). Consulté 21 mars 2021, à l'adresse https://www.**cnil.fr/**fr/lanonymisation-des-donnees-un-traitement-cle-pour-lopen-data*

Les collectivités territoriales et l'open data : Concilier ouverture des données et protection des données personnelles | CNIL. (s. d.). Consulté 21 mars 2021, à l'adresse https://www.cnil.fr/fr/les-collectivites-territoriales-et-lopen-data-concilier-ouverture-des-donnees-et-protection-des

Open data : La protection des données comme vecteur de confiance | CNIL. (s. d.). Consulté 21 mars 2021, à l'adresse https://www.cnil.fr/fr/open-data-la-protection-des-donnees-comme-vecteur-de-confiance

Practical data anonymization.pdf. (s. d.). Consulté 21 mars 2021, à l'adresse https://www.fosteropenscience.eu/sites/default/files/original/47544.pdf

Quel statut pour les données de la recherche après la loi numérique ? – – S.I.Lex –. (s. d.). Consulté 21 mars 2021, à l'adresse https://scinfolex.com/2016/11/03/quel-statut-pour-les-donnees-de-la-recherche-apres-la-loi-numerique/

Quels sont les grands principes des règles de protection des données personnelles ? | Besoin d'aide | CNIL. (s. d.). Consulté 21 mars 2021, à l'adresse https://www.cnil.fr/fr/cnil-direct/question/quels-sont-les-grands-principes-des-regles-de-protection-des-donnees

RESEARCH DATA MANAGEMENT AND OPEN DATA.pdf. (s. d.). Consulté 21 mars 2021, à l'adresse https://www.fosteropenscience.eu/sites/default/files/pdf/1895.pdf

RGPD / open data : Comment concilier les deux ? (s. d.). Smart City Mag. Consulté 21 mars 2021, à l'adresse http://www.smartcitymag.fr/article/306/rgpd-open-data-comment-concilier-les-deux

RGPD : De quoi parle-t-on ? | CNIL. (s. d.). Consulté 21 mars 2021, à l'adresse https://www.cnil.fr/fr/rgpd-de-quoi-parle-t-on

RGPD et open data : Est-il possible de concilier les deux ? (2019, mars 12). Données & RGPD. https://donnees-rgpd.fr/traitement-donnees/concilier-rgpd-open-data/

Ouvrir et partager des données de recherche selon les principes FAIR : comment rédiger un Plan de gestion des données de recherche (PGD ou DMP) ?

Flora CHONG

Dans un contexte favorable à l'Open Data, il y a de plus en plus de mandats pour rendre accessibles les données liées aux publications en particulier de l'édition scientifique, des organismes de financement et des politiques nationale et institutionnelle. Même après leur publication, 80% des données scientifiques sont perdues pour les auteurs, les organismes, les institutions et pour la recherche mondiale. Pour éviter ces pertes, comment les chercheurs devraient-ils publier leurs données ? Et comment publier les données de manière fiable (selon les principes FAIR) ? Le plan de gestion de données est très lié au principe du libre accès aux données de recherche et il est devenu l'outil de gestion incontournable des projets de recherche pour les États et les financeurs.

De manière générale, les **principes FAIR** [1] concernent l'ouverture, la **communication**, l'**appropriation** et la **réutilisation** des données de recherche. Ainsi, une bonne gestion des données est un moyen de soutenir les principes FAIR. En pratique, le plan de gestion de données est un instrument de la "**FAIRification"** [2] de la recherche permettant la découverte des données, leur accessibilité, leur interopérabilité et leur réutilisation.

Mais d'abord, un PGD, qu'est-ce que c'est ?

« Le plan de gestion des données est un outil de gestion. Il se présente sous la forme d'un document structuré en rubriques. Il a pour objectif de synthétiser la description et l'évolution des jeux de données d'un projet de recherche. Il prépare le partage, la réutilisation et la pérennisation des données. » [3] (Doranum).

La gestion des données, en vue de leur partage et de leur réutilisation éventuelle, est un processus qui demande **planification** et **organisation**. Les chercheurs doivent prévoir et allouer du temps pour la gestion des données dès le début de leur projet de recherche. Le ***Data Management Plan*** (DMP) – ou **Plan de gestion des données** – aide à organiser la gestion des données (création, collecte, documentation, description, partage et préservation) tout en abordant les questions juridiques en lien avec leur utilisation ou réutilisation (restriction légale, propriété des données, propriété intellectuelle, obligations contractuelles, données sensibles). Le DMP est un document évolutif qui doit être complété et mis à jour de façon régulière et qui peut prendre différentes formes (document électronique, modèle en ligne, formulaire papier, etc.) et peut amplement varier selon les disciplines et projets de recherche [4]. Le plan de gestion de données s'appuie sur le cycle de vie des données qui désigne les différentes étapes de traitement des données au cours d'un projet de recherche.

En résumé, le DMP c'est :

1 Document évolutif (3 versions minimum)	**2** Aide pour bien organiser les données	**3** Description des données selon leur cycle de vie
4 Définition des responsabilités	**5** Aide pour évaluer les ressources nécessaires	**6** Aide pour obtenir des données fiables

Rédiger un DMP, une approche réglementée ?

Rédiger un DMP est primordial pour plusieurs raisons. Il permet un gain de temps et l'anticipation de plusieurs problématiques : coûts, destruction ou perte des données, infrastructure, etc. Il est parfois obligatoire et exigé par certains bailleurs de fonds publics pour l'octroi de financements. Il favorise considérablement la réutilisation des données, met en valeur les données et travaux de recherche, et enfin soutient une recherche intègre, responsable et transparente. Il est à noter que le DMP est obligatoire pour obtenir un financement du **FNS** [5]. Il est également obligatoire pour obtenir un espace de stockage sécurisé [6] mis en place par la Division calcul et soutien à la recherche [7]. Afin de favoriser la diffusion ouverte des données de recherche, l'**ANR** [8] attire l'attention des déposants sur l'importance de considérer la question des données de recherche au moment du montage et tout au long du projet. Elle impose un DMP pour tous les projets qu'elle finance (Plan d'action ANR 2019, p.9).

Rédiger un DMP est utile pour se poser les bonnes questions dès le départ d'un projet, quitte à évoluer au fur et à mesure de l'avancement du projet. Il consent par exemple à identifier les risques liés à la gestion des données, assurer la sécurité et la préservation des données, prévoir les budgets, matériels, logiciels, personnels, etc., identifier les responsabilités, les rôles de chacun dans la gestion des données, planifier les ressources et compétences nécessaires à cette gestion, garantir

des **données fiables et bien gérées**, compréhensibles, disponibles et préservées sur le long terme pour une réutilisation future (démarche FAIR) [9] ou encore répondre aux exigences d'un financeur.

Pour les financeurs, l'intérêt est la **réutilisabilité des données** (retour sur investissement, ne pas dupliquer inutilement l'effort financier). Pour les organismes de recherche, c'est la **reproductibilité de la recherche** qui prime avant toute chose. Pour les chercheurs, il est fondamental de procéder à une bonne gestion des données au cours d'un projet, et ce pour diminuer les risques, réduire les coûts, augmenter l'efficacité avec la valorisation du travail et permettre les demandes de financement.

Le DMP est un phénomène mondial incontournable qui est de plus en plus recommandé ou exigé, partout dans le monde.
La soumission des DMP dépend des :

- **exigences** de la Commission européenne (Modèles Horizon 2020, ERC)
- **déploiements** d'outils et infrastructures d'ampleur européenne en lien avec la gestion et le partage des données de la recherche (l'entrepôt Zenodo, l'infrastructure OpenAIRE …)

À l'échelle nationale, l'État français a instauré une politique avec le Plan national pour la science. L'ANR a rendu le DMP obligatoire depuis 2019. Au niveau des organismes, il a été mis en place des trames de DMP institutionnelles (CIRAD, INRA, Institut Pasteur, Irstea, Universités…), des politiques d'établissements (INRA…) ainsi que des recommandations intentionnelles (intégrées dans les DMP OPIDoR [10]). L'objectif principal est de "garantir des données fiables et bien gérées tout au long d'un projet, compréhensibles, disponibles et préservées sur le long terme pour une réutilisation future" [11]
D'après une conférence donnée à Amsterdam en 2016 sur la science ouverte, « la gestion et le partage des données doivent devenir l'approche par défaut pour les recherches financées par le secteur public. » (Amsterdam Call for Action on Open Science, 2016).

Comment rédiger un DMP dans les règles de l'art ?

Sa rédaction commence dès le début du projet ! D'ailleurs, il peut être demandé dès la soumission du projet. D'après Inist-CNRS [12], pour bien rédiger un DMP, il faut :
- prévoir **3 versions au minimum** (3 versions successives demandées pour H2020 par exemple) : au début du projet, au milieu du projet et à la fin du projet ;
- **désigner nominativement** les personnes responsables de la gestion des données pour toutes les étapes du projet : saisie des données, production des métadonnées, contrôle de la qualité des données, stockage, partage et archivage des données ;
- **évaluer les ressources nécessaires** (budget, temps alloué, personnels) permettant la mise en œuvre des actions décrites dans le DMP : temps nécessaire à la préparation des données pour le stockage, le partage et l'archivage des données, coûts de matériel, rémunération des personnels, frais de stockage (serveurs dédiés, traitement, maintenance, sécurité, accès…), partage (site web, publication…) et d'archivage des données.

Son contenu informationnel peut varier en fonction du modèle de plan, qu'il soit imposé par un tiers ou choisi. Selon Science Europe [13], les rubriques doivent préciser les aspects suivants [14] :
Contexte : La nature et le contexte du projet de recherche
Description : Le type de données de recherche collectées et produites
Documentation et qualité : Les formats, métadonnées et standards utilisés :

- *Quelles métadonnées et quelle documentation (méthodologie de collecte et mode d'organisation des données) accompagneront les données ?*
- *Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?*

Stockage et sauvegarde : Les informations sur le stockage, la sauvegarde et la sécurisation des données
Exigences légales et éthiques : Les questions éthiques, juridiques et déontologiques (code de conduite) qui se posent :

- *Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?*
- *Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ?*

Partage et conservation à long terme : L'accès, le partage, la réutilisation des données, ainsi que l'archivage et le dépôt utilisé

- *Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (un entrepôt de données ou une archive) ?*
- *Comment l'application d'un identifiant unique et pérenne (DOI) sera réalisée pour chaque jeu de données ?*

Dans le contexte des appels à projet,

il peut être demandé d'expliciter plus spécifiquement comment les principes FAIR sont pris en compte et appliqués.

Un focus sur les modèles de DMP du FNS et H2020…

Exigences du modèle FNS

Depuis 2017, il est **obligatoire d'inclure un DMP** avec chaque requête. Le FNS met à disposition un modèle de DMP dans le compte de chaque chercheur sur MyFNS [15]. En contrepartie, le FNS demande aux chercheurs qu'il finance d'archiver les données de recherche sur lesquelles ils ont travaillé et qu'ils ont produites durant leurs travaux, de partager ces données avec d'autres chercheurs et enfin de déposer leurs données dans des archives (dépôts) publiques existantes, dans des formats accessibles et réutilisables sans restriction par tous, et répondant aux **principes FAIR**.

Le FNS considère le partage des données de recherche « ***comme une contribution fondamentale à l'impact, à la transparence et à la reproductibilité de la recherche scientifique. Les bénéficiaires de subventions doivent donc s'assurer que les données générées par leur projet soient accessibles au public dans des dépôts de données non commerciaux et respectant les exigences FAIR.*** »

Pour aider à la rédaction d'un DMP pour le FNS, le centre de compétences suisse en sciences sociales a élaboré un guide intitulé : ***How to draft a DMP from the perspective of the social sciences, using the SNSF template - june 2019*** [16].

Exigences du modèle H2020

Depuis 2017, tout projet de recherche financé par Horizon 2020 fait partie par défaut de l'***Open Research Data Pilot*** (ORDP) [17]. Ce projet a pour

(Source : Pixabay)

but d'ouvrir les données de recherche tout en protégeant les données sensibles d'accès inappropriés. La rédaction d'un DMP est obligatoire et les chercheurs doivent décrire dans les grandes lignes leur politique de gestion des données en répondant aux questions suivantes :

- Comment les données seront-elles gérées, pendant et après le projet ?
- Quelles données seront collectées, traitées ou générées ?
- Quels méthodologies et standards seront appliqués ?
- Les données seront-elles partagées/rendues accessibles et comment ?
- Comment les données seront-elles archivées, conservées et préservées ?

GUIDES POUR LA RÉDACTION DE DMP DANS HORIZON 2020 : MODÈLES DE DMP EN FRANÇAIS ET EN ANGLAIS (DOCX ET PDF)

GUIDELINES ON DATA MANAGEMENT IN HORIZON 2020. VERSION 3.0 (UNION EUROPÉENNE, 26.07.2016)

RÉALISER UN PLAN DE GESTION DES DONNÉES "FAIR" : GUIDE DE RÉDACTION [V2, 2018] (A.CARTIER, R.DELEMONTEZ, M.MOYSAN, N.REYMONET, 2018)

Quelles recommandations et quels outils à disposition pour rédiger un DMP ?

Quel que soit l'organisme de financement d'un DMP, il existe des recommandations générales pour concevoir un bon DMP. Sur le site de Doranum, il est notamment conseillé de consulter des exemples de DMP, de suivre les conseils des sites de référence et d'opter pour un modèle (nombreux modèles des financeurs et/ou organismes). Cette pratique oblige à respecter les usages de sa communauté. Par ailleurs, il est vivement recommandé de **s'autoévaluer avant l'évaluation externe**, de **partager ses données** (avec des collaborateurs identifiés aux

droits définis copropriétaire/éditeur/lecteur, toutes personnes de son organisme ou avec tout le monde) et de **publier son DMP**. Pour cette dernière étape, vous pouvez utiliser un outil online comme…

DMP-OPIDOR

(France, INIST-CNRS, en français, recommandé par l'ANR et par l'IRD)

DMP ONLINE

(UK, DCC-Digital Curation Centre, en anglais)

DMP-TOOL

(US-University of California, en anglais)

Plusieurs outils sont disponibles pour aider à rédiger le DMP, avec des propositions de trames prédéfinies. Il peut s'agir soit de simples fichiers Word ou Excel contenant les différents champs souhaités, ou bien d'applications en ligne ouvertes à la communauté de recherche. En France, l'outil le plus répandu est Opidor : développé par le CNRS, il propose plusieurs trames de plans de gestion de données ou de logiciels, et offre une possibilité de personnalisation aux établissements de recherche français. Certains établissements de l'Université Paris-Saclay (CEA, INRA) proposent des plans de gestion de données accessibles via les pages de vos services de documentation.

Conclusion

Établir un DMP et partager ses données est une action qui permet d'accroître l'efficience de la recherche tout en facilitant l'accès et l'analyse. Il est essentiel d'assurer la continuité et la reproductibilité de la recherche pour protéger l'intégrité de la recherche elle-même. Cela réduit notablement le risque de perte de données et le gaspillage des ressources. Cette politique accompagne l'évolution actuelle de la publication scientifique, répond aux conditions de financement des projets et atteste la responsabilité scientifique.

La Commission européenne a généralisé le DMP pour tous les projets du programme Horizon 2020. Il en est de même pour les projets ANR (Agence Nationale de la Recherche) depuis 2019 [3]. Le DMP est un élément clé pour produire des données FAIR dans le cadre de la gouvernance des données. Il s'agit d'un document qui décrit comment sont ou seront obtenues, traitées, organisées, stockées, sécurisées, préservées, et partagées les données produites au cours et à l'issue d'un projet de recherche. Ce document synthétique aide à organiser et anticiper la mise en place de bonnes pratiques de gestion à toutes les étapes du cycle de vie des données. Il explique pour chaque jeu de données comment seront gérées les données d'un projet, depuis leur création ou collecte jusqu'à leur partage et leur archivage.

Des exemples de plans de gestions de données publics sont consultables à partir de la page d'accueil de ces outils (Rubriques DMP Publics/Public DMPs et Aide), ou publiés dans des journaux comme Research Ideas and Outcomes (RIO).

▪ **Flora CHONG**

POUR EN SAVOIR PLUS :

- Une présentation des PGD [11]

- Un mode d'emploi de DMP-Opidor
Féret, R., Bracco, L., Cheviron, S., Lehoux, E., Arènes, C., & Li, L. (2020, Avril). Améliorer les chances de succès de son projet ANR grâce à la Science Ouverte. Zenodo. https://dx.doi.org/10.5281/zenodo.3741666

- un jeu pour apprendre à gérer ses données sans douleur : GopenDoRe les pages DMP de l'INRA (Datapartage), du Cirad (CoopIST) et de DoRANum (Données de la Recherche : Apprentissage NUMérique à la gestion et au partage)

Références

[1] *Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18*

[2] *FAIRification Process. (s. d.). GO FAIR. Consulté 10 mars 2021, à l'adresse https://www.go-fair.org/fair-principles/fairification-process/*

[3] *Plan de gestion des données : Fiche synthétique – DoRANum. (s. d.). https://doranum.fr/plan-gestion-donnees-dmp/fiche-synthetique/*

[4] *Data Management Plan (DMP). (s. d.). Consulté 10 mars 2021, à l'adresse https://www.unil.ch/openscience/fr/home/menuinst/open-research-data/gerer-les-donnees-de-recherche-research-data-management/data-management-plan-dmp.html*

[5] *Fonds national suisse de la recherche scientifique.*

[6] *Hébergement de données hors recherche. (s. d.). Consulté 10 mars 2021, à l'adresse https://www.unil.ch/ci/fr/home/menuinst/catalogue-de-services/stockage-et-serveur/hebergement-de-donnees-hors-recherche.html*

[7] *Division calcul et soutien à la recherche—DCSR. (s. d.). Consulté 10 mars 2021, à l'adresse https://www.unil.ch/ci/fr/home/menuinst/calcul--soutien-recherche.html*

[8] *Agence Nationale pour la Recherche*

[9] *Service, U. D. (s. d.). Research data management. UK Data Service. Consulté 10 mars 2021, à l'adresse https://ukdataservice.ac.uk/learning-hub/research-data-management/*

[10] *Modèles de DMP, 10 mars 2021, DMP OPIDoR, site web: https://dmp.opidor.fr/*

[11] *IST IRD - Service IST de l'Institut de Recherche pour le Développement. (14:56:35 UTC). Gerer ses données avec un Plan de Gestion de Donnees (PGD/DMP). 20/09/18 https://fr.slideshare.net/IST_IRD/grer-ses-donnes-avec-un-plan-de-gestion-de-donnes-pgddmp*

[12] *Tutoriel sur l'outil de rédaction DMP OPIDoR – DoRANum. (s. d.). https://doranum.fr/tutoriel-sur%20loutil-de-redaction-dmp-opidor/*

[13] *https://www.scienceeurope.org/*

[14] *Guide pratique pour une harmonisation internationale de la gestion des données de recherche. (s. d.). https://www.ouvrirlascience.fr/guide-pratique-pour-une-harmonisation-internationale-de-la-gestion-des-donnees-de-recherche*

[15] *MySNF. (s. d.). Consulté 10 mars 2021, à l'adresse https://www.mysnf.ch/login.aspx?language=fr*

[16] *Data Management Plan – content of the mySNF form, 10 mars 2021, FN-SNF: https://www.unil.ch/openscience/files/live/sites/openscience/files/Donnees_de_recherche/Files/DMP_content _mySNF_form_en.pdf*

[17] *What is the EC Open Research Data Pilot?, 10 mars 2021, Openaire, site web: https://www.openaire.eu/what-is-the-open-research-data-pilot*

Comprendre l'importance des identifiants persistants dans la construction de données faciles à trouver, accessibles, interopérables et réutilisables.

Julien COLIN

L'emploi d'identifiants uniques et pérennisables dans le temps représente un enjeu fort pour le développement de la science ouverte et des principes FAIR [1]. Aujourd'hui, une très grande quantité de productions scientifiques sont disponibles ou signalées sur le web. Faciliter la recherche et le partage d'information en exploitant des identifiants qualifiés apparaît rapidement comme une nécessité. Mais pour garantir la disponibilité libre et indépendante sur le long terme des informations scientifiques, les formats et la gouvernance de ces identifiants sont également à questionner.

ÉTIQUETTER LE MONDE

1-Desidentifiantsglobalement uniques :

Un identifiant pérenne est une chaîne de caractères alphanumériques. En général ce code n'est pas signifiant en lui-même. Il est souvent abrégé par l'acronyme PID pour *Persistant IDentifier* en anglais. Son rôle est de garantir l'identification d'une ressource ou d'une personne quelle que soit sa nature. Les objets peuvent être des publications, des jeux de données brutes ou traitées, des images, des sons, des concepts dans un thésaurus... Les entités référencées, pour leur part, concernent aussi bien des personnes physiques que des institutions ou des entreprises : des chercheurs, des universités, des organismes financeurs, ou encore des personnages historiques.

Le rôle des identifiants est à la fois de désambiguïser les termes et de gérer les homonymies. Par exemple, dans un univers numérique, il est important de pouvoir différencier les notions attachées au mot « opéra ». Le bâtiment et l'œuvre scénique auront chacun un code différent pour les repérer. De la même manière, Jacques Martin, l'illustrateur de bande dessinée (ISNI 0000 0001 2033 0240) n'aura pas la même suite chiffrée que l'animateur télé (ISNI 0000 0000 7359 228X) pour assurer la distinction. Il s'agit de réaliser une association unique entre un code et une entité de sens. Pour être complètement performante, l'association devra être globalement unique. Son unicité devra se vérifier dans un environnement le plus large possible, être détachée de tout système et partagée avec la communauté utilisatrice.

L'objectif premier des identifiants pérennes est de faciliter la recherche d'information, et leur importance est soulignée dès la première exigence du premier principe FAIR : F1- Attribuer des identifiants uniques et pérennes aux données.

Ce type d'identifiant existe déjà depuis plusieurs années pour des objets physiques comme le livre avec l'ISBN (*International Standard Book Number*) introduit en 1970 pour qualifier de manière unique chaque édition de chaque livre publié. Dans le monde de l'édition, l'ISSN (*International Standard Serial Number*), dont la norme ISO a été créée en 1975 pour identifier les publications en série, est un autre exemple de PID.

2- Localiser sur le long terme :

Avec le développement du numérique et plus particulierement avec l'essor du web, l'identification univoque et durable des données, au sens large du terme, s'est adaptée pour répondre aux spécificités de ce nouvel environnement. Ainsi, Les identifiants pérennes ont une syntaxe commune qui est basée sur la spécification des URI (*Uniform Ressource Identifier*) du W3C [2]. Cette syntaxe est composée de trois parties :

• le protocole dans lequel l'identifiant est attribué (par ex. http:, DOI:, ARK:, etc.)

• un préfixe qui permet de désigner l'autorité nommante qui a attribué l'identifiant au sein de ce système. Les préfixes ont besoin d'être unique au niveau mondial et sont gérés par des institutions internationales. Par exemple l'IANA [3] (*Internet Assigned Numbers Authority*) a attribué le prefixe 12148 pour identifier la BNF (Bibliothèque National de France) dans le système ARK (*Archival Resource Key*)

• un suffixe, c'est-à-dire une chaîne de caractères qui identifie la ressource de manière unique, au sein de ce système et pour cette autorité.

Par exemple : doi:10.1045/january2005-fox

Avec internet, il est également intéressant de pouvoir localiser une ressource et d'avoir un texte actionnable pour pouvoir y accéder. Ainsi, l'identifiant est souvent associé à une adresse web. Avec la volubilité du numérique, la question de l'accès sur le long terme vient s'ajouter à la problématique de la localisation. En effet, le système d'adressage par URL (*Uniform Resource Locator*) lie le document à son emplacement physique sur un serveur connecté à internet. Si ce document est déplacé ou le serveur modifié, il devient impossible de retrouver le document. Ce problème est souvent matérialisé par l'erreur 404, code renvoyé par un serveur pour indiquer qu'aucune ressource n'a été trouvée à l'adresse demandée. Ainsi pour garantir la disponibilité d'une ressource, identifiée de façon univoque, les identifiants pérennes se basent sur une gestion active des liens au travers d'un résolveur. L'objectif est de faire correspondre au nom de la ressource son adresse réelle de manière la plus durable possible. Le résolveur peut être interne à l'institution qui donne les noms, ou externe et géré par une autorité indépendante.

Pour rendre actionnable un identifiant, il sera précédé de l'adresse d'un résolveur. Son rôle est de rediriger l'internaute vers la ressource, quel que soit son emplacement sur le web. Si on prend le résolveur d'identifiant de type HANDLE de la Bibliothèque du Congrès à Washington et la ressource de l'exemple précédent on peut construire l'adresse suivante, alors que la ressource n'a aucun lien avec l'institution du résolveur.Par exemple : http://hdl.loc.gov/doi:10.1045/january2005-fox

Ainsi, les identifiants pérennes permettent à la fois de rendre facile le fait de trouver une ressource et de garantir son accessibilité. Ils participent donc aux principes FAIR.

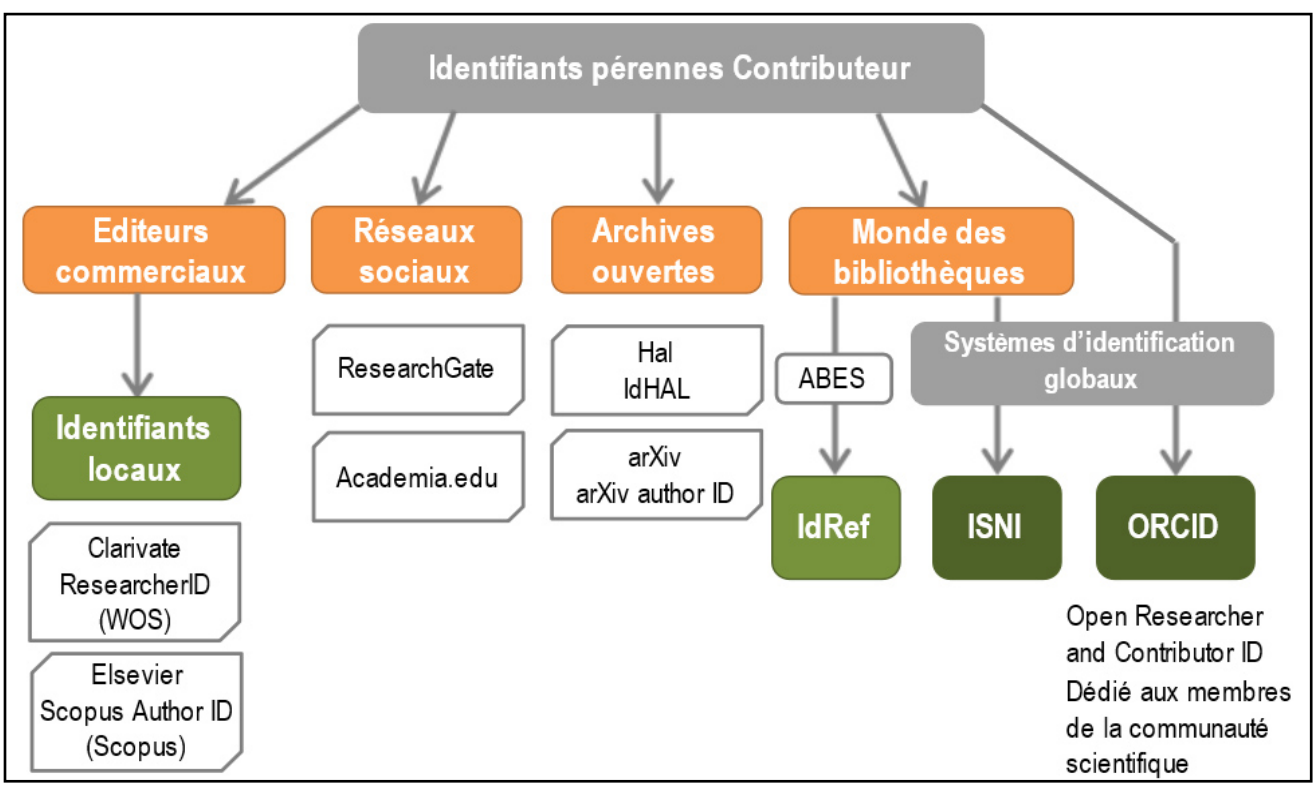
3- Une prolifération des identifiants

Le système PID s'est largement développé. Aujourd'hui de nombreuses plateformes et organismes attribuent des identifiants aux données hébergées et à leurs auteurs. Concernant les contributeurs, il existe une multitudes d'identifiants locaux fournis par les éditeurs commerciaux comme par exemple Clarivate avec l'identifiant ResearcherID ou Elsevier avec Scopus Author ID. Les sites de réseaux sociaux tels que ResearchGate ou Academia.edu délivrent aussi leur propre identifiant. Les plateformes d'archives ouvertes peuvent proposer la création d'un identifiant local, par exemple idHAL pour l'archive ouverte HAL, ou *arXiv author ID* pour l'archive ouverte Arxiv.

A cela s'ajoutent des identifiants globaux comme l'ISNI (*International Standard Name Identifier*) défini par une norme ISO ou ORCID (*Open Researcher and Contributor ID*) dédié spécifiquement aux membres de la communauté scientifique.

Les objets, publications ou jeux de données, ont aussi leur propre système comme HANDLE, DOI, ARK ...

Si le système des identifiants pérennes répond aux principes FAIR, leur prolifération tend à limiter les avantages. Pour répondre aux objectifs de la science ouverte et garantir l'accessibilité sur le long terme des informations scientifiques, il est nécessaire de s'assurer que les identifiants s'appuient sur des protocoles standards, libres, ouverts et soutenus par la communauté scientifique. Il est donc important de prendre en considération l'ensemble du système pour s'assurer de répondre, à tous les niveaux, aux différentes prérogatives énoncées par les principes FAIR.



(DoRANum. Données de la recherche : apprentissage numérique. [en ligne]. France : DoRANum 2020. Identifiants pérennes : un aperçu. 12 décembre 2020 (consulté le 17/03/2021) Disponible sur :https://doranum.fr/identifiants-perennes-pid-apercu/)

Fédérer les initiatives pour relier les données

1- Un engagement politique

Un plan national pour la science ouverte a été initié en France suite au discours du 4 juillet 2018, de Frédérique Vidal, Ministre de l'Enseignement supérieur, de la Recherche et de l'Innovation. Ce plan divisé en 3 grands axes vise entre autre à favoriser le développement des principes FAIR. À ce titre, la France s'est notamment engagée dans diverses entreprises pour s'assurer que les identifiants pérennes répondent à ces principes. L'objectif général est de soutenir l'utilisation de standards et de protocoles ouverts et partagés par la communauté scientifique (cf. R4, A1 et A2 des principes FAIR),

Ainsi, le comité pour la science ouverte participe au consortium ORCID, regroupant de nombreux acteurs scientifiques au niveau international pour promouvoir et co-gérer le développement de ce système. Cet identifiant est ainsi indépendant et gratuit pour les chercheurs. En plus d'identifier de manière unique n'importe quel auteur scientifique qui en fait la demande, l'ORCID associe des métadonnées décrivant le profil du chercheur. Présenté sous la forme d'une page web (ORCID record), ce CV permet au chercheur de gérer lui-même ses informations et de les lier à ces différents autres identifiants. Ce numéro est interconnecté avec la plupart des plateformes comme Web Of Science (WoS), Zenodo, figshare. Il est reconnu, voire exigé, par de nombreux acteurs de la publication scientifique et permet ainsi d'associer les travaux à leur auteur : soumission d'article, demande de financement, évaluation, publication...

Pour les objets, l'Inist-CNRS (Institut National de l'Information Scientifique et Technique du Centre National de la Recherche Scientifique) s'est associé à l'organisme DataCite pour devenir l'agence d'attribution des DOI pour la France. DataCite est lui-même lié à l'IDF (*International DOI Foundation*) qui est une organisation à but non lucratif. L'appui d'un consortium comme DataCite permet de maintenir un résolveur commun à un grand nombre d'institutions. L'attribution d'un DOI demande la mise en place d'un contrat via le portail OPIDor [4]. L'identifiant est gratuit pour les chercheurs car c'est l'organisme ou l'institution de recherche qui souscrit un abonnement. Le DOI fournit un accès stable et durable à la ressource quelle que soit son emplacement sur le web. Il est aussi associé à un fichier de données dont le schéma est libre et construit par la communauté. Le schéma est consultable en ligne. [5] L'ensemble répond à la norme ISO 26324:2012 garantissant sa solidité et sa pérennité.

L'usage d'identifiants pérennes est un pilier du web des données liées. Le choix du système doit en plus répondre aux principes FAIR pour garantir l'accessibilité et l'interopérabilité nécessaire pour la science ouverte.

2- Organiser les correspondances

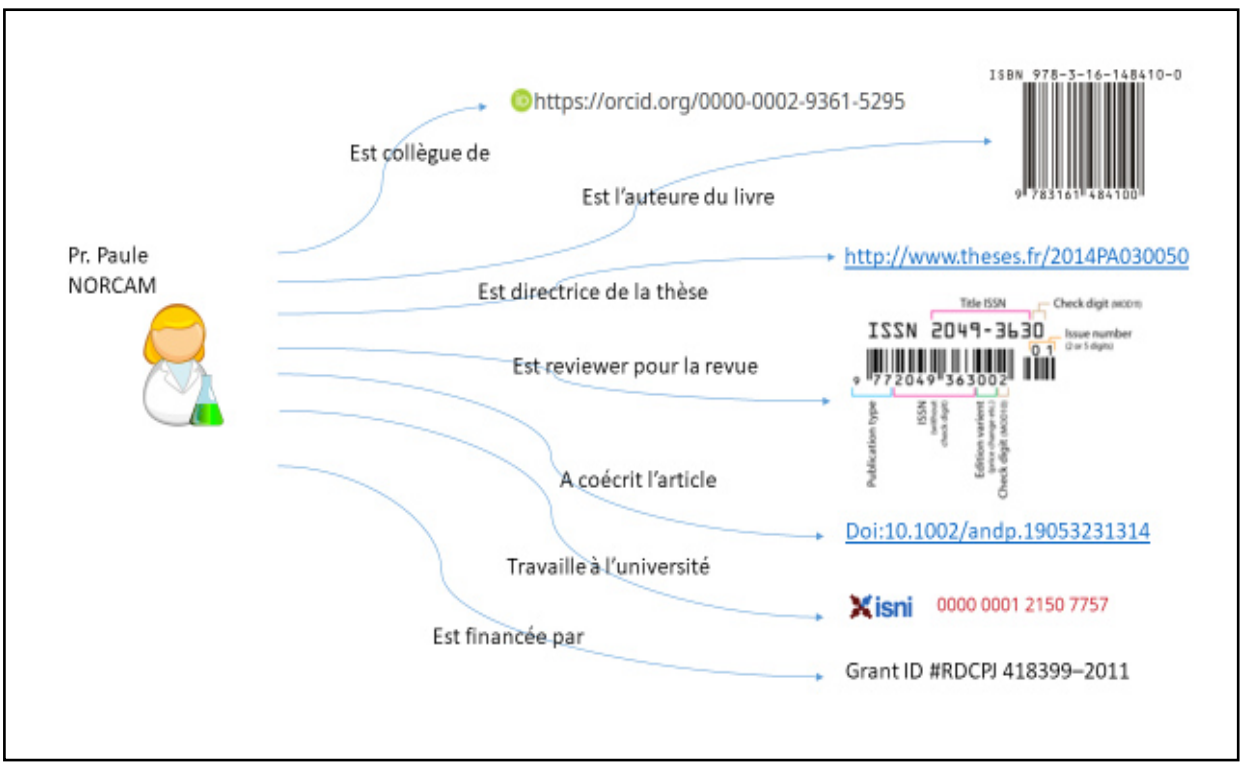
Un identifiant unique et durable dans un système libre et ouvert ne suffit pas forcément. En effet, aujourd'hui de nombreux systèmes coexistent, rendant l'idée d'unicité globale quelque peu caduque. Cette multiplicité s'explique par des objectifs et des besoins différents. Si certains identifiants semblent proches, en réalité ils se

distinguent notamment par les métadonnées associées. Les différents schémas permettent ainsi d'avoir plus au moins de finesse selon les besoins. C'est le cas, par exemple, entre les identifiants du RNSR (Répertoire National des Structures de Recherche) et ceux d'AuréHAL (Accès Unifié aux Référentiels HAL). De même, ORCID est exclusivement destiné à l'identification des auteurs et contributeurs des domaines de l'enseignement supérieur et de la recherche alors que l'ISNI représente l'identité publique d'une personne de façon plus générale dans le domaine de la création intellectuelle. De son côté, l'ABES (Agence Bibliographique de l'Enseignement Supérieur) a développé son identifiant l'IDRef permettant par exemple de qualifier tous les auteurs de thèse au nom du dépôt légal. Mais l'ORCID comme IDref sont étroitement liés à l'ISNI. L'ABES ou la BNF assure alors un rôle d'alignement au niveau international pour tenter de faire correspondre tous ces identifiant au travers de projet comme le VIAF (*Virtual International Authority File* : Fichier d'autorité international virtuel) ou des plateformes comme Wikidata. L'ORCID propose aussi de lien d'autres identifiants pour favoriser l'interopérabilité.

3- Relier toutes les données

L'utilisation généralisée des PID de toutes sortes permet donc d'identifier et de localiser un grand nombre de données sur le web. L'usage de ces identifiants dans les métadonnées permet alors de relier toutes ces informations dans un vaste réseau. En répondant à l'exigence d'emploi de métadonnées avec des attributs selon les principes FAIR, il est alors possible de relier un auteur à ses travaux ou à son institution, de faire correspondre les données et la publication qui les exploitent. Les métadonnées liées favorisent la réutilisation des informations tout en garantissant la paternité et la provenance.

Les identifiants pérennes, en participant à un système global, augmentent la visibilité sur les web aussi bien des chercheurs et des institutions que des travaux scientifiques. L'identification d'une ressource, en étant unique, induit sa répétition à chaque utilisation, renforçant sa présence. De même, les identifiants liés facilitent la recherche d'information que ce soit par le principe même d'une identification univoque mais aussi par les rebonds permis. D'une publication, je peux retrouver l'ensemble des travaux d'un chercheur, ou l'ensemble des domaines d'une institution, ou encore l'ensemble des articles liés à une revue...



(Comité pour la science ouverte [En ligne]. France : Comité pour la science ouverte : 2020 Des identifiants ouverts pour la science ouverte : note d'orientation (consulté le 17/03/2021) Disponible sur :https://www.ouvrirlascience.fr/des-identifiants-ouverts-note-dorietation)

Conclusion

Les principes FAIR sont un ensemble de recommandations pour gérer les données de la recherche visant à les rendre faciles à trouver, accessibles, interopérables et réutilisables aussi bien par l'homme que par la machine. Le principe d'identifiant persistant est la pierre angulaire du déploiement de la science ouverte. Aujourd'hui, et surtout demain, des millions d'objets scientifiques (publications, données et autres objets numériques), produits par autant de chercheurs, affiliés à leur tour à des centaines de milliers d'organisations, peuvent être reliés grâce à des systèmes d'identifiants pérennes ouverts et alignés. Ce vaste réseau d'information favorise la diffusion, la découverte et le partage des données scientifiques, éléments fondamentaux de l'innovation.

▪ Julien COLIN

À LIRE :

- **Sur le site de DORAnum :**

- les ressources consacrées aux principes FAIR :
https://doranum.fr/enjeux-benefices/principes-fair/
- les ressources consacrées aux identifiants pérennes :
https://doranum.fr/identifiants-perennes-pid/

- **Les différentes ressources proposées sur le site du comité pour la science ouverte :**
- *https://www.ouvrirlascience.fr/category/ressources/*

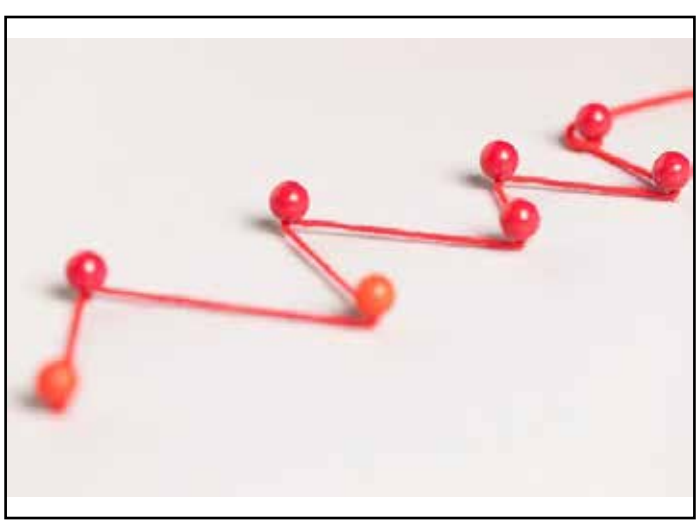
- **Le site de l'ABES consacré au réseaux IDref/ORCID :**
https://abes.fr/reseaux-idref-orcid/le-reseau/

- **Le site de la BNF présentant les identifiants internationaux :**
https://www.bnf.fr/fr/identifiants-internationaux

- **Vade-mecum Identifiants pérennes pour les ressources culturelles**
https://www.culture.gouv.fr/Sites-thematiques/Innovation-numerique/Donnees-publiques/ Identifiants-perennes-pour-les-ressources-numeriques

- **Projet de PID Graph du consortium Freya :**
https://www.project-freya.eu/en/pid-graph/the-pid-graph

- **Le site wikidata :**
https://www.wikidata.org



(Source : Freepic)

Références :

[1] *Énoncé des différents principes FAIR : https://www.go-fair.org/fair-principles/*
[2] *RFC 3986 : http://www.ietf.org/rfc/rfc3986.txt*
[3] *Internet Assigned Numbers Authority. La liste des préfixes enregistrés est accessible en ligne : http://www.iana.org/assignments/uri-schemes.html*
[4] *Portail mis en place par l'Inist-CNRS : https://opidor.fr/identifier/*
[5] *Portail mis en place par l'Inist-CNRS : https://opidor.fr/identifier/*
[6] *Schéma de métadonnées DataCite : https://schema.datacite.org/*

Dossier : stru
et groupes
de t

Les Plateformes Universitaires de données

Cécile DESCAMPS

PROGEDO accompagne les chercheurs dans la gestion des données d'enquêtes et de statistiques en Sciences Humaines et Sociales. La Très Grande Infrastructure de Recherche (TGIR) PROGEDO a pour objectif de développer la culture des données de la recherche mais également d'organiser et d'encourager une stratégie des données recueillies lors des enquêtes réalisées dans le cadre de la recherche en sciences sociales. Les Plateformes Universitaires de Données (PUD) ont été créées dans les Maisons des Sciences de l'Homme (MSH), elles-mêmes situées au sein des universités françaises, afin de répondre à cette volonté politique. Elles apportent leur support local aux acteurs de la recherche qui utilisent les données quantitatives en Sciences Humaines et Sociales (SHS) tout en respectant l'utilisation des données confidentielles et en proposant un environnement de travail sécurisé.

Science ouverte et TGIR

La création de différents types d'infrastructures de recherche scientifique s'est développée avec la science ouverte afin de favoriser la gestion des données tout en respectant les principes du modèle FAIR (Facile à trouver, Accessible, Interopérable et Réutilisable). Celles-ci sont encadrées par des politiques nationales et européennes afin d'articuler une stratégie autour des résultats de la recherche entre industriels, innovation et projets de recherche. En 2008, suite à la Feuille de route nationale [1] élaborée par le Ministère de l'Enseignement Supérieur et de la Recherche et de l'Innovation (MESRI), les TGIR ont pour attribution de mettre en place un système d'organisation des données de la recherche (description, archivage, mise à disposition) tout en sécurisant les accès aux données confidentielles et en participant à la production de grandes enquêtes pluridisciplinaires d'intérêt national. Afin de construire des connaissances et reproduire des raisonnements, l'accès aux données est devenu un enjeu de plus en plus important pour la recherche scientifique. Cependant, on peut voir que la méthode de facilitation d'accès aux données est également primordiale pour cet enjeu.

PROGEDO, une TGIR pour produire et gérer des données quantitatives en sciences humaines et sociales

Suite au rapport « Les sciences sociales et leurs données » de 1999 [2] soumis au ministre de l'Éducation nationale et de la Technologie qui proposait la création d'un institut de diffusion des données en sciences sociales, le **Comité Consultatif des Données en Sciences Humaines et Sociales (CCDSHS)** a été créé (Décret n°. 2001-139 du 12 février 2001). Sa mission était de **définir une politique de données pour les sciences sociales**. En 2008, La feuille de route des infrastructures de recherche françaises prévoyait l'existence d'une infrastructure de recherche appelée PROGEDO (Production et gestion des données en sciences sociales) ayant pour mission de développer en France une politique des données des enquêtes initiées dans le cadre de la recherche en sciences humaines et sociales. Elle a également pour ambition de valoriser les statistiques et les enquêtes publiques produites en France et en Europe. Elle doit également organiser l'accès à ces données provenant de la communauté de recherche. Elle intervient au niveau national sur le portail Quetelet Progedo Diffusion, au niveau régional sur les plateformes universitaires des données et au niveau international sur le partage de grandes enquêtes et bases de données.

Les PUD

Qu'est-ce que c'est ?

Les **PUD sont situées au cœur des universités dans les facultés de Sciences Sociales et Humaines** et sont au nombre de 14 : 12 hébergées dans des Maisons des Sciences Humaines (MSH) en France et 2 à l'Université de Paris et à Sciences Po. Elles sont financées à hauteur de 1 million d'euros par an par le MESRI. La première PUD a vu le jour à Lille en 2002. Elles fonctionnent en réseau et font évoluer leurs activités en relayant de l'information.Leur rôle est d'**encourager au niveau régional l'utilisation des enquêtes et des données quantitatives issues des statistiques françaises, européennes et internationales** proposées par PROGEDO tout en tenant compte des spécificités de chacune au niveau territorial. Chaque PUD a sa couleur locale, par exemple la Plateforme Universitaire de Données d'Aix-Marseille Université (PUD-AMU) est spécialisée sur les données de l'aire méditerranéenne, celle de Rennes (PUD R) sur les données en Bretagne, celle de Grenoble (PUD GA) sur les données du bassin alpin etc.Elles accompagnent également les chercheurs dans les différentes étapes d'une recherche qui nécessite des données quantitatives, et cherchent à atteindre de nouveaux publics comme les doctorants en organisant des colloques, des journées d'études, des ateliers - par exemple, la semaine Data SHS qui a eu lieu simultanément dans l'ensemble des PUD en 2020.

Comment fonctionnent-elles ?

Chaque PUD est une **plateforme de compétences** constituée a minima d'un ingénieur d'études et d'un référent scientifique. L'ingénieur est nommé par l'université dans laquelle est implantée la PUD. Il organise et anime l'activité, développe l'action de la PUD en accompagnant les chercheurs dans les différentes étapes liées à l'analyse quantitative : au moment de la recherche de données existantes, de la prise en main des données, du traitement des données, de l'interprétation et de la présentation des résultats des données concernées. Il est doté de fortes compétences statistiques et de motivations pédagogiques. Quant au référent scientifique, il s'agit d'un enseignant-chercheur du paysage local, qui pilote l'action de la PUD et la relaie auprès des tutelles.

Quelles sources de données y trouve-t-on ?

Les données françaises

(Source : Freepik)

Le portail Quetelet-PROGEDO-Diffusion est le département de la diffusion des données françaises en sciences humaines et sociales à destination de la communauté de recherche, mis en œuvre par la TGIR PROGEDO. Les données sont issues de la statistique publique nationale (grandes enquêtes, recensements, bases de données) et de grandes enquêtes provenant de la recherche française. Le portail permet l'accès à 3 catalogues ADISP, CDSP et INED (cf. rubrique sur les données issues de la statistique publique). Ce portail recense également les accès privilégiés aux enquêtes internationales négociés par PROGEDO comme ESS (*European Social Survey*), ISSP (*International Social Survey Programme*), SHARE (*Survey on Health, Ageing and Retirement in Europe*), MAFE (Migration entre l'Afrique et l'Europe) etc. Quetelet Progedo Diffusion est positionné dans la perspective des données FAIR en offrant l'accès aux données en fonction de leur niveau de sensibilité. Le standard international DDI [3] (*Data Documentation Initiative*) est un protocole de documentation qui est utilisé pour rendre les données réutilisables et permettre l'interopérabilité.

Le CASD (Centre d'Accès Sécurisé aux Données) met à disposition des données confidentielles sur les entreprises et les personnes physiques en proposant un équipement garantissant un accès hautement sécurisé aux données qui lui sont confiées. Il est l'interface entre les producteurs déposants de données et leurs utilisateurs. Le CASD est un groupement d'intérêt public (GIP) de l'Etat. Il est représenté par l'INSEE (Institut National de la Statistique et des Études Économiques), le GENES (Groupe des Écoles Nationales d'Économie et Statistique), le CNRS (Centre National de la Recherche Scientifique), l'École polytechnique et HEC (École des Hautes Études Commerciales) Paris et a été créé par arrêté interministériel du 29 décembre 2018. Selon l'arrêté du 20 décembre 2018 « portant approbation de la convention constitutive du groupement d'intérêt public Centre d'accès sécurisé aux données », le GIP « à vocation industrielle et commerciale,(il) a pour objet principal d'organiser et de mettre en œuvre des services d'accès sécurisé pour les données confidentielles à des fins non lucratives de recherche, d'étude, d'évaluation ou d'innovation, activités qualifiées de « services à la recherche » [Journal officiel de la République Française, 29 décembre 2018].

D'autre part, le GIP a de nombreux rôles. Il doit participer à la création des banques de données en regroupant les données venant de différentes sources (opération appelée l'appariement de données) et en les rendant anonymes. Il doit décrire et stocker les données confidentielles mais également assister aux séances organisées par le Comité du Secret Statistique afin de donner un avis sur la communication de données couvertes par le secret statistique ou fiscal. Il a aussi pour mission d'accompagner l'ensemble des utilisateurs de ces données confidentielles, d'aider à l'homologation des résultats de la recherche venant des données sensibles. Pour finir, le CASD prend part au déploiement permettant d'accéder aux données confidentielles en France, en Europe et à l'international.

Les données confidentielles très détaillées sont accessibles via le CASD en faisant une demande d'accès adressée au Comité du secret statistique pour les données de l'INSEE, du Ministère des Finances-DGFiP, du Ministère du Travail-DARES, du Ministère de l'Environnement-SDES et DPMA, du Ministère de l'Agriculture-SSP, du Ministère de l'Éducation Nationale-DEPP et du CEREQ. De plus, l'utilisateur doit assister à une séance d'information et de sensibilisation obligatoire, appelée séance d'enrôlement. Au cours de cette séance, une carte d'accès à puce sur laquelle sera enregistrée une empreinte digitale sera remise à l'utilisateur.

Zoom sur la certification de résultats : Cascad-CASD [4]

Le CASD et l'agence de certification CASCAD se sont réunis afin de proposer à la communauté de recherche une certification permettant de mentionner la reproductibilité des données confidentielles d'une publication scientifique qui sont hébergées au CASD. L'objectif de ce partenariat est de sécuriser la certification : le code portant sur les données confidentielles est exécuté par une personne habilitée. Cela peut être le DOI (Digital Object Identifier), la version, les sources, les produits…
Cascad est adossée au CNRS et est financée par trois instituts : l'université d'Orléans, le CNRS et HEC Paris.

Les données issues de la statistique publique

L'ADISP (Archives de Données Issues de la Statistique Publique) diffuse des enquêtes et bases de données produites par l'INSEE, plusieurs services statistiques ministériels et institutions de recherche publique.

Le CDSP (Centre de Données Socio-Politiques de SciencesPo) diffuse des enquêtes quantitatives et qualitatives, ainsi que les résultats électoraux du Ministère de l'intérieur.

INED (Institut National d'Études Démographiques) [5] est un catalogue d'enquêtes socio-démographiques de l'Institut depuis 1945.

ELIPSS (Étude Longitudinale Par Internet Pour les Sciences Sociales) [6] est un dispositif d'enquêtes par internet destiné à la communauté scientifique. Il vise à combler l'absence de moyens d'enquête par questionnaires dédiés aux chercheurs en sciences humaines et sociales. Ce panel est mis en place par le CDSP.

(Source : freepik)

Les enquêtes qualitatives BeQuali [7] est une banque d'archivage d'enquêtes qualitatives en science politique et en sociologie proposée par le CDSP.

Les données européennes

Le CESSDA (*Consortium of European Social Science Data Archives*) [8] est le réseau européen des centres d'archives de données en sciences sociales. Son catalogue contient les métadonnées des données dans un environnement sécurisé. Il adhère aux principes de données FAIR pour rendre les données trouvables et fournir des informations sur les données (où elles se trouvent, comment elles sont accessibles). Il propose des outils et des services à la fois aux producteurs de données et aux réutilisateurs de données.

EUROSTAT [9] est le fournisseur de statistiques en sciences sociales de la Commission Européenne. Il produit des statistiques européennes en partenariat avec les instituts nationaux de statistique et d'autres autorités nationales des États membres de l'UE. Ce partenariat est connu sous le nom de Système Statistique Européen (SSE).

Il existe également ESS (*European Social Survey*), SHARE (*Survey of Health, Ageing and Retirement in Europe*), GGP (*Generation and Gender*

PROGEDO
DATA INFRASTRUCTURE

Program), EU-SILC (*Survey on Income and Living Conditions*), ISSP (*International Social Survey Programme*), EVS (*European Values Study*), EWCS (*European Working Conditions Survey*), MAFE (*Migrations between Africa and Europe*), LIS (*Cross-national data center in Luxembourg*), EES (*European Election Studies*).

Les données internationales

Au niveau international, il y a de nombreux sites de centralisation de données : l'ICPSR (*Inter-university Consortium for Political and Social Research*), l'OCDE (Organisme de Coopération et de Développement Economique), WVS (*World Values Survey*) etc.

Quel outil de traitements de données est utilisé ?

La SD-Box est un boîtier qui donne accès à l'infrastructure centrale du Centre d'Accès Sécurisé aux Données (CASD). Il permet aux personnes qui utilisent ou déposent des données de travailler dans un environnement de travail sécurisé et certifié par le comité du secret statistique.

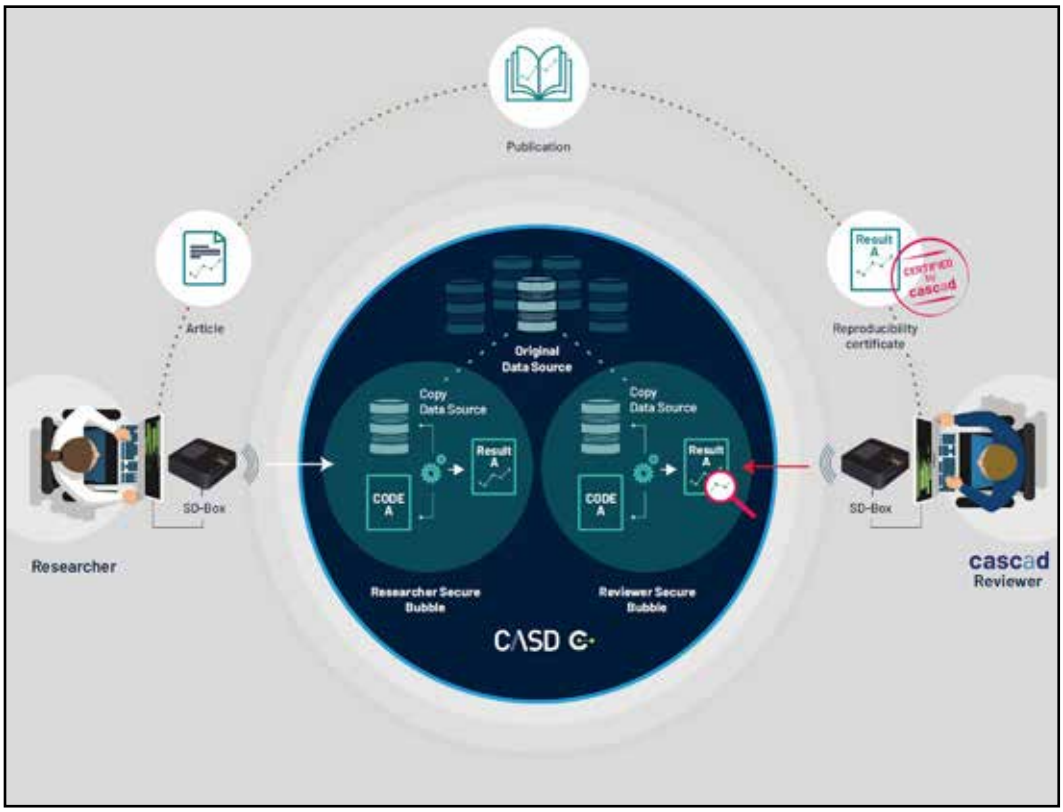
La PUD va accompagner les utilisateurs de cette box pour explorer des données confidentielles, les exploiter et croiser les différentes sources de données. Enfin, elle les aide également à préparer le dossier d'habilitation auprès du comité du secret statistique.

Comme nous avons pu le voir, les PUD permettent d'informer le plus largement possible sur l'ensemble des enquêtes documentées disponibles en France. La majorité des jeux de données à échelle nationale sont accessibles via le portail Quetelet Diffusion, mais également sur les points d'accès aux enquêtes européennes et internationales. Par ailleurs, force est de constater que l'accompagnement humain des chercheurs aux outils informatiques dans la gestion des données (choix des données disponibles et des méthodes pour en tirer le meilleur parti) est la mission remplie par les PUD. Elles permettent de rendre les données statistiques disponibles numériquement en alliant formation en sciences des données et en sciences numériques.

Zoom sur le comité du secret statistique

Il veille à l'accès des données couvertes par le secret statistique ou par le secret fiscal. Le respect des règles du secret statistique permet d'assurer aux personnes, dont les informations sont données à l'établissement de statistiques, la confidentialité sur leur vie personnelle et familiale, et aux entreprises, le secret commercial.

▪ Cécile DESCAMPS



(Circuit de la certification CASD
Source : CASD, https://www.casd.eu)

Références

[1] *https://cache.media.enseignementsup-recherche.gouv.fr/file/Infrastructures_de_recherche/62/2/feuille_route_tgir_2008_527622.pdf*
[2] *https://www.education.gouv.fr/les-sciences-sociales-et-leurs-donnees-12923*
[3] *https://ddialliance.org/*
[4] *https://www.casd.eu/wp/wp-content/uploads/joe_20181229_0301_0053.pdf*
[5] *https://www.comite-du-secret.fr/*
[6] *https://www.casd.eu*
[7] *http://nesstar.ined.fr/webview/*
[8] *https://quanti.dime-shs.sciences-po.fr/fr/*
[9] *https://bequali.fr/fr/*

Bibliographie

Buléon, P. dir. de la publication (2018). Progedo Actu (n°14). https://www.mshb.fr/sites/default/files/PGDactu14.pdf

Chauvel, B., Pannetier, J., Tandar, S. & Tudoux, B. (2021). La Plateforme Universitaire de Données de Nanterre : un service dédié aux données d'enquêtes et aux statistiques à Paris Nanterre. https://pointcommun.parisnanterre.fr/medias/fichier/pudn-point-commun-06012021_1609936415962-pdf

Cour des comptes. (2019). Le pilotage et le financement des très grandes infrastructures de recherche. https://www.ccomptes.fr/system/files/2019-07/20190717-cahier-annexe-TGIR-2.pdf

Da Costa, A., Elegbede, C. (2019). Ouverture des données : spécificités dans le domaine des Sciences Humaines et Sociales. https://databfc2.sciencesconf.org/data/program/12_DataBFC2A_DaCosta_CElegbede_SHS.pdf

Donati, CS (2020.30.03). Entretien avec Clément de Belsunce. https://oaamu.hypotheses.org/1919

Dury, C. (2009.05.06). Jean-Marie Duprez, responsable de la Plateforme Universitaire de données de Lille (PUDL) présente les missions de la PUDL. 25 images SHS. https://25images.msh-lse.fr/data_shs/video/intervention-de-jean-marie-duprez/fr

Marie, M., Niaré, A. (2019). Plate-forme universitaire de données de Caen (PUDC) : rapport d'activité 2019-2019. https://www.unicaen.fr/recherche/mrsh/sites/default/files/public/node/docs/Rapport%202018-2019%20PUDC.pdf

MESRI. (2018). La Feuille de route nationale des Infrastructures de recherche. Enseignement supérieur de la recherche. https://www.enseignementsup-recherche.gouv.fr/cid70554/la-feuille-de-route-nationale-des-infrastructures-de-recherche.html

Ministère de l'économie et des finances. (2018). Arrêté du 20 décembre 2018 portant approbation de la convention constitutive du groupement d'intérêt public « Centre d'accès sécurisé aux données ». https://www.casd.eu/wp/wp-content/uploads/joe_20181229_0301_0053.pdf

Oliveau, S. (2020). Charte des Plateformes Universitaires de Données. http://www.progedo.fr//app/uploads/2020/07/Charte_des_PUD_2020-06-15.pdf

Oliveau S., Blöss-Widmer I., Doignon Y., Belsunce C. de (2020). Aix-Marseille University SSH data platforms : Skills to support research in social sciences and humanities (SSH) in the Mediterranean. Égypte/Monde arabe, 2 (22), pp. 95-105. https://www.cairn.info/revue-egypte-monde-arabe-2020-2-page-95.htm

Plates-formes universitaires de données. (s. d.). Progedo. http://www.progedo.fr/promouvoir/plates-formes-universitaires-de-donnees/.

Progedo. (2019). Quetelet Progedo Diffusion : Enquête de satisfaction. http://www.progedo.fr//app/uploads/2019/07/EnqueteQueteletProgedoDiffusion2019.pdf

PUD Maison Méditerranéenne des Sciences de l'Hommes (2020). Présentation Progedo : tour national des données de la recherche. https://pud.mmsh.univ-aix.fr/wp-content/uploads/2020/01/Pr%c3%a9sentation-PROGEDO.pdf

Silberman, R. (1999). Les sciences sociales et leurs données. https://www.education.gouv.fr/les-sciences-sociales-et-leurs-donnees-12923

Varieras, F. (2017). Présentation Plateforme universitaire de données Strasbourg (PUD-S). https://glisss.hypotheses.org/files/2017/07/Pr%C3%A9sentation_PUDS.pdf

Zolotoukhine, E. (2019). Les catalogues des bases de données de Quetelet-PROGEDO Diffusion : Webinar de Tuto@Mate. https://mate-shs.cnrs.fr/wp-content/uploads/2020/04/tuto19-slides-zolotoukhine-catalogue-progedo.pdf

Huma-Num : une infrastructure au service des SHS.

Nadia NEMER-BRUN

La science ouverte et le principe de l'accessibilité des données et des résultats de la recherche a vu son essor avec l'Appel de Jussieu. Nous pourrions cependant dater sa naissance au 17ème siècle avec la création de l'Académie des sciences et l'apparition de la Revue académique. Aujourd'hui, différents programmes ministériels et européens jouent un rôle majeur dans la construction de l'Espace Européen de Recherche en finançant, notamment au travers de différents projets, les Infrastructures de Recherche et les Très Grandes Infrastructures de Recherche. Ces dernières relèvent de stratégies gouvernementales nationales ou font l'objet de partenariats nationaux et/ou internationaux. Parmi celles-ci, la TGIR Huma-Num. Outre son implication à l'échelle européenne, elle met à la disposition des utilisateurs différents services tout au long du cycle de vie des données résultant de recherches en sciences humaines et sociales.

Huma-Num : un engagement européen au service des sciences humaines et sociales

En tant que Très Grande Infrastructure de Recherche (TGIR), Huma-Num [1] a pour principal intérêt de mutualiser, diffuser et favoriser l'accès aux fichiers et données. Mais elle a également pour mission première d'assurer la **préservation du patrimoine scientifique.**

Huma-Num s'inscrit, par le biais de ses partenaires et au titre du programme **Horizon 2020**, dans plusieurs projets européens [2] visant à soutenir le développement de pratiques conformes aux **principes FAIR**.

Le **projet Triple** a pour ambition de développer une plateforme favorisant la **visibilité** des données de publications en SHS mais également de pouvoir les **réutiliser**. Le but de ce projet, en partenariat avec **OpenEdition**, est également de mettre en relation des communautés disciplinaires, par le biais des profils des chercheurs, et linguistiques différentes tout en travaillant sur une interface basée sur le principe de l'UX Design. Cette plateforme européenne aura pour nom **GO TRIPLE.** Lancé en octobre 2019, ce projet prendra fin en mars 2023.

En partenariat avec le **CINES**, Huma-Num est engagé dans le projet **OPERAS-P** soutenant la mise en œuvre d'une infrastructure de recherche européenne pour les publications SHS en accès libre, sur la base de **l'interopérabilité** entre les services de publication et les différents marché, **OPERAS**. *"OPERAS est l'infrastructure européenne de recherche pour le développement de la communication savante ouverte dans les sciences sociales et humaines."*[3] Ce projet s'achèvera en juin 2021.

(Source : Montage réalisé par Nadia NEMER-BRUN)

De 2013 à 2017, la Commission Européenne a financé un projet de recherche en archéologie appelé **ARIADNE**. Celui-ci a abouti à la création de **ARIADNEplus**, catalogue consultable en ligne et regroupant des jeux de données en archéologie.
Depuis janvier 2019, dans le cadre du programme Horizon 2020, et constatant l'abondance des jeux de données hétérogènes dans leurs structures et formats, deux consortiums de la TGIR apportent leurs contributions dans ce projet : le **Consortium MASA**, dont l'objectif est la **diffusion** et **la mise en oeuvre des principes FAIR** dans la communauté archéologique, et le **Consortium 3D-SHS** qui s'est donné pour principale mission de diffuser **les bonnes pratiques** (méthodologie, formats, schéma métadonnées, etc) en matière d'**acquisition**, de **diffusion** et d'**archivage** des modèles 3D. Ils ont jusqu'en janvier 2022, date butoir du financement, pour concrétiser leurs objectifs.

Une des ambitions d'Horizon 2020 est de mettre en place un cloud destiné aux SHS. Différents acteurs européens, parmi lesquels le **CNRS** (en partenariat avec Huma-Num et MAPS [4]), Sciences Po et **EURISHFIRM** [5], se sont donné pour

mission de créer un environnement sécurisé pour le partage et l'utilisation des données sensibles et confidentielles. **SSHOC**, qui est le nom de ce projet, a également pour but de mettre à la disposition des chercheurs une plateforme, appelée **SSH Open Marketplace**, de services et d'outils répondant aux différentes étapes du cycle de vie de la donnée. Ce projet, en lien avec EOSC[6], s'achèvera en janvier 2022 et répondra aux principes FAIR.

Le dernier projet H2020 dans lequel est impliquée la TGIR est **EOSC-Pillar**[7]. Ce projet ambitieux, réunissant 18 partenaires répartis dans 5 pays européens, a été lancé en juillet 2019, pour une durée de 36 mois, en soutien à EOSC dont l'objectif est d'offrir aux professionnels des SHS un environnement virtuel répondant aux **principes FAIR**, notamment au principe de **réutilisation** des données de recherche.

Si Huma-Num marque son engagement au niveau européen à travers différents projets, la TGIR n'en oublie pas moins son rôle premier qui est de mettre à la disposition des communautés SHS une infrastructure dédiée à leur domaine de recherche.

Huma-Num : des solutions humaines et techniques au service des SHS

Huma-Num est une Très Grande Infrastructure de Recherche (TGIR) impliquée dans le virage numérique de la recherche en sciences humaines et sociales. Son fonctionnement est basé sur une organisation alliant à la fois la **concertation collective** et les **services numériques pérennes**.
Elle apporte son aide aux chercheurs en Sciences Humaines et Sociales et met à leur disposition tout un ensemble de services répondant aux besoins des étapes du cycle de vie des données. Elle propose également des listes d'entraides et d'échanges autour de certains outils de gestion et diffusion de corpus et données de la recherche en SHS sur son serveur de listes. Il est à noter que pour bénéficier des différentes solutions proposées par Huma-Num, l'utilisateur doit appartenir à la communauté académique et le projet scientifique concerné doit être validé par la TGIR. D'autre part, ces mêmes projets doivent s'inscrire dans une démarche d'archivage à long terme des données mais également s'engager sur l'interopérabilité des données de la recherche et des métadonnées associées. Ils doivent également s'inscrire dans une démarche de traitement des données (comme l'enrichissement par exemple). Enfin, il est nécessaire de préciser que l'ensemble des solutions proposées nécessitent au préalable des ouvertures de comptes par le biais du service d'authentification centralisé HumanID [8].

Sa mission est de proposer des solutions d'**organisation**, de **collecte**, de **traitement**, de **préservation**, de **publication** et de **réutilisation des données** [9].

Le projet s'articule autour de différents *Work Packages* (WP) et Huma-Num, pour sa part, est engagée dans les WP 5 et 6. Son implication dans ces WP n'est pas anodine et reprend son principe de base qui est celui de l'application des principes FAIR :

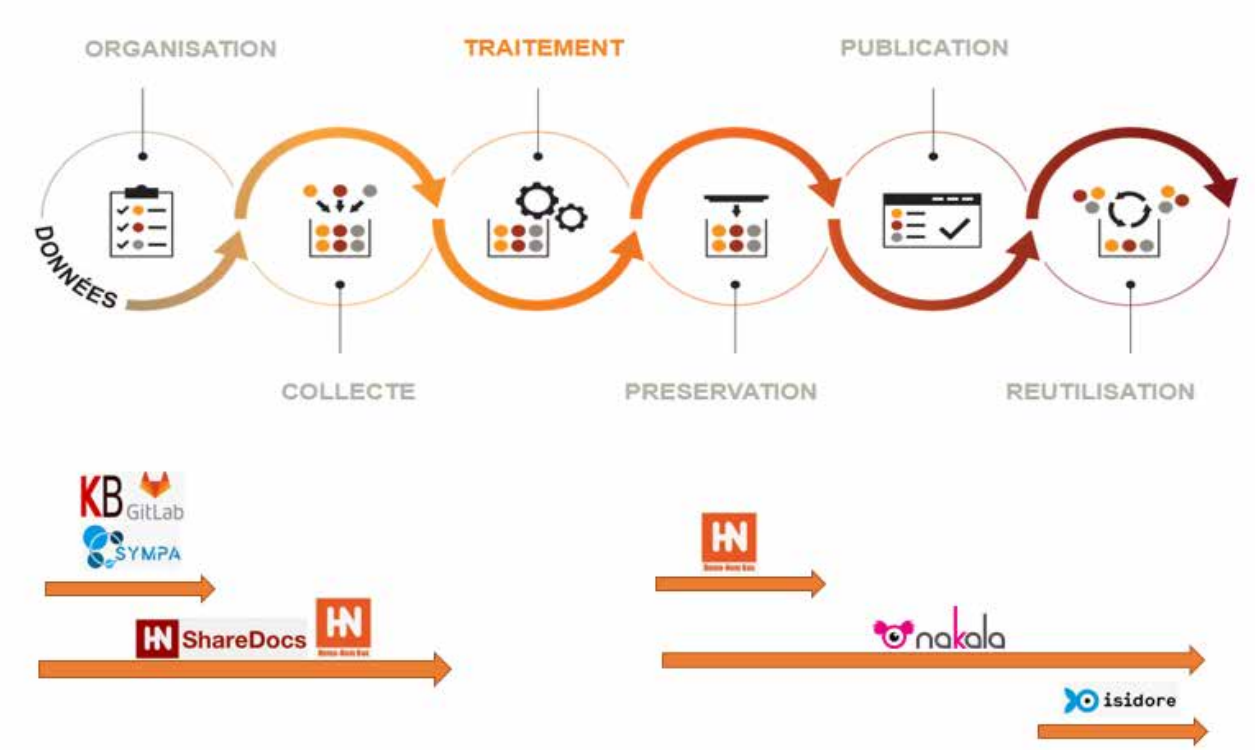
◊ le WP 5, intitulé "*The data layer: establishing FAIR data services at the national and transnational level*" s'intéresse essentiellement à l'utilisateur final en élaborant des supports pédagogiques;

◊ le WP 6, "*EOSC in action : Use cases and community-driven pilots*", a pour finalité de rassembler des cas d'utilisation pour analyser différents outils et service pour la FAIRisation des données.

Les 6 étapes du cycle de vie des données de recherche *[10]*

◊ Création ou collecte des données
◊ Traitement des données
◊ Analyse des données
◊ Conservation des données
◊ Accès aux données données
◊ Réutilisation des données

Cycle de vie des données de la recherche (adapté de UK Data Archive)



Des services à chaque étape
Les différentes étapes ci-dessus se basent sur celles du cycle de vie des données de la recherche, de leur création à leur réutilisation (montage : Nadia NEMER-BRUN à partir des logos disponibles sur le site Huma-Num)

Voyons de plus près ce que sont ces solutions.

Avant toute chose, le chercheur doit **organiser** les données qui seront déposées. Pour ce faire, il dispose de 4 outils : ShareDocs, GitLab, Kanboard et Mattermost. Pourquoi 4 outils? Tout simplement parce que leurs finalités sont différentes.

ShareDocs est un gestionnaire de fichiers (texte, pictural, multimédia...). Sa fonction est de préparer les fichiers pour une édition en ligne ou une diffusion.
GitLab permet de déposer des fichiers de code et d'en maîtriser le partage. Sa particularité est qu'il s'agit là d'une solution open source.
À cette étape d'organisation des données, Huma Num met à disposition trois outils collaboratifs : **Kanboard** qui est un gestionnaire de projet. Mais également en permettant à l'utilisateur de créer, publiposter et gérer des listes de diffusion par le biais de **Sympa**. Il est cependant important de souligner ici que les documents partagés par ce biais ne sont pas pourvus d'un identifiant pérenne et qu'à la dissolution de la liste de diffusion, il ne sont pas sauvegardés. Enfin, il est proposé un service de messagerie instantanée et d'échange de fichier par le biais de l'instance **Mattermost**.

***

L'étape suivante est la **collecte** de

données. Les données présentées pouvant être diversifiées, Huma-Num offre la possibilité de les stocker "telles quelles" et de travailler, par

exemple, sur des fichiers par le biais de **ShareDocs**. L'autre service de stockage sécurisé proposé est celui de la **Huma-Num Box**. Son usage est destiné aux données qui n'ont pas vocation à être consultées fréquemment. Ce sont les données dites froides ou tièdes [11].

Afin de répondre au mieux aux besoins

de l'utilisateur, la TGIR lui propose de les décrire en termes de transformation et d'analyse des données en vue d'optimiser leur **traitement** (comme par exemple la conversion d'un format de fichier vers un autre). Un comité interne d'Huma-Num, le **Comité de la grille**, est chargé d'analyser ces besoins et de proposer les solutions logiciels les plus adéquates en vue de favoriser l'interopérabilité et la pérennité des données [12].

***

A cette étape, il est important de faire un distinguo entre la notion de sauvegarde et celle de **préservation**. La sauvegarde, ou stockage sécurisé, a pour but d'assurer l'intégrité d'un contenu tandis que la préservation a pour fonction première d'assurer la pérennisation de l'information. Ce maintien de lisibilité du contenu et des métadonnées est cependant lié à la problématique de l'obsolescence

des supports et peut nécessiter une conversion des formats de fichiers. Cette étape est majeure car d'elle dépendra celle de la réutilisation dans le temps des données. Ici nous retrouvons **Hum-Num Box** qui est, comme nous l'avons vu à l'étape de "collecte", un service de **stockage sécurisé**.

Concernant la **préservation**, Huma-Num met à la disposition des utilisateurs l'entrepôt **NAKALA** qui est un entrepôt de données sécurisé de deux niveaux : la donnée est décrite, contextualisée et sécurisée dès qu'elle est enregistrée dans Nakala ; la donnée est préservée à long terme par le biais d'un partenariat avec le **CINES** [13]qui assurera, entre autres, l'intégrité du fichier et la migration des formats vers des formats plus pérennes si nécessaire pour le maintien de leur lisibilité.

***

Engagée dans divers projets européens, Huma-Num met à la disposition des utilisateurs un ensemble d'outils en faveur de la **publication** et de **l'éditorialisation** de leurs données, contribuant ainsi à leur structuration et à leur visibilité.

Parmi ces outils nous retrouvons l'entrepôt **NAKALA** et son module de publication **NAKALA-PRESS.** Ce dernier fait suite au pack NAKALONA (qui associait NAKALA et OMEKA [14], CMS permettant de créer des sites web, notamment dans le domaine des éditions scientifiques en ligne). NAKALA-PRESS permet d'éditorialiser les données dans un site web public. En parallèle de cette solution de publication, Huma-Num propose un hébergement web mutualisé et un hébergement de machines virtuelles [15].

***

La dernière étape reprend le 4ème principe FAIR : le "reuse". Le fonctionnement de l'entrepôt **NAKALA** est garant de la **réutilisation** des données hébergées. Par le biais de la standardisation de la description des données, l'attribution d'une licence de diffusion et l'interopérabilité garantie notamment par l'attribution d'un identifiant pérenne, les données deviennent visibles et réutilisables [16].

L'autre service proposé et développé, entre autres, par la TGIR est le moteur de recherche en SHS, **ISIDORE**. Celui-ci moissonne des données en SHS, en français, anglais et espagnol, et les enrichit également dans ces mêmes langues en se basant sur des référentiels métiers. S'inscrivant dans le mouvement de l'*open science*, ISIDORE privilégie *"l'accès*

à des données en libre accès (open access) produites par des organismes de recherche et de l'enseignement supérieur, des laboratoires, des équipes de recherche (...)" [17]. ISIDORE va au-delà du web syntaxique. Il accède à toutes sortes de données en reliant les données et les documents entre eux.

La réutilisation des données à long

terme n'est cependant possible que si leur préservation est assurée. La pérennisation de l'accès à l'information, quel que soit le type de fichier, est un enjeu majeur et est à envisager dès la production et/ou la collecte des données. Le CINES, partenaire de Huma-Num à l'étape "préservation" est l'un des acteurs, avec les Archives nationales et départementales, de cet enjeu national et international [18].

Les différentes solutions présentées ne sont pas destinées à rester telles quelles. Pour exemple, le projet Huma-Num Science Ouverte [19] (HNSO) consiste à améliorer les plateformes NAKALA et ISIDORE afin de renforcer leur adéquation avec les principes FAIR et accroître ainsi la visibilité des données en SHS.

Pour conclure, si nous devions résumer ce qu'est Huma-Num :

Huma-Num est une Très Grande Infrastructure de Recherche, mettant en œuvre des dispositifs humains et technologiques, au service des Sciences Humaines et Sociales, inscrite dans un engagement en faveur d'une science ouverte, appliquant les principes FAIR, et notamment le "reuse", au travers de différents projets européens financés par le programme Horizon2020.

▪ **Nadia NEMER-BRUN**

Références

[1] *Huma-Num est un acronyme pour Humanités Numériques : domaine de recherche regroupant les Sciences Humaines et Sociales, l'Informatique, l'Ingénierie et les Arts et les Lettres.*

[2] *Projets. (s. d.). Huma-Num. Consulté 12 mai 2021, à l'adresse https://site2020.huma-num.fr/projets-internationaux/*

[3] *"OPERAS is the European Research Infrastructure for the development of open scholarly communication in the social sciences and humanities." https://operas.hypotheses.org/category/projects/operas-p*

[4] *Réseau MAPS : Réseau thématique de modélisation multi-agents appliquée aux phénomènes spatialisés https://maps.hypotheses.org/*

[5] *EURHISHFIRM a pour projet de concevoir une infrastructure de recherche de classe mondiale pour collecter, fusionner, extraire, rassembler, aligner et partager des données historiques détaillées de haute qualité au niveau des entreprises pour l'Europe. https://eurhisfirm.eu/*

[6] *EOSC, en étant le portail d'accès aux données, services et ressources des instituts de recherches européens, facilite l'interopérabilité et l'échange. https://www.ouvrirlascience.fr/portail-web-de-leosc/*

[7] *Pour en savoir plus : https://www.eosc-pillar.eu/*

[8] *https://humanid.huma-num.fr/*

[9] *L'ensemble des étapes et des outils présentés ici sont consultables à l'adresse suivante : https://www.huma-num.fr/les-services-par-etapes/*

[10] *Une introduction à la gestion et au partage des données de la recherche—Le cycle de vie des données. (s. d.). Consulté 11 mai 2021, à l'adresse https://www.inist.fr/wp-content/uploads/donnees/co/module_Donnees_recherche_7.html*

[11] *Les données sont classées en trois catégories : les données "chaudes" désignent les données dont l'accès est fréquent, les données "tièdes" sont consultées de façon plus modérée et les données "froides" sont les données qui n'ont plus lieu d'être consultées (https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-fra.html?lang=fra&i=&index=frt&srchtxt=DONNEES%20CHAUDES)*

[12] *https://www.huma-num.fr/presentation/#cogrid*

[13] *https://www.cines.fr/archivage/typologies/donnees-scientifiques/*

[14] *Pour en savoir plus : https://omeka.fr/presentation-omeka*

[15] *Pour en savoir plus : https://documentation.huma-num.fr/hebergement-web/*

[16] *Pour en savoir plus : https://cat.opidor.fr/index.php/Nakala*

[17] *Pour en savoir plus : https://isidore.science/about*

[18] *Différents groupes de travail œuvrent actuellement sur cette problématique des formats de fichiers, et donc de la préservation et la pérennisation des données. C'est le cas, en France, la Cellule nationale de veille instituée par le Groupe PIN, de l'Association ARISTOTE, dont le CINES et la BnF, pour ne citer qu'eux, font partie.*

[19] *rédaction. (s. d.). Huma-Num Science Ouverte, un projet soutenu par le Fonds national pour la science ouverte (FNSO) [Billet]. Le blog d'Huma-Num et de ses consortiums. Consulté 12 mai 2021, à l'adresse https://humanum.hypotheses.org/6407*

Analyse comparée des entrepôts de données européens et africains. Comment opérer un choix pour les données de recherche?

Gérard BODO

Cet article s'intéresse au sujet des entrepôts de données : il entend donner une explication claire de cette notion, et des spécificités propres à ce sujet. De plus, il dresse une liste claire des exigences normatives actuelles et des actions qui peuvent être mises en œuvre pour optimiser un entrepôt de données au-delà de sa localisation géographique.

Le mouvement d'ouverture des données de recherche sur la scène européenne ou africaine semble entraîner une ruée des chercheurs vers les entrepôts de données. Dans ce contexte, des facteurs comme l'obsolescence des formats de fichier, le développement des nouvelles pratiques informationnelles, c'est-à-dire la manière dont l'on s'approprie un dispositif d'information donné, font en sorte que l'analyse qui précède la mise en place ou le choix d'un entrepôt de données de recherche paraît complexe. Car le chercheur qui souhaite stocker les données de recherche devra cerner les usages d'un dispositif par rapport à la donnée qu'il souhaite valoriser. Mais aussi identifier la structure de l'outil choisi pour en déterminer les spécificités. Dans cet article, il s'agira pour nous de cerner la notion d'entrepôt de données ; et du comment les choisir pour l'hébergement des données de recherche. Ensuite nous dresserons une analyse comparée de quelques entrepôts de données de la scène européenne et africaine.

Tout savoir sur les entrepôts de données

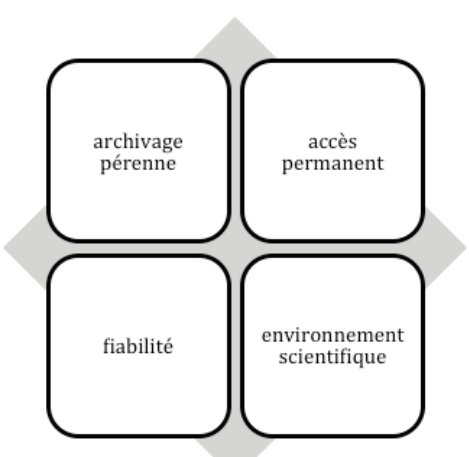
Pourquoi un entrepôt ?

Dans le monde de la recherche scientifique, il est important de collaborer et de faire connaître les travaux de recherche. Cela atteste de l'avancée de la recherche et de la crédibilité du chercheur (autorité scientifique). C'est pourquoi les entrepôts de données sont des dispositifs web qui permettent d'identifier la structure (formats, métadonnées, contenus etc.) des données de recherche stockées par les chercheurs. De plus, ils permettent une conservation pérenne et une utilisation ultérieure des données stockées. D'après le CIRAD : « un entrepôt de données de recherche (*Research Data Repository ou Data Repository*) est une base de données destinée à accueillir, conserver, rendre visibles et accessibles des données de recherche » [1].
Autrement dit, il s'agit d'un réservoir susceptible de stocker diverses informations de manière structurée et dans lequel cette information peut être retrouvée.

Quelle est la structure d'un entrepôt de données ?

Un entrepôt de données de recherche est composé d'une structure unique qui prend en compte 4 éléments essentiels d'après Hélène Prost et Joachim Schöpfel [2] (fig. 1).
L'archivage pérenne tels que décrit par la norme (ISO 14641:2018, relative à l'Archivage électronique - Conception et exploitation d'un système informatique pour la conservation intègre de documents électroniques – Spécifications) [3] doit répondre à un ensemble d'exigences liées à la conservation du document, l'accessibilité et l'intelligibilité.
En d'autre termes, les supports des données déposés doivent maintenir leurs intégrités comme au moment du dépôt par leurs auteurs, les données doivent êtres exploitables par les terminaux de lecture et d'écriture, et enfin les données doivent comporter des éléments additionnels (métadonnées) répondant à des standards de description des données (Dublin Core, KBart, Marc XML, TEI, METS etc.) qui faciliteront leur compréhension et leur exploitation par n'importe quel utilisateur ou système d'information.

De plus, l'accès permanent fait référence à la disponibilité en toute circonstance des données. C'est la raison pour laquelle, chaque donnée possède un identifiant pérenne appelé DOI (Digital Object Identifier) qui permet de la différencier d'une autre donnée ayant les mêmes propriétés en termes de format, support etc. La fiabilité est la capacité pour une donnée de pouvoir être citée et vérifiée.
Et enfin l'environnement scientifique prend en compte l'idée selon laquelle chaque donnée produite a été réalisée dans le cadre d'une activité scientifique, qui appartient à un domaine de connaissance précis ayant un objet d'étude et des méthodes.



Structure d'un entrepôt de données

Qui sont les utilisateurs des entrepôts ?

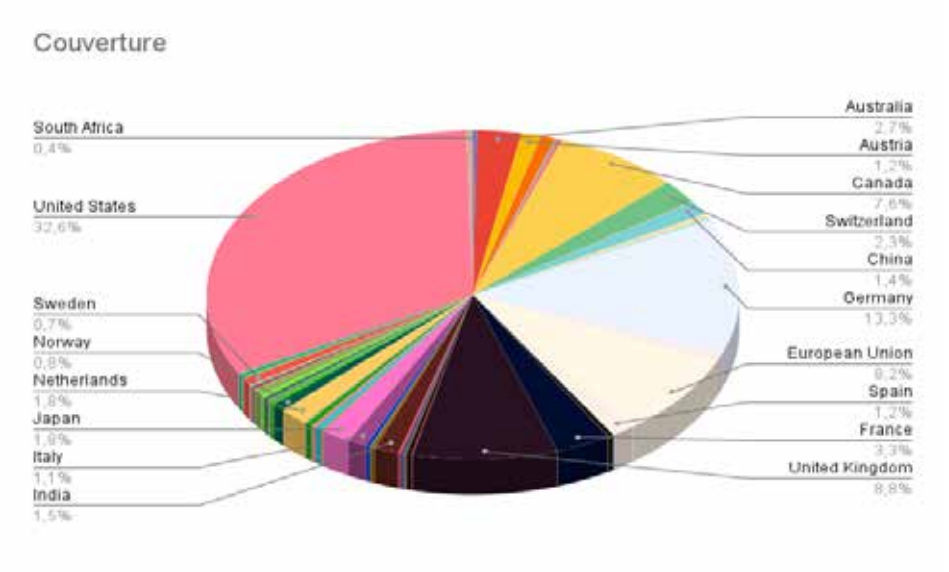
Lorsque vient le moment de la publication ou la consultation des données de recherche, on aperçoit généralement 5 types de publics à savoir :

- Professionnels des Sciences de l'Information et de la Communication (SIC) ;
- Chercheurs ;
- Laboratoires de recherche ;
- Étudiants ;
- Organisations publiques ou privées.

En fonction du profil de l'utilisateur, le type d'entrepôt utilisé peut être différent. Les laboratoires de recherche, chercheurs, étudiants seront attirés par des entrepôts disciplinaires ou multidisciplinaires. Par ailleurs, les professionnels des SIC et ou les organisations utilisent les entrepôts disciplinaires ou multidisciplinaires, et mixtes.

Pouvons-nous chiffrer les entrepôts existants ?

Les données statistiques de l'annuaire re3data.org [4] en date du 20/05/2021 (Fig.2) révèlent que les États Unis (USA) possèdent le volume le plus élevé en termes d'entrepôts de données. Ensuite l'Allemagne, la Grande Bretagne, l'Union européenne et le Canada. En ce qui concerne les pays du SUD, l'Afrique du Sud, est le pays qui est le mieux couvert en termes d'entrepôts de données. Comment le chercheur doit s'y prendre pour opérer son choix ?



Répartition géographique des entrepôts de données dans le monde

Mieux choisir son entrepôt de données
Quels critères pour le choix d'un entrepôt de données ?

D'entrée de jeu, choisir un entrepôt pour héberger les données de recherche n'est pas une mince affaire, et avec l'évolution galopante des principes de la science ouverte il est nécessaire de repasser ces critères de choix au peigne fin. C'est pourquoi, il faut évoluer d'une manière graduelle, sous forme de check-list. Il faut se poser les bonnes questions, ci-dessous quelques questions qui devraient guider vos choix. :

• *Est-ce que les données de recherches à disposition sont FAIR (Findable, Accessible, Interoperable, Reusable) ?*
Initié en 2014 par FORCE 11, et publié en 2016, les Principes FAIR [5] sont un ensemble de résolution ayant pour but de faciliter la découverte, l'accès, l'interopérabilité et la réutilisation des données de recherche mis à disposition par les chercheurs.
Cette opération d'évaluation peut être automatisée à l'aide d'outil, comme *FAIR data assessment tool* [6] qui permet d'évaluer le caractère « FAIR » d'un jeu de données.

• *Est-ce que cet entrepôt attribue des identifiants pérennes aux ressources hébergées ? A titre d'exemple nous avons (idHAL, DOI, URI etc.)*

• *Existe-t-il des métadonnées descriptives qui décrivent à la fois le support et le contenu de la donnée afin de faciliter son repérage et sa compréhension ?*
Les métadonnées, descriptives obéissent à plusieurs standards, nous pouvons trouver en fonction des dispositifs : *Text Encoding Initiative* (TEI), *Dublin Core* (DC), *Metadata Encoding and Transmission Standard* (METS) etc.

• *Existe-t-il des conditions de diffusion, de partage ou de réutilisation des données hébergées ?*
L'un des ratios sur lequel vous pouvez vous appuyer, c'est la licence. Il en existe plusieurs désignés sous le nom de Creative Commons.

• *L'entrepôt permet-il de citer systématiquement les jeux de données ?*
Chaque donnée a un niveau de communicabilité, fonction de la politique de diffusion définie par l'auteur, l'utilisateur peut avoir accès ou non au contenu de la ressource. Mais cela ne doit en aucun cas empêcher la ressource d'être citée.

• *Existe-t-il un Data Management Plan (DMP) ?*
Ce processus basé sur le circuit et le cycle de vie des données, prend en compte les délais légaux de conservation, les différentes politiques de conservation, etc.

• *Est-ce un entrepôt de données certifié ?*
Le *Coretrusteal* est un organisme de certification des entrepôts de données de la recherche.
Toutefois, il existe des outils qui répondent aux questions ci-dessus et qui sont compatibles aux principes FAIR. C'est le cas de *Search Repositories* [7] sur le site du ***Coretrusteal, Repository Finder*** [8] de ***Data Cite***. Et l'annuaire re3data. org. [9] que nous recommande d'utiliser Sylvie Cocaud et Pascal Aventurier en 2017 lors du séminaire de l'INRA sur « les entrepôts de données de recherche » [10].

• *Comment un entrepôt peut passer FAIR ?*
Cette question revient assez souvent, il n'existe pas de solution miracle pour rendre un entrepôt FAIR. Tout débute par la mise en application des exigences desdits principes au sein de votre entrepôt ou la prise en compte de ces aspects dans votre stratégie projet, c'est plus une question de bonne pratique d'exploitation et d'ouverture des données.

Quelques exemples d'entrepôts Européen et Africain
Zenodo [11] : est un entrepôt de l'Union européenne développé par l'Organisation Européenne pour la Recherche Nucléaire (CERN) créé en 2013, il est parfaitement interopérable avec GitHub, une infrastructure américaine d'hébergement de code source. Actuellement sur cette plateforme, il est possible de charger jusqu'à 50 GB par jeu de données. Chaque jeu de données publiés possède un DOI ayant le préfix « 10.5281 » et le radical « zenodo ». Les données hébergées couvrent les sciences humaines et sociales, les sciences de la vie, et enfin les sciences de l'ingénieur. En matière de contenu, il est courant de retrouver sur zenodo des codes sources, des archives, des documents sonores, textuels, des images, des fichiers de données et enfin des vidéos. En terme de métadonnées, nous n'allons pas aborder l'ensemble, mais nous présenterons celles qui nous semblent les plus significatives à savoir : titre, auteur, date, doi, mots clés, résumé.

Nakala [12]: est une infrastructure française de recherche en sciences humaines et sociales, conçue par Huma-num en 2016. Disponible en 3 langues à savoir le français, l'espagnol et l'anglais. Les contenus pris en charge sont variés, nous avons les images, textes, données et sons. Il est possible d'affiner la recherche d'un jeu de données en fonction des critères : type, licence, année et pertinence. Les identifiants pérennes sont composés d'un préfix « 11280 » et un radical spécifique au jeu de donnée décrit.

Quant au standard utilisé pour la description des métadonnées, il s'agit du Dublin Core, nous retrouvons des champs tels que : DOI, titre, date, auteur, licence, type.

Datafirst [13]: Conçu par l'université de Cape Town afin d'offrir un accès libre aux données en Open Data du gouvernement Sud-Africain, et aux gouvernements ou institutions africaines. Ce dispositif ne fournit pas suffisamment d'information sur sa politique de protection ou de conservation des données, il s'appuie sur le standard Dublin Core pour décrire l'ensemble des métadonnées qui sont hébergées au sein de celui-ci, en guise de métadonnées nous pouvons rencontrer le titre, l'identifiant pérenne, l'année, le pays de dépôt, l'éditeur, l'auteur, et la date etc.

Egypt's Information Portal (EIP) [14] : est géré par l'Egyptian Cabinet, Information and Decision Support Center. Ce dispositif sert de vitrine au gouvernement égyptien en matière d'*Open Data*. Depuis quelques mois le gros des contenus de la plateforme a été migré sur les systèmes embarqués Android et IOS. Les domaines disciplinaires, comme les sciences humaines et sociales, les sciences naturelles, les sciences de la vie, les sciences de l'ingénieur sont couverts. Les métadonnées identifiées nous avons titre, éditeur, lien du fichier, résumé, et enfin date.

Évaluation du niveau de FAIR des entrepôts Zenodo, Nakala, Datafirst, EIP

Notre analyse s'appuie sur plusieurs critères évoqués au sein de cet article afin d'identifier avec précision les dispositifs les plus conformes aux principes FAIR.

Critères FAIR	Zenodo	Nakala	Datafirst	EIP
Accès libre	✓	✓	✓	✓
Identifiant pérenne	✓	✓	✓	×
Métadonnées	✓	✓	✓	✓
Licence	✓	✓	✓	✓
Citabilité	✓	✓	✓	×
DMP	✓	✓	✓	×
Certification	×	×	✓	×
Politique de dépôt	✓	✓	✓	✓

Analyse comparée

Cette analyse révèle que les entrepôts des deux aires géographiques semblent ne pas être totalement conformes aux exigences FAIR. D'où l'absence de certification dans la plupart des cas. Cependant, seul ***Data First*** est totalement conforme aux exigences FAIR. Pour pallier les manquements des autres entrepôts nous proposons un ensemble de stratégie à mettre en place.

En définitive, pour faciliter l'accès et la réutilisation des données de recherche, nous devons chacun apporter notre contribution ceci passe par :

a) L'usage des technologies libres dans nos projets web de mise en place d'entrepôt. En guise d'exemple HTML, PHP, PYTHON, CSS, etc. ;

b) Privilégier les formats libres pour nos entrepôts : TAR, XML, ODT, PDF, etc. ils sont facilement interopérables, et reconnus par la majeure partie des terminaux de lecture ;

c) Exploiter les standards ou normes connus de description des métadonnées à savoir : Dublin Core [15], TEI [16], KBart [17], METS [18], etc. ;

d) Privilégier les protocoles favorisant l'archivage pérenne au sein des dispositifs en *Open Access*. Par exemple OAI-PMH [19] ;

e) Rendre les serveurs de données disponibles ;

f) Sensibiliser les propriétaires d'entrepôts aux bienfaits de la certification des entrepôts «***Why certification*** » [20].

▪ Gérard BODO

(Source : Freepik)

Références

[1] *Centre de coopération internationale en recherche agronomique pour le développement (CIRAD). 1 - Qu'est-ce qu'un entrepôt de données de recherche [en ligne]. Disponible sur : https://coop-ist.cirad.fr/gerer-des-donnees/deposer-des-donnees-dans-un-entrepot/1-qu-est-ce-qu-un-entrepot-de-donnees-de-recherche (Consulté le 20/03/2021)*
[2] *PROST Hélène, SCHÖPFEL Joachim. Les entrepôts de données en sciences de l'information et de la communication (SIC). Une étude empirique. Études de communication [En ligne], 2019, mis en ligne le 01 janvier 2021, 52. Disponible sur : http://journals.openedition.org/edc/8604 (Consulté le 20/05/2021)*
[3] *ISO. L'Archivage électronique - Conception et exploitation d'un système informatique pour la conservation intègre de documents électroniques - Spécifications . Disponible sur : https://www.iso.org/fr/standard/74338.html*
[4] *https://www.re3data.org/browse/by-country/ (Consulté le 20/05/2021)*
[5] *FAIR Principles [en ligne]. go-fair, 2016 . Disponible sur : https://www.go-fair.org/fair-principles/*
[6] *https://www.surveymonkey.com/r/fairdat (Consulté le 20/03/2021)*
[7] *Search Repositories - CoreTrustSeal [en ligne] . Disponible sur : https://www.coretrustseal.org/why-certification/search-repositories/ (Consulté le 20/03/2021)*
[8] *DataCite Repository Selector [en ligne] . Disponible sur : https://repositoryfinder.datacite.org/ (Consulté le 20/03/2021)*
[9] *Home | re3data.org [en ligne]. Disponible sur : https://www.re3data.org/ (Consulté le 20/03/2021)*
[10] *https://anfdonnees2017.sciencesconf.org/data/pages/Entrepots_ANFRenatis_2017_Cocaud_Aventurier_1.pdf*
[11] *Zenodo - Research. Shared. [en ligne] . Disponible sur : https://zenodo.org/ (Consulté le 20/03/2021)*
[12] *Nakala [en ligne] . Disponible sur : https://www.nakala.fr/ (Consulté le 20/03/2021)*
[13] *DataFirst – Home [en ligne]. Disponible sur : https://www.datafirst.uct.ac.za/ (Consulté le 20/03/2021)*
[14] *http://www.eip.gov.eg/Default.aspx*
[15] *https://dublincore.org/*
[16] *https://tei-c.org/*
[17] *https://www.openedition.org/26973*
[18] *https://www.loc.gov/standards/mets/mets-home.html*
[19] *http://www.openarchives.org/pmh/*
[20] *https://www.coretrustseal.org/why-certification/*

Bibliographie

AVENTURIER Pascal, DESCONNETS Jean-Christophe.(2019). FAIRISATION des Entrepôts. Atelier JNSO 2019. IRD – Institut de Recherche pour le Développement (France) https://jnso2019.sciencesconf.org/data/pages/INTROD_1.PDF

COCAUD Sylvie, AVENTURIER Pascal.(2017). Les entrepôts de données de recherche. Participer à l'organisation du management des données de la recherche, gestion de contenu et documentation des données. Action Nationale de Formation organisée par les réseaux Renatis et Médici, Centre National de la Recherche Scientifique (CNRS), pp.63 . https://hal.archives-ouvertes.fr/hal-01595599

PROST Hélène, SCHÖPFEL Joachim.(2019). Les entrepôts de données en sciences de l'information et de la communication (SIC). Une étude empirique. Études de communication, 52. http://journals.openedition.org/edc/8604

Des données FAIR & des entrepôts de données TRUST : la combinaison parfaite pour la science ouverte.

Selin GUDER

Les technologies de l'information et de la communication étant devenues omniprésentes dans notre société, nous sommes de plus en plus dépendants des données numériques et des dispositifs techniques qui permettent d'accéder à ces ressources et de les utiliser. Les infrastructures de recherche doivent gagner la confiance des communautés qu'ils entendent servir et démontrer qu'elles sont fiables et capables de répondre aux critères d'excellence scientifique et technologique. Dans le cadre de la science ouverte et de l'ouverture des publications et données de la recherche, la RDA (Research Data Alliance) a publié l'article The TRUST Principles for digital repositories dans lequel sont décrits les principes que doivent suivre les entrepôts de données pour maintenir leur fiabilité et pour lequel nous vous en proposons une note de lecture.L'objectif de cet article est de rappeler aux responsables de la gestion des données de recherche - c'est-à-dire - les chercheurs, la nécessité d'adopter le modèle FAIR à savoir rendre les données faciles à trouver, accessibles, interopérables et réutilisables tout en ayant des entrepôts fiables, dotés d'une gouvernance et de cadres organisationnels durables via les principes TRUST.

Les principes TRUST, de quoi s'agit-il ?

(Source : Research Data Alliance)

Transparency (Transparence) : « Faire preuve de transparence en ce qui concerne les services d'entrepôt spécifiques et les collections de données, vérifiables par des preuves accessibles au public [1] ».
Pour se conformer à ce principe, les entrepôts de données doivent s'assurer que, au minimum, leur déclaration de mission et leur champ d'application soient clairement énoncés. En outre, les aspects tels que les conditions d'utilisation, le délai minimum de conservation numérique pour les fonds de données ou encore toute caractéristique ou service supplémentaire pertinent, par exemple la capacité à gérer de manière responsable les données sensibles doivent être déclarés de manière transparente.

La communication claire des politiques relatives aux dépôts et, en particulier, des conditions d'utilisation des fonds de données permet d'informer les utilisateurs de toute limite qui pourrait restreindre leur utilisation des données ou du dépôt. De même, le fait de pouvoir évaluer facilement si un dépôt peut traiter des données sensibles de manière responsable éclairerait également leur décision d'utiliser ou non les services de données disponibles.

Responsibility (Responsabilité) : « Être responsable de l'authenticité et de l'intégrité des données détenues, ainsi que de la fiabilité et de la pérennité de son service [2]. »

Les dépôts dignes de confiance assurent la responsabilité de la gestion de leurs fonds de données et du service de leur communauté d'utilisateurs. Cette responsabilité est démontrée par :
- L'adhérence aux métadonnées et aux normes de conservation de la communauté désignée, tout en assurant la gestion des fonds de données, par exemple la validation technique, la documentation, le contrôle de la qualité, la protection de l'authenticité et la persistance à long terme ;
- L'offre des services de données, par exemple des interfaces de portail et de machine, le téléchargement de données ;
- La gestion des droits de propriété intellectuelle des producteurs de données, la protection des ressources d'information sensibles et la sécurité du système et de son contenu.

Les utilisateurs de l'infrastructure de recherche doivent avoir l'assurance que les déposants de données sont invités à fournir toutes les métadonnées conformes aux normes de la communauté. Le fait de savoir qu'un entrepôt de données vérifie l'intégrité des données et des métadonnées disponibles garantit aux utilisateurs potentiels que les fonds de données sont plus susceptibles d'être interopérables avec d'autres ensembles de données pertinents. Les déposants et les utilisateurs doivent avoir l'assurance que les données resteront accessibles au fil du temps et qu'elles pourront donc être citées et référencées dans des publications.

User focus (Orientation vers l'utilisateur) : "Veiller à ce que les normes de gestion des données et les attentes des communautés d'utilisateurs cibles soient respectées." [3].

Les entrepôts de données ont un rôle essentiel dans l'application et le respect des normes et standards de la communauté d'utilisateurs cible, car la conformité facilite l'interopérabilité et la réutilisation des données. Les normes de données que les dépôts dignes de confiance doivent appliquer comprennent les schémas de métadonnées, les formats de fichiers de données, les vocabulaires contrôlés, les ontologies et d'autres sémantiques lorsqu'elles existent dans la communauté des utilisateurs.

Un entrepôt de données peut démontrer son adhésion à ce principe en implantant des données métriques pertinentes et en les mettant à la disposition des utilisateurs, en contribuant à des catalogues communautaires pour faciliter la découverte de données et en surveillant l'évolution des attentes de la communauté afin de répondre, si

nécessaire, à ces besoins nouveaux.

L'utilisation et la réutilisation des données de recherche font partie intégrante du processus scientifique et, par conséquent, ces entrepôts devraient permettre à leur communauté de trouver, d'explorer et de comprendre leurs fonds de données en ce qui concerne la (ré)utilisation potentielle. Les dépôts devraient encourager les utilisateurs à décrire pleinement les données au moment du dépôt et faciliter leur retour d'information sur tout problème lié aux données (par exemple, la qualité ou l'aptitude à l'emploi) qui pourrait apparaître après la mise à disposition des données.

Sustainability (Durabilité) : "Maintenir les services et préserver les collections de données à long terme." [4]
Un dépôt digne de confiance peut démontrer la durabilité de ses fonds en :
- Planifiant suffisamment l'atténuation des risques, la continuité des activités, la reprise après sinistre et la succession ;
- Fournissant des fonds pour permettre une utilisation continue et maintenir les propriétés souhaitables des ressources de données que le dépôt a été chargé de préserver et de diffuser ;
- Assurant une gouvernance pour la préservation à long terme des données, afin que les ressources de données restent accessibles et utilisables à l'avenir.

Assurer sa durabilité est nécessaire pour assurer un accès ininterrompu à ses précieux fonds de données pour les communautés d'utilisateurs actuelles et futures. L'accès continu aux données dépend de la capacité du dépôt à fournir des services au fil du temps et à répondre avec des services nouveaux ou améliorés pour répondre aux besoins changeants de la communauté des utilisateurs.

Technology (Technologie) : « Fournir l'infrastructure et les capacités nécessaires pour soutenir des services sécurisés, pérennes et fiables » [5].
Une infrastructure de recherche dépend de l'interaction des personnes, des processus et des technologies pour prendre en charge des services sécurisés, persistants et fiables. Ses activités et fonctions sont soutenues par des logiciels, du matériel et des services techniques. Ensemble, ils fournissent les outils nécessaires à la mise en œuvre des principes TRUST.
Un entrepôt de données peut démontrer l'adéquation de ses capacités technologiques avec la mise en œuvre de normes, d'outils et de technologies pertinents et appropriés pour la gestion et la conservation des données ainsi que la mise en place de mécanismes pour prévenir, détecter et répondre aux menaces de sécurité cybernétique ou physique.

Le modèle FAIR & TRUST : des principes complémentaires

Les entrepôts FAIR sont des entrepôts qui favorisent la découverte, l'accès, l'interopérabilité et la réutilisation des données qu'ils hébergent et fonctionnent comme une ligne directrice pour faciliter la découverte et la réutilisation des connaissances scientifiques. Les entrepôts de données TRUST garantissent la conservation à long terme des données ouvertes et comportent au minimum les métadonnées FAIR, et leur gestion des données prend en compte de manière adéquate la vie privée, la sécurité et la propriété intellectuelle, notamment en ce qui concerne les données personnelles. Dans ce but, et pour contribuer à augmenter le nombre d'entrepôts de données certifiés TRUST, les académies des sciences recommandent leur utilisation et soutiennent le travail du *World Data System* (WDS) du Conseil international pour la science.

Les deux principes s'entrecoupent et se renforcent mutuellement de manière à favoriser l'élaboration de données FAIR et leur préservation dans les dépôts qui adoptent les principes TRUST.

Les scientifiques doivent être en mesure de trouver, de réutiliser, de déposer et de partager des données via des dépôts de données fiables qui mettent en œuvre les principes de données FAIR et qui garantissent la durabilité à long terme. Les entrepôts de données doivent être faciles à trouver et à identifier, et offrir aux utilisateurs une transparence totale sur leurs services. Ces dispositifs techniques doivent avoir des politiques transparentes, une organisation, des ressources financières et humaines adéquates pour assurer leurs missions de mise à disposition des données de manière durable et sécuritaire.
"L'adoption des principes FAIR et la mise en œuvre des principes TRUST donnent ainsi aux utilisateurs l'assurance qu'ils bénéficient d'entrepôts sûrs avec des moyens durables. Pour les auteurs, les principes TRUST constituent un moyen mnémotechnique de rappeler la nécessité de développer et d'entretenir les infrastructures afin de favoriser une gestion continue des données et de permettre l'utilisation future des collections de données." [6]

Comment identifier un entrepôt de données TRUST ?

Avec la Certification coretrustseal

Le ***CoreTrustSeal*** [7] est une organisation de certification mise en place conjointement par le DSA (*Data Seal of Approval*) [8] et le ICSU WDS (*the International Council for Science's World Data System*) [9]. Cette certification garantit aux déposants que leur données seront protégées et gérées de manière optimale.
La *certification CoreTrustSeal* évalue des critères relatifs aux entrepôts de données selon plusieurs niveaux de conformité. Cette étape de certification est importante pour garantir la fiabilité et la durabilité des dépôts de données ainsi que l'archivage et le partage à long terme des données. La Research Data Alliance (RDA) fournit un cadre commun pour la mise en œuvre et la maintenance des dépôts numériques selon 16 critères organisés en 3 thèmes (Infrastructure organisationnelle, Gestion des objets numériques des données et des métadonnées et Technologie et Sécurité).

Les avantages qu'apporte la certification

La certification offre de nombreux avantages à une infrastructure de recherche et à ses parties prenantes. "La certification *CoreTrustSeal* est envisagée comme la première étape d'un cadre mondial pour la certification des référentiels qui comprend la certification de niveau étendu (nestor-Seal DIN 31644) et la certification de niveau formel (ISO 16363). En fin de compte, *CoreTrustSeal* s'efforcera également de fournir une certification de niveau de base à d'autres entités de recherche telles que les services de données et les logiciels." [10] Par ailleurs, la certification implique un processus peu intensif par lequel les entrepôts de données fournissent la preuve qu'ils sont durables et dignes de confiance. Un entrepôt effectue d'abord une auto-évaluation interne, qui est ensuite examinée par des pairs de la communauté. Ces évaluations permettent d'améliorer la qualité et la transparence de leurs processus, et à mieux connaître et respecter les normes établies.

En plus des avantages externes, tels que le renforcement de la confiance des parties prenantes, l'amélioration de sa réputation et la démonstration que l'entrepôt de données suit de bonnes pratiques, la certification de base offre un point de référence pour la comparaison et aide à déterminer les forces et les faiblesses de l'infrastructure de recherche.

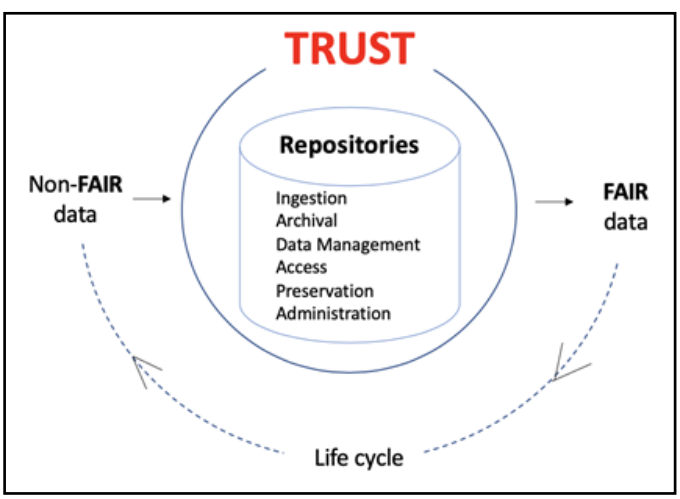
Enfin, le CoreTrustSeal donne aux producteurs de données l'assurance que les données soient préservées et restent réutilisables - à savoir FAIR– à l'avenir et donne aux organismes de financement l'assurance que les investissements sont optimisés. Les propriétaires veulent s'assurer que leurs données sont en lieu sûr et restent accessibles, utilisables et signifiants au fil du temps. Les utilisateurs veulent avoir la certitude que les données ont été préservées correctement et sont de haute qualité.

Les prérequis de la certification

Les entrepôts qui disposent d'une certification CoreTrustSeal permettent de garantir un bon niveau de compatibilité des données avec les principes FAIR, les prérequis de la certification étant compatibles avec les principes FAIR. Pour les entrepôts ne disposant pas de certification, les questions ci-dessous permettent d'évaluer leur niveau de compatibilité avec les principes FAIR.

- Des identifiants uniques et pérennes (par exemple DOI) sont-ils attribués aux jeux de données et/ou aux fichiers composant les jeux de données ?

- L'entrepôt permet-il de documenter les données avec des métadonnées (auteurs, description du contenu du jeu de données, publications associées, etc.) et des informations permettant



(Source : RDC DRC)

(Source : coretrustseal.org)

de mieux comprendre et utiliser les données (définition des variables, logiciels associés, provenance, etc.) ? Les métadonnées (et idéalement les données) sont-elles indexées pour permettre leur recherche ?

-L'entrepôt permet-il de mentionner clairement la licence (licence ouverte, CC BY, etc.) ou les conditions spécifiques sous lesquelles les données sont utilisables ?

-L'entrepôt rend-il accessibles publiquement les citations et les métadonnées sont-elles toujours accessibles, même dans le cas de jeux de données dont les fichiers associés sont à accès restreint ?

-L'entrepôt utilise-t-il des métadonnées compatibles avec des standards de métadonnées reconnus ?

-L'entrepôt dispose-t-il d'un plan de préservation à long terme des données ?

Quels sont les Entrepôts de données « dignes de confiance »?

Actuellement, 107 entrepôts de données [11] sont enregistrés dans le registre Re3Data et sont certifiés par le CoreTrustSeal, répondant aux 16 exigences reflétant les caractéristiques que l'on peut conférer à des entrepôts fiables.

En France, deux entrepôts de données adoptent les principes TRUST et sont considérés « dignes de confiance » : le Centre de Données astronomiques de Strasbourg [12], un centre de données voué à la collecte et à la distribution dans le monde entier de données astronomiques hébergeant la base de référence mondiale pour l'identification d'objets astronomiques et le IFREMER-SISMER Portail de données marines [13] contribuant à la gestion de données des programmes de l'Ifremer et en particulier au programme centre de données océanographiques de l'Ifremer et aux programmes d'océanographie opérationnelle.

"De nombreux entrepôts n'ont pas de certification mais sont cependant largement reconnus par la communauté scientifique et offrent des garanties de conservation à long terme. Ces entrepôts sont d'ailleurs souvent recommandés voire imposés par les éditeurs." [14] Par exemple, NAKALA : entrepôt national français spécialisé en Sciences Humaines et Sociales ; SEANOE : entrepôt national français en sciences marines ; PANGAEA spécialisé en sciences de la Terre et de l'environnement ; le Réseau Quetelet : entrepôt national français en sciences sociales ; ORTOLANG : entrepôt national français spécialisé en sciences du langage ; TreeBASE (phylogénétique): données sous-jacentes aux publications.

La majorité des exigences de CoreTrustSeal se réfèrent (indirectement) aux principes FAIR des infrastructures [15]. Nous devons partager nos données afin de rendre la science ouverte réelle. Les principes FAIR nous aident à définir une gestion des données de recherche de haute qualité et transparente dans la gestion des données de recherche. Les mécanismes de certification, comme CoreTrustSeal pour les entrepôts de données, nous aident à créer le principe TRUST dans l'infrastructure de données de recherche dont nous avons besoin pour faire de la science une réalité. Les infrastructures de données de recherche sont ce dont nous avons besoin pour sauvegarder l'accessibilité et la durabilité de nos données FAIR.

« Les données de recherche ne deviendront ni ne resteront FAIR par magie. Nous avons besoin de personnes compétentes, de processus transparents, de technologies interopérables et d'une collaboration pour construire, exploiter et maintenir des infrastructures de la recherche »

Mari Kleemola, membre du RDA.

▪ **Selin GUDER**

Références

[1] *Ouvrir la Science. Les principes « TRUST » des entrepôts de données. [en ligne] URL : https://www.ouvrirlascience.fr/les-principes-trust-des-entrepots-de-donnees/ (Consulté le 28 mars 2021)*
[2] *Ibid.*
[3] *Ibid*.
[4] *Ibid*.
[5] *Ibid*
[6] *Ibid*
[7] *CoreTrustSeal. (s. d.). CoreTrustSeal. [en ligne]. URL : https://www.coretrustseal.org/ (Consulté le 13 juin 2021)*
[8] *Leeuw, L. de (DANS) (2019). Data Seal of Approval (DSA). DANS. https://doi.org/10.17026/dans-28z-njxq*
[9] *World Data System (WDS). International Science Council. (s. d.). [en ligne]. URL : https://council.science/what-we-do/affiliated-bodies/world-data-system-wds/ (Consulté le 13 juin 2021)*
[10] *CoreTrustSeal. (s. d.). CoreTrustSeal. [en ligne]. URL : https://www.coretrustseal.org/about/ (Consulté le 13 juin 2021)*
[11] *Re3data.org. (s. d.). [en ligne] URL: https://www.re3data.org/ (Consulté le 28 mars 2021)*
[12] *CDS. Centre de données astronomiques de Strasbourg. (s. d.). [en ligne] URL : https://cds.u-strasbg.fr/ (Consulté le 13 juin 2021)*
[13] *SISMER. Portail des données marines. (s. d.).[en ligne] URL : http://data.ifremer.fr/SISMER (Consulté le 13 juin 2021)*
[14] *La minute entrepôt. DoRANum. [en ligne] URL : https://doranum.fr/depot-entrepots/minute/https://doi.org/10.1038/s41597-020-0486-7 (Consulté le 28 mars 2021)*
[15] *Pôle ODATIS. Généralités. [en ligne] URL : https://www.odatis-ocean.fr/donnees-et-services/principes-de-gestion-des-donnees/generalites (Consulté 28 mars 2021)*

Bibliographie

CoreTrustSeal+FAIR : Statement of Cooperation & Support. (2020, octobre 27). CoreTrustSeal. https://www.coretrustseal.org/why-certification/coretrustsealfair-statement-of-cooperation-support/

CoreTrustSeal : Criteres de conformite. (2019, septembre 7). RDA. https://www.rd-alliance.org/coretrustseal-criteres-de-conformite

CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022 | Zenodo. Consulté 28 mars 2021, à l'adresse https://zenodo.org/record/3638211#.YGDPN68zbIV

Datapartage—Les entrepôts FAIR. Consulté 28 mars 2021, à l'adresse https://datapartage.inrae.fr/Produire-des-donnees-FAIR/Comment-FAIR-en-pratique/Les-entrepots-FAIR

Mini-symposium sur les principes TRUST : L'avenir des dépôts de données numériques | Research Data Canada. (2020, juillet 29). Research Data Canada | Avec Le Concours de Ses Intervenants, Données de Recherche Canada Veille à Ce Que Les Données Scientifiques Engendrent Des Innovations Dont Profitera Chaque Canadien. https://www.rdc-drc.ca/fr/mini-symposium-sur-les-principes-trust-lavenir-des-depots-de-donnees-numeriques/

Paillassard, P. Certification des entrepôts de données – DoRANum. Consulté 28 mars 2021, à l'adresse https://doranum.fr/2019/02/05/certification-des-entrepots-de-donnees/

À lire

« The TRUST Principles for digital repositories » consulter le site : https://www.nature.com/articles/s41597-020-0486-7

Consulter le site : https://www.coretrustseal.org/

Dispositifs techniques d'anonymisation et pseudonymisation des données de santé.

L'explosion du numérique et de l'ouverture des données a contribué à mettre en place plusieurs réglementations et obligations quant à la gestion des données sensibles et/ou à caractère personnel. les données de santé sont particulièrement impactées, car bien qu'indispensables, la collecte et le stockage de ces données demandent une gestion toute particulière. pour protéger ces données dites « sensibles », deux solutions : la pseudonymisation, ou plus sûr encore : l'anonymisation de celles-ci. Tour d'horizon des dispositifs permettant la mise en place de ces solutions.

Laurent TRIPOLI

La particularité des données de santé

Les données de santé (DDS) sont toutes les données médicales et/ou qui portent de manière générale sur la santé. Ces données peuvent concerner la santé d'un individu en particulier, d'un groupe de personnes, voire même d’une population dans son ensemble. Ces données peuvent être utilisées à plusieurs fins (suivi et évaluation des systèmes et politiques de santé, budgets prévisionnels, projections, statistiques, etc.).

Les données sensibles que représentent les données personnelles de santé sont protégées par le Règlement Général sur la Protection des Données (RGPD). La Commission nationale de l'informatique et des libertés (CNIL) quant à elle définit les données personnelles de la santé de la façon suivante : "Les données à caractère personnel concernant la santé sont les données relatives à la santé physique ou mentale, passée, présente ou future, d’une personne physique (y compris la prestation de services de soins de santé) qui révèlent des informations sur l’état de santé de cette personne. " (CNIL). On comprend ainsi que du fait de leur caractère particulièrement personnel et intime, ces données, si elles sont partagées ou stockées, doivent l’être dans un cadre très réglementé.
Effectivement la particularité des données de santé est qu’elle recouvre à la fois des données personnelles et des données sensibles.

« Dans le domaine des données de santé, on est souvent confronté au problème de devoir faire un choix entre la sécurité et le partage. Soit on sécurise les données, au détriment de leur qualité, ce qui les rend moins efficientes. Soit elles restent telles quelles mais on compromet leur confidentialité » explique Olivier Breillacq, dirigeant de WeData (Mathoux, 2020).

Car en effet, dans un monde tourné de plus en plus vers la science ouverte et l’*Open Data*, les données de santé ne restent plus forcément stockées par l’organisme qui les a récoltées. À des fins utiles (pour la société, les citoyens, les organisations), elles peuvent se retrouver diffusées et partagées bien au-delà, quelquefois même à l’international.

La pseudonymisation

La première des solutions afin de ne plus lier visiblement l’individu à ses données personnelles est le processus de pseudonymisation.

L’article 4 du RGPD définit la pseudonymisation ainsi : « […] on entend par pseudonymisation : le traitement de données à caractère personnel de telle façon que celles-ci ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires, pour autant que ces informations supplémentaires soient conservées séparément et soumises à des mesures techniques et organisationnelles afin de garantir que les données à caractère personnel ne sont pas attribuées à une personne physique identifiée ou identifiable. » (Galichet, 2017).

La pseudonymisation consiste donc à supprimer les champs des données directement identifiants, et à les remplacer par un nouveau champ, un nouveau type de données. Par exemple, le nom et le prénom seront remplacés par un numéro, un identifiant. Ce processus doit rendre impossible le lien entre le pseudonyme et l’identification réelle de l’individu. La pseudonymisation est souvent mise en place au travers d’une fonction de hachage. Cette dernière, dérivée de l’anglais "*hash function*" (traduction de pagaille, désordre, recouper et mélanger), désigne le fait qu’à partir d'une donnée fournie en entrée, on calcule une empreinte numérique servant à identifier rapidement la donnée initiale, au même titre qu'une signature. C’est le cas par exemple d’un identifiant numérique (numéro de passeport, numéro de sécurité sociale, etc.)...

Le principal avantage de la pseudonymisation est que, du fait du remplacement de la donnée confidentielle par un pseudonyme, le traitement des données (non personnelles) peut se faire à l’identique, comme avec une base de données non anonymisée.

Cependant, la pseudonymisation n’est pas reconnue comme un moyen efficace de « dé-identification », car elle ne donne pas un niveau de protection suffisamment élevé. En effet, il est toujours techniquement possible de réattribuer les identifiants pseudonymes à des personnes physiques. Par exemple, des séquences temporelles de positions géographiques (successions de visites médicales dans tel établissement par exemple) constituent très vite une trace unique. Ainsi, bien que cela soit interdit, à partir d’informations connues sur un individu on peut alors rapidement le ré-identifier.

Dans certains cas, la police peut par exemple retrouver un individu à partir de l’historique des coordonnées GPS du téléphone de ce dernier.

L'anonymisation

L'autre solution la plus efficace, pour pouvoir à la fois partager les données personnelles tout en respectant le caractère confidentiel et sensible de celles-ci, est l'anonymisation.
Pour la CNIL, l'anonymisation est "un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible." (CNIL, 2019).

Euris group explique sur sa page internet que : « Pour profiter des avantages qu'offrent l'accroissement et l'ouverture des données, les solutions d'anonymisation sont aujourd'hui essentielles pour adopter une gouvernance qui respecte la confidentialité et la protection des données de santé. » (EURIS).

L'anonymisation des données de santé est un processus complexe qui vise à préserver les données de tous les risques « raisonnables » de ré-identification d'un individu. Pour ce faire, l'anonymisation supprime l'association entre l'ensemble de données l'identifiant et le sujet des données. L'objectif étant, bien-sûr, qu'une fois le processus d'anonymisation réalisé, il ne soit plus possible d'associer les données récoltées avec un individu en particulier. L'individu n'est plus identifiable, mais l'utilité des données est préservée. De ce fait, l'anonymisation est un processus irréversible. C'est en cela que l'anonymisation diffère de la pseudonymisation, le remplacement d'un nom par un pseudonyme. Le processus technique d'anonymisation est assez compliqué, et son principal risque est qu'il soit mal réalisé, et qu'il puisse alors y avoir divulgation de données sensibles et confidentielles.

Des techniques d'anonymisation variées

Il existe plusieurs outils et dispositifs d'anonymisation. La professeure Latanya Sweeney de l'université de Harvard aux États-Unis a proposé une méthode proche de la pseudonymisation, mais plus efficace : la **k-anonymisation.**

La k-anonymisation peut être comparée à un processus permettant de flouter. Il s'agit de réduire les détails des données sensibles. Ainsi, des individus d'un même groupe auront la même valeur donnée, c'est ce que Sweeney appelle le « quasi-identifiant ». Il s'agit de généraliser: par exemple, toutes les personnes d'un certain âge, (18 à 32 ans, 32 à 45 ans, etc.) auront une valeur donnée identique. La précision de la donnée à caractère personnel est ainsi réduite. Il est rendu impossible avec ce processus de relier précisément un individu à une donnée… du moins en partie (Nguyen, 2014).

Effectivement il peut s'avérer possible, à partir du moment où l'on connaît le « quasi identifiant » d'un individu, par exemple son âge si l'on garde l'exemple précédent, d'exclure plusieurs données ou valeurs. Il est de même tout à fait envisageable d'établir une probabilité qu'un individu ait telle ou telle valeur pour telle ou telle donnée, d'autant plus si tous les individus d'un même groupe possèdent la même donnée sensible. Dans ce processus la modification des attributs du jeu de données est contraignante surtout lorsque l'on traite des données sensibles comme le sont celles de la santé.

Tandis que la technique dite de la "**l-diversité**" ne fait que "flouter" un peu plus la technique de la k-anonymisation, celle de la « **t-proximité** » va un peu plus loin. Sans doute trop loin même, puisque toutes les données considérées comme sensibles sont fractionnées en parts égales, de sorte que toutes les partitions se ressemblent en termes de distribution. Cela présente donc un très gros désavantage, surtout dans le domaine de la santé, puisqu'il devient impossible d'établir des statistiques pertinentes. Par exemple, le fait de savoir si telle maladie touche plus les personnes âgées, les plus jeunes, une certaine classe d'âge, etc. (Nguyen, 2014).

Dans ces trois cas de **généralisation** (famille de techniques donnée à ce genre de dispositifs), le jeu de données est trop large, et perd en pertinence.

L'autre famille de techniques est celle de la **randomisation**. Il s'agit ici d'altérer la précision des informations d'un individu. Par exemple, si une personne est mesurée au centimètre près, on peut lui attribuer au final une donnée de + ou – 10 cm. Cette modification de l'attribut de la donnée ayant pour objectif de réduire sa précision est appelée « **l'ajout du bruit** ». Le niveau de bruit que l'on souhaite est proportionnel au caractère personnel et privé de la donnée affectée. Cependant, attention à ne pas ajouter trop de bruit, ce qui entraînerait une baisse de la pertinence de la donnée récoltée.

Bien sûr, pour accroître la difficulté d'identification de l'individu, il est tout à fait possible d'allier cette technique à celle de la généralisation.

Cette technique de randomisation est par exemple mise en avant par l'entreprise Wedata qui a été approuvée récemment par la CNIL. Sur son site internet, la société

(Source : Pixabay)

française explique : « chacune de ces données est légèrement modifiée. Elles deviennent assez différentes des données de base pour garantir l'anonymat mais quand même assez proches des données de base afin de ne pas perdre leur valeur. L'idée reste de garder la granularité du jeu de données, avec toujours le même nombre de patients. Cette technologie permet d'assurer qu'on va garder les corrélations entre les individus et les distributions sur une variable. Lorsqu'une variable est modifiée, toutes les autres doivent être impactées. Admettons qu'il existe une corrélation entre la taille et le poids dans un jeu de données, alors il faudra modifier les deux afin de rester statistiquement pertinents. » (Mathoux, 2020).

C'est également cette technique que semble proposer l'offre « Cloud Santé® Anonymisation » d'Euris Group qui propose pour sa part de « transformer les données et ajuster la stratégie d'anonymisation pour obtenir la valeur analytique optimale des données dé-identifiées, en fonction du contexte d'utilisation. » (EURIS).

Une autre technique de randomisation est la **permutation**. Ce processus vise à garder les attributs exacts de chaque donnée, mais à les donner de manière aléatoire à d'autres individus. La corrélation entre les données et les individus n'existe donc plus, mais l'exactitude de toutes les données en tant que telles en revanche reste.

Cependant, tout comme pour l'ajout du bruit, la pertinence des attributs pour des données médicales doit être bien réfléchie. En effet, si par exemple les différents attributs d'un champ de données sont très liés entre eux, la pertinence du processus de permutation n'est plus. Ainsi, si l'on prend pour exemple, comme champs d'attributs, « motifs d'hospitalisation/ symptômes/service concerné », on se rend bien compte alors que les 3 champs sont particulièrement liés les uns aux autres. Une "dé-identification" n'est plus forcément garantie. C'est pourquoi, tout comme pour l'ajout du bruit, cette technique ne garantit pas en elle-même l'anonymat et doit souvent être combinée avec d'autres techniques.

Supprimer les données à caractère personnel ?

Et si la méthode la plus simple était, tout simplement, de supprimer les données à caractère personnel et/ ou sensible ? Sans les données identifiantes, alors a priori il n'y a aucun moyen d'identification. C'est vrai, mais cela soulève un autre problème. Supprimer ces données peut entraîner la suppression de données qui peuvent, elles, s'avérer utiles. Aussi, cela enlève évidemment le caractère personnel de la donnée, ce qui d'un point de vue statistique ou d'étude peut également être problématique. Dans le domaine de la santé la suppression de données à caractère personnel n'est bien évidemment pas envisageable.

Alors, comment anonymiser des données sensibles de manière plus fiable, et garantir ainsi la sécurité de la vie privée des individus en procédant à une réelle « **dé-identification** » ?

Nous avons vu que finalement aucune anonymisation n'emploie une méthode

véritablement parfaite et fiable. Elles ont toutes leurs avantages et leurs inconvénients.

La CNIL et les autorités de protection des données européennes ont défini trois critères qui permettent de s'assurer qu'un jeu de données est véritablement anonyme :

- « **L'individualisation** : il ne doit pas être possible d'isoler un individu dans le jeu de données »
- « **La corrélation** : il ne doit pas être possible de relier entre eux des ensembles de données distincts concernant un même individu »
- « **L'inférence** : il ne doit pas être possible de déduire, de façon quasi certaine, de nouvelles informations sur un individu. »

La CNIL ajoute que si les trois critères ne sont pas remplis, le responsable de traitement souhaitant anonymiser son jeu de données devra démontrer, par une évaluation des risques d'identification, que le risque de ré-identification est nul (CNIL, 2020).

L'importance de l'anonymisation dans le domaine de la santé, dans lequel la garantie de la vie privée est primordiale, n'est plus à démontrer. Aussi, de son côté, la CNIL rappelle également que, pour tout processus d'anonymisation, « si un jeu de données publié en ligne comme « anonyme » contient en réalité des données personnelles et qu'aucune des exceptions mentionnées à l'article L.312-1-2 du Code des relations entre le public et l'administration (CRPA) n'est applicable, cela peut être considéré comme une violation de données. Il est alors nécessaire de :

- Procéder au retrait du jeu de données en question dans les plus brefs délais
- En informer la CNIL si cette violation est susceptible d'engendrer un risque pour les droits et libertés des personnes ;
- D'en informer les personnes concernées si ce risque est élevé. » (CNIL, 2020)

C'est ainsi que pour certains chercheurs la protection des données sensibles et/ou à caractère personnel est peut-être à imaginer dans le « cloud ». Un stockage des données décentralisé, directement géré par les individus concernés serait pour eux une solution intéressante. L'individu est finalement le cœur du « problème », étant donné qu'aucune anonymisation fiable à 100% n'existe ; l'important est que la personne soit bien informée sur le modèle utilisé, la manière dont sont gérées ses données personnelles, et puisse à tout moment du processus, de la collecte au stockage, en passant par l'utilisation, avoir un droit de regard et de retrait de ses données.

▪ **Laurent TRIPOLI**

(Source : Pixabay)

Bibliographie

Agence esanté Luxembourg. Service de pseudonymisation en santé (sps). https://www.esante.lu/portal/fr/je-m-informe/services-pour-professionnels-de-sante-188-205.html

Ben Fredj, F. (2018). Méthode et outil d'anonymisation des données sensibles. [Thèse de doctorat, Université de Sfax]. https://tel.archives-ouvertes.fr/tel-01783967/document

CNIL. (2019, 17 octobre). L'anonymisation des données, un traitement clé pour l'open data. https://www.cnil.fr/fr/lanonymisation-des-donnees-un-traitement-cle-pour-lopen-data

CNIL. Quelles formalités pour les traitements de données de santé à caractère personnel ? https://www.cnil.fr/fr/quelles-formalites-pour-les-traitements-de-donnees-de-sante-caractere-personnel

CNIL. Santé. https://www.cnil.fr/fr/sante

CNIL. (19 mai 2020). L'anonymisation de données personnelles. https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

CNIL. (10 avril 2014). Avis 05/2014 sur les Techniques d'anonymisation. https://www.cnil.fr/sites/default/files/atoms/files/wp216_fr.pdf

Documentation du SNDS. Pseudonymisation. (16 décembre 2019) https://documentation-snds.health-data-hub.fr/glossaire/pseudonymisation.html#interet

EURIS. Anonymisation. https://www.euris.com/fr/cloud-sante-marketplace/anonymisation/

French Tech Central. (2020, 3 juin). Webinar : data et santé les bénéfices de l'anonymisation et de la pseudonymisation CNIL [vidéo]. YouTube. https://www.youtube.com/watch?v=70EOF4xvhXE

Galichet, C. (2017, 17 octobre). Données personnelles : anonymisation ou pseudonymisation ? Village de la justice. https://www.village-justice.com/articles/donnees-personnelles-anonymisation-pseudonymisation,26194.html

Groupe de travail « article 29 » sur la protection des données. Avis 05/2014 sur les Techniques d'anonymisation. https://www.dataprotection.ro/servlet/ViewDocument?id=1288

Mathoux, J. (2020, 5 novembre). La Cnil approuve WeData pour l'anonymisation des données de santé. Usine Digitale. https://www.usine-digitale.fr/article/la-cnil-approuve-wedata-pour-l-anonymisation-des-donnees-de-sante.N1024644

Matusek, F. Anonymisation statique vs dynamique. https://ressources.genetec.com/blog/anonymisation-statique-vs-dynamique

Nguyen, B. (2014, décembre). Techniques d'anonymisation. Statistique et société. https://www.benjamin-nguyen.fr/papers/ss.pdf

Oracle. Comment anonymiser des données ? https://www.oracle.com/fr/cloud/comment-anonymiser-donnees.html

Piquette-Muramatsu, S. (2021, 3 février). Anonymisation/Pseudonymisation. Université de Lille. https://pod.univ-lille.fr/ethique-et-protection-des-donnees-en-recherche/video/16591-s17-anonymisation-pseudonymisation/

Sham. Les données de la santé. https://www.sham.fr/blog/nouveaux-risques/la-donnee-de-sante

Wikipedia. Données de santé. https://fr.wikipedia.org/wiki/Donn%C3%A9es_de_sant%C3%A9

Wikipedia. Fonction de hachage. https://fr.wikipedia.org/wiki/Fonction_de_hachage

L'importance des protocoles de communication (HTTP, FTP) POUR LE RESPECT DES PRINCIPES FAIR

Guillaine POON

Comment accède-t-on à des données dans l'espace numérique qu'est le web ? Il est nécessaire de faire usage d'un ensemble de règles formelles décrivant comment transmettre et échanger des données, en particulier à travers un réseau. Ces règles sont aussi connues sous le nom de protocoles de communication. On utilisera un protocole spécifique selon l'action à faire. Les protocoles standards sont ceux les plus usités et codifiés comme tels. Seront décrits, les protocoles HTTP (HyperText Transfert Protocol) et FTP (File Transfer Protocol) car largement répandus, libres, gratuits, ouverts et libellés comme standards. De par leur nature, les protocoles respectent ainsi le A pour Accessibilité, des principes FAIR, à savoir permettre l'accès aux données et leur téléchargement.

HTTP et FTP dans un modèle de couches [1]

Les protocoles communiquent les uns avec les autres par une modélisation sous forme de couches superposées l'une sur l'autre, chaque couche dépendant de celles étant géographiquement plus bas suivant un axe vertical.

Le modèle TCP/IP (*Transmission Control Protocol/Internet Protocol*) est une suite de protocoles représentant l'ensemble des règles de communication sur Internet. Il peut être divisé en quatre couches principales :

- La couche Réseau (équivalent de "*link*" en anglais sur le schéma ci-dessus) assure la connexion des liaisons physiques entre matériels (entre un ordinateur et un routeur par exemple).
- La couche Internet assure le routage entre les réseaux.
- La couche Transport assure le routage entre deux appareils.
- La couche Application est utilisée pour la communication entre les applications de deux ordinateurs.

Plus la couche est située en haut de pile, plus elle assure la communication entre deux entités abstraites. Dans le cas de nos données, les protocoles HTTP et FTP sont utilisés dans la couche Application. Cela est cohérent car c'est à ce niveau-là que les données peuvent transiter. Dans tous les cas, il est nécessaire que ces protocoles soient standards, c'est-à-dire ouverts, libres et d'usage universel pour pouvoir procéder à la récupération des données échangées.

Le rôle de HTTP [2]

HTTP est un protocole client-serveur qui permet la consultation de sites Web. Il va chercher des ressources qui peuvent être des documents HTML (*HyperText Markup Language*), des vidéos ou des images pour ensuite les afficher selon une certaine mise en page spécifiée en HTML. Il peut aussi envoyer du contenu à stocker sur des serveurs (via des formulaires par exemple) ou permettre la mise à jour d'une page web en allant chercher certaines ressources sur un serveur pour actualiser une partie de la page.
Par exemple, un client (généralement un navigateur web) souhaite accéder à une page web, une requête est faite à un serveur Web qui répond en fournissant la page demandée.

HTTP peut aussi être utilisé pour échanger des données via des API (*Application Programming Interfaces*) web. Ce sont des solutions logicielles qui permettent aux applications informatiques de communiquer entre elles et donc de partager et d'accéder à des données lisibles par machine. Ces API peuvent autoriser un partage d'une partie uniquement de leurs données dans le cas où cela serait pertinent (données sensibles ou volume trop important non nécessaire pour la finalité recherchée). De nombreuses API bien documentées concernant l'échange de données et de métadonnées sont dénombrées.

On citera OGC WMS pour les images de cartes géographiques. Dans un but principalement d'alimentation et d'augmentation de la visibilité de certaines ressources dans des portails numériques, les échanges de métadonnées entre entrepôts de données sont possibles via OAI-PMH [3] (*Open Archives Initiative – Protocol for Metadata Harvesting*). Plus exactement, les institutions souhaitant intégrer les données des entrepôts – on parlera de moissonnage de données – à leurs bibliothèques numériques, font une requête au serveur hébergeant ces données et les reçoivent sous format XML.

Le rôle de FTP [4]

FTP permet le partage de fichiers entre un client et un serveur sur un réseau. Il s'agit aussi d'un protocole client-serveur. Il permet à un client de téléverser un fichier sur un serveur ou de télécharger un fichier depuis un serveur avec l'ouverture d'une session entre le client et le serveur pendant une durée fixée. Par exemple, un client souhaite stocker un fichier de son ordinateur sur un serveur web, le protocole FTP va permettre la communication entre ces deux machines et allouer un espace de stockage au serveur pour qu'il puisse héberger le contenu de l'utilisateur.

L'importance de ces protocoles

Les protocoles HTTP et FTP permettant la communication entre un navigateur et un serveur, sont des éléments essentiels pour l'échange de données sur le web. Le fait que leur dénomination s'affiche dans l'URL (*Uniform Resource Locator*), qu'une connexion de données soit établie entre le client et le serveur et que les deux protocoles servent au transfert de fichiers, sont des caractéristiques qui les rassemblent.

On différenciera HTTP de FTP dans le sens où FTP est efficace pour le transfert de fichiers volumineux alors que HTTP est plus usité pour

l'affichage dynamique et rapide de sites web et donc est pertinent pour des transferts de fichiers de petite taille. Cette spécificité de FTP se modélise avec une connexion de données ajoutée à celle de contrôle ainsi que la nécessité d'une authentification par mot de passe pour le transfert de fichiers.

Ces dispositifs sont cruciaux pour le respect des principes FAIR [5], les données sont accessibles dans le sens de récupérables via HTTP et déposables via FTP. Ces protocoles doivent donc être implémentés dans les entrepôts de données pour assurer l'ouverture, l'échange et le partage des données. Les deux protocoles présentés sont libres, gratuits et ouverts, cela signifie qu'ils sont libres d'utilisation, libres de propriété et interopérables avec d'autres logiciels. L'interopérabilité n'est pas une caractéristique négligeable car cela signifie la mise à disposition d'une documentation open source constamment enrichie et mise à jour par une grande communauté ainsi qu'une compatibilité possible avec de futurs outils.

L'aspect sécurité, indispensable pour le partage des données

Bien que certaines données issues de la recherche ne nécessitent pas d'authentification pour y accéder, ceci n'est pas généralisable à toutes les données.

En effet, une authentification est parfois nécessaire pour accéder aux données : ***"As open as possible, as closed as necessary".*** Le A des principes FAIR n'est pas synonyme d'ouverture complète des données. L'accessibilité est à prendre au sens de communiquer les conditions exactes dans lesquelles les données sont accessibles |6].

Par exemple, celles considérées comme sensibles [7], doivent être partagées avec des restrictions qui s'y appliquent, comme celles qui suivent :

- données à caractère personnel ;
- données relevant de la sécurité nationale ;
- données sujettes à un dépôt de brevet.

Dans ce cas, les règles d'accès aux données et métadonnées doivent être clairement explicitées. Quand l'autorisation est accordée, un processus d'authentification doit exister pour qu'un humain ou une machine puisse accéder aux données de façon sécurisée.

Ce processus préalable de vérification se fait grâce à des variantes des protocoles standards avec la suffixation d'un "S" pour "*Secure*" à l'acronyme de base. Cet ajout signifie que l'aspect Sécurité est pris en compte dans ces nouveaux protocoles, l'ancien protocole est complété par un protocole de sécurisation, soit SSL (*Secure Sockets Layers*) et plus récemment TLS (*Transport Layer Security*).

Dans le cas de HTTPS [8] par exemple, le visiteur peut vérifier à l'aide d'un certificat d'authentification validé par une autorité fiable, l'identité d'un site web et de même le site peut vérifier l'identité du visiteur si celui-ci possède un certificat d'authentification client. Ceci permet d'une part une sécurisation de l'intégrité des données et d'autre part une confidentialité des données par restriction des droits accordés selon l'identité de l'utilisateur.

Un autre cas nécessitant l'authentification serait l'accès pour gérer les données (dépôts et gestion de dépôts).

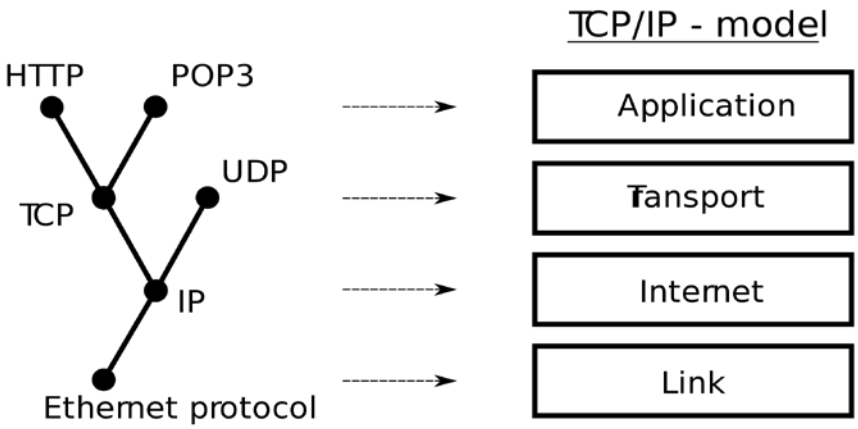
Généralement, une authentification par pseudonyme et mot de passe avec HTTP est à faire pour le propriétaire souhaitant déposer des données, impliquant la création préalable d'un compte sur le site web. Celui-ci peut également imposer une authentification à des contributeurs par exemple.

Cela permet d'identifier le propriétaire des jeux de données et celui-ci peut spécifier, si besoin, les droits d'accès à ces données. Quant aux API, la connexion est faite de façon sécurisée via un système de clé API, une clé formée aléatoirement de caractères et de chiffres, l'équivalent d'un mot de passe.

Les données et métadonnées doivent être récupérables par le biais de protocoles de communication standards libres et ouverts selon les principes FAIR. Le bon fonctionnement du partage de données sur les entrepôts de données dépend fortement des protocoles de communication implémentés. Chaque protocole a un rôle bien attribué, HTTP permet d'afficher correctement le contenu demandé d'un site web ainsi et d'accéder à des API, interfaces d'échanges de données entre applications informatiques. FTP permet quant à lui, le dépôt et le téléchargement des données avec comme source ou destinataire l'entrepôt de données. Les deux protocoles utilisent un système d'authentification, pouvant être renforcé par des protocoles de sécurisation supplémentaires. Les protocoles utilisés par les entrepôts de données peuvent donc être un critère déterminant pour le choix final de l'entrepôt.

▪ **Guillaine POON**

(Source : Freepik)



(Source : Wikipédia by Jsoon eu)

(Source : Pixabay)

Références

[1] *Australian Research Data Commons (ARDC). (s. d.). Standardised communications protocols. ARDC. Consulté le 25 septembre 2021, à l'adresse https://ardc.edu.au/resources/standardised-communications-protocols/*

[2] *Mozilla Developer Network Web Docs. (2021, août 14). An overview of HTTP - HTTP | MDN. Disponible à l'adresse : https://developer.mozilla.org/en-US/docs/Web/HTTP/Overview*

[3] *BNF. (s. d.). Protocole OAI-PMH. BnF - Site institutionnel. Consulté le 25 septembre 2021, à l'adresse : https://www.bnf.fr/fr/protocole-oai-pmh*

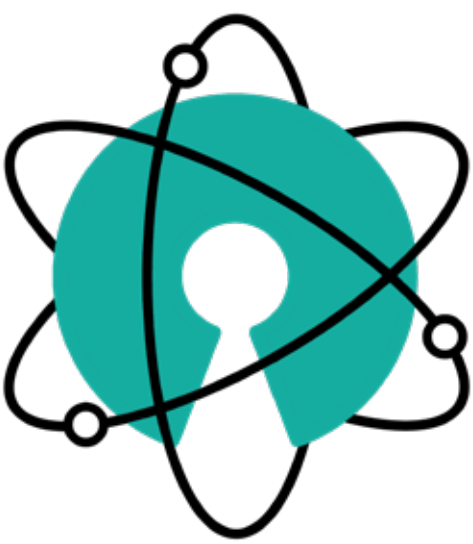
[4] *Speedcheck. (s. d.). FTP. Consulté le 27 mars 2021, à l'adresse : https://www.speedcheck.org/fr/wiki/ftp/*

[5] *Fair data. (2021). In : Wikipédia. Disponible à l'adresse : https://fr.wikipedia.org/w/index.php?title=Fair_data&oldid=186437479*

[6] *GO FAIR. (s. d.). A1.2 : The protocol allows for an authentication and authorisation where necessary. Consulté 25 le mars 2021, à l'adresse: https://www.go-fair.org/fair-principles/a1-2-protocol-allows-authentication-authorisation-required/*

[7] *DoRANum. (2019, décembre 4). Les principes FAIR. Disponible à l'adresse: https://doranum.fr/enjeux-benefices/principes-fair/*

[8] *HyperText Transfer Protocol Secure. (2021). In : Wikipédia. Disponible à l'adresse : https://fr.wikipedia.org/w/index.php?title=HyperText_Transfer_Protocol_Secure&oldid=184664024*



Consortiums : des groupes de travail dédiés au partage des données scientifiques

Pierre POUGET

Dans un contexte fortement marqué par l'ouverture des données scientifiques, différents groupes de travail appelés « consortiums » voient le jour et s'organisent afin de guider les chercheurs dans le partage de leurs données. Ces groupes, à la fois européens et internationaux, proposent différents outils et émettent des recommandations permettant de rendre les données « FAIR ». Cet article vise à établir un état des lieux des groupes de travail déjà existants et de se concentrer sur les différents outils et plateformes mis en place.

Des consortiums interdisciplinaires

Ces dernières années, plusieurs groupes interdisciplinaires ont été créés dans l'optique de faciliter le partage et la réutilisation de données.

Parthenos

C'est notamment le cas du projet européen Parthenos [1]. La mission principale de ce dernier est d'œuvrer à une certaine cohésion de la recherche dans des champs disciplinaires très variés : histoire, archéologie, linguistique….

Dans son plan d'action, Parthenos prévoit le développement de normes communes, la coordination d'activités conjointes, l'harmonisation de la définition et la mise en œuvre des politiques, la mise en commun des services et le partage de solutions aux mêmes problèmes. Concrètement, Parthenos s'engage à mettre en place des normes de données communes dans un contexte fortement marqué par l'interdisciplinarité de la recherche.

Research data alliance

La Research Data Alliance (RDA) [2] est une organisation internationale qui a vu le jour en 2013 en tant qu'initiative communautaire. Elle a pour objectif de permettre le partage et la réutilisation des données.
Elle regroupe des groupes de travail réunissant des experts issus du monde universitaire et du secteur privé. Ces différents groupes développent des infrastructures de données destinées à être utilisées par des communautés issues de domaines très variés (santé, agriculture, économie…).

Des consortiums centrés sur une discipline spécifique

D'autres groupes de travail se sont développés autour d'un domaine de recherche en particulier.

Cahier

C'est par exemple le cas de Cahier [3], dédié principalement aux corpus d'auteurs pour les Humanités.

Les corpus traités dans ce consortium sont le plus souvent associés à une activité éditoriale, intégralement numérique ou sous forme de deux supports (papier et en ligne). En outre, Cahier offre des formations aux nouveaux outils et méthodes, privilégie le partage d'expériences et veille au bon respect des règles concernant l'échange, l'interopérabilité et la réutilisation des données de recherche scientifique.
Pour ce dernier point, il a mis en place des normes concernant les corpus d'auteurs ainsi que différents guides et recommandations à destination des chercheurs.

Un groupe de travail spécifique a été constitué en 2016 : il s'agit de **Data_Cahier**. [4] Les principaux objectifs de ce groupe de travail reposent sur l'accompagnement des projets membres du consortium Cahier dans l'exposition de leurs données ainsi que sur l'organisation du stockage de leurs données dans l'optique de mettre en place un archivage de ces données à long terme.

CORLI

Un autre consortium relatif aux corpus existe : il s'agit de CORLI [5] (Corpus, Langues, Interactions). Ce dernier constitue une aide à la diffusion des corpus, des outils et des méthodes de travail, et d'exploration de ces corpus.

Organisé sous la forme d'un réseau de recherche, il permet le partage de méthodes, techniques et données de ses participants. Il promeut grandement l'utilisation de standards et d'outils ouverts et partageables. Pour cela, des journées de formation sont organisées dans lesquelles sont abordées différentes thématiques parmi lesquelles la gestion des métadonnées ou encore l'utilisation de logiciels pour l'annotation de corpus.

Sur son site internet, CORLI propose de nombreuses ressources : inventaire des outils, guides d'annotation, bonnes pratiques juridiques... [6]

MASA

Concernant le domaine de l'archéologie, le consortium MASA [7] (Mémoires des archéologues et des sites archéologiques) s'impose comme un incontournable.

Le traitement des données en archéologie est en effet primordial car les données de terrain constituent une mine précieuse d'information. De plus, les documents produits au cours des fouilles archéologiques sont le plus souvent très fragiles et le fait de les numériser facilite grandement leur consultation.
Pour faciliter ces démarches, le consortium MASA propose un accès unifié à des corpus variés de données et de documentations produites par les archéologues.
Il propose des outils destinés à la communauté archéologique, qui veillent au respect des standards internationaux. Parmi ces outils, on retrouve notamment **Opentheso** [8], conçu par la Maison de l'Orient et de la Méditerranée de Lyon. Il s'agit d'un gestionnaire de thésaurus multilingue et multi-hiérarchique. Il comporte de nombreuses fonctionnalités telles que la gestion des thésaurus collaborative avec quatre niveaux d'authentification ("superadmin", "admin", "manager" et "contributeur"), l'interopérabilité (génération automatique d'identifiants pérennes), ou encore l'inclusion d'un module d'alignement paramétrable qui permet l'alignement semi-automatique vers des thésaurus externes. Il est important de noter que l'import et l'export peuvent se réaliser sous différents formats (SKOS, Turtle, Json-LD et CSV).

ImaGEO

Le consortium ImaGEO [9] a été conçu pour répondre à un besoin concernant la mise à disposition de données historiques sur les infrastructures de données géographiques (IDG).
Ce consortium se donne ainsi pour mission de rendre accessibles, consultables et mobilisables des données cartographiques et photographiques. Pour simplifier la consultation des données, il a mis en place la base **Navigae** [10]. Cet outil permet en effet de rechercher et de visualiser les cartographies produites et de les réutiliser grâce à la mise en place de métadonnées Dublin Core.

3D SHS

En 2014, un consortium dédié aux pratiques de la 3D en Sciences Humaines et Sociales voit le jour : il s'agit de 3D SHS [11].
Créé pour accompagner l'émergence de nouvelles pratiques au sein des Sciences Humaines et Sociales, ce consortium permet aussi de rapprocher les différents acteurs du consortium 3D SHS. Il centre ainsi ses missions autour de la technologie 3D.
Parmi ces missions, on retrouve l'archivage des données qui est réalisé par l'intermédiaire du Conservatoire National des Données 3D.
Cet environnement, hébergé par l'infrastructure d'Huma-Num, assure une sauvegarde totalement sécurisée des données produites dans le cadre de projets de l'Enseignement Supérieur et de la Recherche en Sciences Humaines et Sociales. Les données qui seront sauvegardées dans le conservatoire devront impérativement être accompagnées de métadonnées décrivant le cadre scientifique du projet ainsi que les types de production 3D réalisées. Par ailleurs, si les données sont destinées à être déposées au CINES (Centre Informatique National de l'Enseignement Supérieur), il faudra être attentif aux différents formats de fichiers.
Pour guider les différents groupes projets dans l'archivage de leurs données 3D, le consortium 3D SHS a rédigé un guide de bonnes pratiques. [12]

(Source : NAVIGAE)

Références :

[1] *https://www.parthenos-project.eu/*
[2] *https://rd-alliance.org/*
[3] *https://cahier.hypotheses.org/*
[4] *https://cahier.hypotheses.org/activites/groupe-data_cahier*
[5] *https://corli.huma-num.fr/*
[6] *https://corli.huma-num.fr/bonnes-pratiques/*
[7] *https://masa.hypotheses.org/*
[8] *https://github.com/miledrousset/opentheso/releases*
[9] *https://imageo.hypotheses.org/*
[10] *https://www.navigae.fr/*
[11] *https://shs3d.hypotheses.org/*
[12] *https://hal.archives-ouvertes.fr/hal-01683842v4/document*

Bibliographie

À propos. (s. d.). Consortium ImaGEO. Consulté le 11 mai 2021. [En ligne] https://imageo.hypotheses.org/a-propos

About RDA. (2016, mars 22). RDA. https://www.rd-alliance.org/about-rda

Baudry, J. (2020). ImaGEO, un consortium au service du géographe. Arabesques, 98, 18-19. https://doi.org/10.35562/arabesques.1886

Bonnes pratiques – Consortium CORpus, Langues et Interactions. (s. d.). Consulté le 10 mai 2021. [En ligne] https://corli.huma-num.fr/bonnes-pratiques/

CND3D. (s. d.). Consulté le 10 mai 2021. [En ligne] https://3d.humanities.science/

Consortium CORpus, Langues et Interactions – CORpus, Langues, Interactions. (s. d.). Consulté le 10 mai 2021. [En ligne] https://corli.huma-num.fr/

FAIR Data Maturity Model : Specification and guidelines - draft. (2020, avril 10). RDA. https://www.rd-alliance.org/group/fair-data-maturity-model-wg/outcomes/fair-data-maturity-model-specification-and-guidelines

Galonnier, J., Le Courant, S., Pecqueux, A. & Noûs, C. (2019). Ouvrir les données de la recherche ? Tracés. Revue de Sciences humaines. ENS Éditions, Décembre 2019, no #19, p. 17-33. ISBN 9791036202278

GP1 – Interopérabilité / Pratique et outils d'exploration de corpus – Consortium CORpus, Langues et Interactions. (s. d.). Consulté le 10 mai 2021. [En ligne] https://corli.huma-num.fr/les-groupes-projets/gp1/

Groupe « Data_Cahier ». (s. d.). Consortium Cahier. Consulté le 10 mai 2021. [En ligne] https://cahier.hypotheses.org/activites/groupe-data_cahier

Idmhand, F., Galleron, I., (2020). Guide pour la FAIRisation des données des corpus d'auteurs préparé par Fatiha Idmhand et Ioana Galleron pour le [Groupe de travail Data_Cahier]. [Rapport de recherche]. Huma-Num. [En ligne] https://halshs.archives-ouvertes.fr/halshs-02889777/document

Le consortium. (s. d.). Consortium MASA. Consulté le 10 mai 2021. [En ligne] https://masa.hypotheses.org/le-consortium

Opentheso. (s. d.). Consortium MASA. Consulté le 10 mai 2021. [En ligne] https://masa.hypotheses.org/opentheso

Outils. (s. d.). Consortium MASA. Consulté le 10 mai 2021. [En ligne] https://masa.hypotheses.org/outils

PARTHENOS, Hollander, H., Morselli, F., Uiterwaal, F., Admiraal, F., Trippel, T., & Di Giorgio, S. (2019). PARTHENOS Recommandations pour FAIRiser vos données. https://doi.org/10.5281/zenodo.3463521

RDA for Disciplines. (2016, mai 26). RDA. https://www.rd-alliance.org/rda-disciplines

RDA Groups. (s. d.). RDA. Consulté le 10 mai 2021. [En ligne] https://www.rd-alliance.org/groups

Outils. Dans : Consortium MASA. (s. d.). Consulté le 10 mai 2021. [En ligne] https://masa.hypotheses.org/outils

RDA for Disciplines. Dans : RDA. (2016, mai 26). Consulté le 06 mai 2021. [En ligne] https://www.rd-alliance.org/rda-disciplines

RDA Groups. Dans : RDA. (s. d.). Consulté le 06 mai 2021. [En ligne] https://www.rd-alliance.org/groups

Research Data Alliance FAIR Data Maturity Model Working Group - 2020 - FAIR Data Maturity Model specification and guidel.pdf. (s. d.). Consulté le 10 mai 2021. [En ligne] https://halshs.archives-ouvertes.fr/halshs-02889777/document

DOSSIER : EXEMPLES D'APPLICATIONS EN SCIENCE OUVERTE

La FAIRisation des données de la recherche médicale

Zaina NZOGU LOZANO

Dans le secteur médical, la protection de l'information est un enjeu crucial puisque les chercheurs manipulent fréquemment des données sensibles. Par ailleurs, la pandémie actuelle a mis en lumière l'urgence à les partager et/ou à les réutiliser. Les principes FAIR, conçus principalement par des spécialistes des sciences du vivant [1], apparaissent donc comme particulièrement pertinents pour inscrire la recherche médicale dans le mouvement de la science ouverte.

Source : Pixabay

L'Union européenne (UE) rend désormais obligatoire la FAIRisation des données de recherches qu'elle finance. [2] La mise en œuvre de ce modèle repose sur des infrastructures techniques qui permettent de stocker des (méta)données, ainsi que sur l'implication des acteurs du terrain. Comme le soulignait en 2019 un groupe d'experts de la Commission européenne, l'un des défis majeurs est de faire évoluer les pratiques des chercheurs. [3] Ils ne pensent en effet pas forcément à réaliser un Plan de Gestion des Données (PGD) ou *Data Management Plan* (DMP) en amont à leur projet. En France, deux études récentes [4] ont par exemple mis en lumière un faible recours aux dispositifs institutionnels existants pour le stockage et l'archivage pérenne des données de la recherche. La réussite de l'implémentation des principes FAIR dépend également d'une prise en compte des spécificités des champs disciplinaires. Il existe en effet des différences en matière de méthodes de collecte et de partage des données, d'origine des financements, de modèles économiques éditoriaux ou d'importance accordée au facteur d'impact. Qu'en est-il du fonctionnement de la recherche médicale ?

Les caractéristiques de la recherche médicale

Des disciplines diverses

La recherche médicale concerne les disciplines liées à la santé humaine et animale. Si l'on se réfère au classement disciplinaire du Conseil national des universités (CNU) français, l'enseignement et la recherche dans ce domaine impliquent bien évidemment les spécialités médicales, mais également, par exemple, les sciences infirmières (section 92), les sciences physico-chimiques et l'ingénierie appliquée à la santé (section 85), les neurosciences (section 69) ou la biochimie et la biologie moléculaire (section 64). L'équivalent helvétique du CNRS, le Fonds national suisse de la recherche scientifique (FNS), a dressé une liste des disciplines représentées en son sein et distingue notamment la médecine expérimentale (pathophysiologie, immunologie, éthologie, etc.), clinique (chirurgie, neurologie, médecine vétérinaire, gériatrie, etc.), préventive (toxicomanie, accidents, maladies cardio-vasculaires et infectieuses, etc.) et sociale (diététique, problèmes médico-sociaux, statistiques, santé publique, etc.). En principe, une collaboration active entre chercheurs provenant de filières différentes est donc possible, mais elle dépend de l'organisation interne des structures et/ou de la nature des projets. Dans tous les cas, l'interdisciplinarité s'étend désormais aux sciences dites "dures", via l'application de méthodes issues des sciences sociales ou de la psychologie. La Bibliothèque nationale de médecine des Etats-Unis a d'ailleurs intégré dans l'arborescence de MeSH, son thésaurus de référence mondiale, des termes issus des sciences humaines et sociales qui permettent de rendre compte de cette approche lors de l'indexation [5].

La littérature scientifique évoque fréquemment la recherche biomédicale. À quoi se réfère-t-on exactement ? Il s'agit d'une catégorie transdisciplinaire qui, pour certains chercheurs, apparaît comme une "convergence entre biologie et médecine (…) rendue possible par la croissance de l'industrie pharmaceutique" [6] et, pour d'autres, le résultat des "transformations technico-scientifiques qui sont intervenues dans le domaine de la médecine" [7]. Selon l'article L11221-1 du code de la santé publique français, la recherche biomédicale

concerne « les recherches organisées et pratiquées sur l'être humain en vue du développement des connaissances biologiques ou médicales » [8].

Recherche fondamentale versus recherche clinique

Comme l'indique l'Inserm, la recherche médicale se divise en deux volets. La recherche fondamentale, d'une part, vise à produire des connaissances sur le fonctionnement des organismes vivants et s'appuie sur l'expérimentation. La recherche clinique, d'autre part, concerne les études menées sur l'être humain [9]. L'industrie, qui peut mobiliser plus de moyens et attend un retour sur investissement, finance essentiellement de la recherche clinique.

Le poids du secteur privé

Contrairement aux sciences humaines et sociales, le secteur médical compte un nombre important de financements provenant du secteur privé. En 2001, J. Martin revenait déjà sur les débats concernant les conflits d'intérêts qui traversent le domaine médical en raison des liens entre les chercheurs et les industriels. L'auteur faisait remarquer que de nouveaux traitements, servant à traiter des maladies répandues dans les pays riches, étaient régulièrement commercialisés tandis que les malades des pays pauvres n'avaient toujours pas accès à des traitements pour soigner des maladies les touchant particulièrement (VIH, paludisme, etc.) [10]. La crise sanitaire engendrée par le Covid-19, a mis en évidence pour l'opinion publique les problèmes éthiques liés à la présence d'industriels dans le secteur pharmaceutique. Les brevets, qui permettent aux "big pharma" de s'assurer des bénéfices considérables, mettent en péril le bien commun. Le PDG de la société Moderna détient d'ailleurs une fortune estimée à 3,5 milliards d'euros [11]. Au regard de ce contexte commercial, on peut donc s'interroger sur les

pratiques en matière de partage des données de la recherche médicale.

Pratiques éditoriales

Le marché de l'édition médicale génère des revenus colossaux, évalués à 11,9 billions de dollars en 2017 et dont la croissance est estimée à 4,6% en 2021 [12]. D'après une étude britannique de Jisc [13], le champ biomédical se caractérise par un taux de publication élevé, un recours fréquent à la co-écriture et la prépondérance des articles sur les monographies. L'enquête Couperin 2019 sur les pratiques des chercheurs français en matière de science ouverte révèle les différences disciplinaires et quelques spécificités des sciences du vivant et de la médecine (SVM). Tout d'abord, la langue dominante de communication dans les SVM est l'anglais [14]. Dans ce domaine, les chercheurs paient fréquemment des frais de publication dans des revues en *open access* (*Article Processing Charges*, APC), contrairement aux lettres et sciences humaines, où cette pratique est marginale [15]. Enfin, le facteur d'impact est un critère de choix fondamental pour plus de 90% des répondants issus des SVM [16]. Le dépôt de preprints ou d'articles dans des archives ouvertes peut être un indicateur du degré de maturation des disciplines vis-à-vis de la FAIRisation des données. On remarque que la pratique du dépôt dans des preprints est plus rare en médecine, pharmacologie et sciences du vivant [17]. En outre, seul un peu plus de 20% des répondants issus des SVM dépose régulièrement ses travaux dans une archive ouverte (taux le plus faible) [18].

Les dispositifs techniques de gestion des données de la recherche médicale

La FAIRisation des données de la recherche fait partie du deuxième axe du Plan national français pour la science ouverte de 2018. Dans ce cadre, le comité pour la science ouverte a récemment constitué deux groupes de travail pour la recherche en santé. Le premier concerne les plans de partage des données issues des essais cliniques et le second un projet de portail des études individuelles en santé [19]. La mise en œuvre du modèle FAIR repose justement sur une gestion réfléchie du cycle de vie de la donnée, ainsi que sur des infrastructures qui facilitent le partage et la réutilisation des (méta)données.

L'interopérabilité

Afin de faciliter le partage et la réutilisation des (méta)données, les entrepôts de données doivent intégrer des langages et syntaxes garantissant leur interopérabilité. En voici deux exemples.

- *Fast Healthcare Interoperability Resources* (FHIR)

Le FHIR est un standard dont la dernière version a été publiée

en 2019. Il permet l'échange des données médicales dans le respect de leur intégrité. Il repose sur la notion de ressource, un format devant permettre de la décrire (UML, XML, JSON). Des métadonnées lui sont associées. En outre, elle doit pouvoir être lisible par l'humain [20]. L'un des projets de la communauté vise justement à lier les modèles FHIR et FAIR [21].

- *Set of Common Data Elements for Rare Diseases Registration*

Ce CDE a été conçu par un groupe de travail de l'Union européenne. Il est composé d'une série de 16 éléments servant à décrire les données disponibles dans les dépôts dédiés aux maladies rares. Un code correspond à chaque élément. Parmi ceux-ci, on peut citer : le pseudonyme du patient, l'âge d'apparition des premiers symptômes, l'âge du diagnostic, le diagnostic génétique ou le consentement pour la réutilisation de ses données à des fins de recherche.

Les entrepôts

En tant que pays associé à l'Espace européen de la recherche (EER), la Suisse participe à de nombreux projets de l'UE. Le FNS a amorcé sa transition vers la FAIRisation des données dans le cadre de sa politique institutionnelle *Open Research Data* (ORD). La présentation d'un PGD est obligatoire pour toute demande de financement depuis 2017. Les coûts liés à la mise en œuvre du PGD sont pris en compte dans les financements, mais les chercheurs doivent sélectionner des dépôts qui répondent aux critères FAIR. Le FNS a donc dressé une liste d'entrepôts satisfaisants [22], ainsi qu'une check-list permettant d'en identifier d'autres qui respecteraient aussi ces critères. En effet, comme le note un dernier rapport du groupe d'experts sur les principes FAIR de l'EOSC, bien que ces principes ne soient souvent pas évoqués de manière explicite, certains dispositifs, de fait, les mettent en pratique [23]. Quels sont donc les critères du FNS ? Premièrement, un identifiant pérenne doit être attribué aux jeux de données. Deuxièmement, les dépôts doivent permettre le téléchargement des métadonnées, publiquement accessibles indépendamment des restrictions sur le texte intégral. Troisièmement, la licence doit être clairement stipulée. Quatrièmement, le dépôt doit spécifier dans quel format soumettre les données pour garantir leur interopérabilité. Cinquièmement, ce dépôt doit avoir un plan d'archivage pérenne.

Dans la liste d'entrepôts respectant les principes FAIR, on en retrouve justement pour les sciences du vivant.

- *ArreyExpress*

Cet entrepôt qui est en train de fusionner avec BioStudies [24] stocke des données issues de recherches expérimentales en génomique, pouvant être réutilisées par la communauté.

- GenBank

Cet entrepôt états-unien contient des séquences d'ADN.

- *PRoteomics IDEntifications Database* (PRIDE)

Le dépôt de données se fait grâce à un outil fourni par PRIDE.

- RCSB Protein Data Bank

Cet entrepôt contient des données biologiques sur les protéines et les acides nucléiques. Le dépôt s'effectue à l'aide de divers outils permettant de convertir aux bons format (PDB, PDBx/mmCIF, PDBML/XML, etc.) les données puis de les valider.

- *Sequence Read Archive* (SRA)

Il s'agit de la plus grande base de données publique donnant accès à des séquences d'ADN.

L'infrastructure ELIXIR

Ce dispositif coordonne et soutient des ressources bioinformatiques (bases de données, outils, logiciels, etc.) dans le domaine des sciences du vivant qui respectent les principes FAIR. ELIXIR met à la disposition de la communauté scientifique cinq plateformes : *Tools, Data,*

Source : Pixabay

Compute, Interoperability et *Training*. En définitive, cette infrastructure européenne de grande envergure aborde toutes les questions liées à la FAIRisation des données : standards, interopérabilité, reproductibilité, logiciels, *workflows*, stockage, formation, etc. Parmi les bases de données soutenues par ELIXIR, on retrouve ArrayExpress et PRIDE.

Le projet FAIR4Health

Ce projet transnational coordonné par le CHU de Séville mobilise 17 pays d'Europe. Financé par le programme Horizon 2020, il vise à implémenter les principes FAIR dans la recherche médicale. L'un des objectifs du projet est de développer une plateforme dédiée à la FAIRisation des données de la recherche [25]. Dans ce cadre, les participants ont créé des outils permettant aux chercheurs de transformer leurs données brutes en jeux de données respectant les principes FAIR. Ces outils intègrent les spécifications FHIR.

Conclusion

De nombreuses initiatives existent, essentiellement portées par des fonds européens, pour mettre en œuvre la FAIRisation des données issues de la recherche médicale. Les entrepôts actuels concernent particulièrement le domaine de la génétique. Systématiser l'application des principes FAIR permettra sur le long terme de faciliter le partage et la reproductibilité de données essentielles pour la recherche en santé publique. En outre, la mise à disposition au plus grand nombre de ces données contribuera à favoriser la confiance citoyenne en la science, ainsi qu'à mettre en œuvre des politiques publiques de manière plus transparente. Le deuxième Plan national pour la science ouverte français, présenté en juillet 2021, évoque d'ailleurs les recherches en santé et la nécessité de « réduire le biais de publication, qui est la tendance à ne publier que les études ayant obtenu un résultat positif, au détriment des résultats peu concluants ou négatifs » [26]. Rappelons néanmoins que la FAIRisation des données n'implique pas leur ouverture automatique car des restrictions validées par la loi persistent, d'autant plus dans le champ médical.

▪ **Zaina NZOGU LOGANO**

Références:

[1] *Jaime Delgado et al., Approaches to the integration of TRUST and FAIR principles, présentation du 24 mars 2021 au SWForum.*
[2] *Cathrin Stöver et Karel Luyben, EOSC strategic implementation plan, p. 18.*
[3] *Turning FAIR into reality, p.11.*
[4] *Les enquêtes de Hans Dillaerts et al. et CommonData.*
[5] *Voir les branches I et K de MeSH: https://meshb.nlm.nih.gov/treeView*
[6] *Martin Benninghoff et al., p.12.*
[7] *Id.*
[8] *https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000006685827/2008-01-29/*
[9] *La recherche à l'Inserm : https://www.inserm.fr/recherche-inserm*
[10] *Jean Martin, p.91.*
[11] *Le Figaro : https://www.lefigaro.fr/flash-eco/quatre-nouveaux-milliardaires-francais-entrent-au-classement-forbes-20210406*
[12] *Rob Johnson et al., p.22.*
[13] *Ibid., p.61.*
[14] *Françoise Rousseau, p.12.*
[15] *Ibid., p.29.*
[16] *Ibid., p.69-70.*
[17] *Ibid., p.47.*
[18] *Ibid., p. 40.*
[19] *Comité pour la science ouverte : https://www.ouvrirlascience.fr/deux-nouveaux-groupes-de-travail-pour-la-recherche-en-sante/*
[20] *https://hl7.org/FHIR/overview.html*
[21] *FAIRness for FHIR (FHIR4FAIR): https://confluence.hl7.org/pages/viewpage.action?pageId=91991234*
[22] *FNS : http://www.snf.ch/fr/leFNS/points-de-vue-politique-de-recherche/open_research_data/Pages/depots-de-donnees.aspx*
[23] *Six recommentations for implementation of FAIR practice, p.17.*
[24] *https://www.ebi.ac.uk/biostudies/*
[25] *Vasiliki : https://mediaserver.unige.ch/play/137375*
[26] *MESRI : https://www.ouvrirlascience.fr/wp-content/uploads/2021/06/Deuxieme-Plan-National-Science-Ouverte_2021-2024.pdf*

Bibliographie

BENNINGHOFF Martin, RAMUZ Raphaël et LUTZ Andrea. La recherche biomédicale en Suisse : espace social, discours et pratiques. Document du Conseil suisse de la science (CSSI). 2/2014. [En ligne : https://www.swir.ch/images/stories/pdf/fr/SWIR_2_2014_Recherche_biomedicale.pdf]. Consulté le 10 février 2021.

Code de la santé publique. Partie législative. [En ligne : https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000006685827/2008-01-29/]. Consulté le 5 juin 2021.

Comité pour la science ouverte. Deux nouveaux groupes de travail pour la recherche en santé. 27 avril 2021. [En ligne : https://www.ouvrirlascience.fr/deux-nouveaux-groupes-de-travail-pour-la-recherche-en-sante/]. Consulté le 5 juin 2021.

Comité pour la science ouverte. Plans de partage des données issues des essais cliniques. Ouvrir la science. [En ligne: https://www.ouvrirlascience.fr/plan_de_partage_des_donnees_issues_des_essais_cliniques/?menu=1]. Consulté le 6 juin 2021.

Commission européenne. Turning FAIR into Reality. Final report and action plan from the European Commission expert group on FAIR data. 26 novembre 2018. [En ligne : https://op.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1]. Consulté le 10 février 2021.

DELGADO Jaime, ALVAREZ ROMERO Celia, MARTINEZ GARCIA Alicia et al. Approaches to the integration of TRUST and FAIR principles, présentation du 24 mars 2021 au SWForum. [En ligne : https://www.swforum.eu/sites/default/files/1stSwForumWs_paper_6.pdf]. Consulté le 27 mars 2021.

DILLAERTS Hans, PAGANELLI Céline, VERLAET Lise, et al. Usages et pratiques en lien avec les données de recherche. Une enquête menée auprès des chercheurs de l'université Paul-Valéry Montpellier 3. 20 juillet 2020. [En ligne : https://halshs.archives-ouvertes.fr/halshs-02902710]. Consulté le 10 février 2021.

ELIXIR. [En ligne : https://elixir-europe.org/]. Consulté le 10 juin 2021.

European Open Science FAIR working group. Six recommendations for implementation of FAIR practice. Luxembourg: Office des publications de l'Union européenne. Octobre 2020. [En ligne : doi: 10.2777/986252]. Consulté le 27 mars 2021.

Fast Healthcare Interoperability Resources. [En ligne : https://hl7.org/FHIR/index.html]. Consulté le 12 juin 2021.

Fonds national suisse de la recherche scientifique. Quels dépôts de données peuvent être utilisés ? [En ligne : http://www.snf.ch/fr/leFNS/points-de-vue-politique-de-recherche/open_research_data/Pages/depots-de-donnees.aspx] Consulté le 27 mars 2021.

FOUFI Vasiliki. FAIR4Health: Improving Health Research in EU through FAIR Data. Présentation au Swiss Research Data Day. 22 octobre 2020. [En ligne : https://mediaserver.unige.ch/play/137375]. Consulté le 3 février 2021.

JOHNSON Rob, WATKINSON Anthony et MABE Michael (éd.). The STM report : an overview of scientific and scholarly publishing. Octobre 2018. [En ligne : https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf]. Consulté le 8 février 2021.

Quatre nouveaux milliardaires français entrent au classement Forbes. Le Figaro. 6 avril 2021. [En ligne : https://www.lefigaro.fr/flash-eco/quatre-nouveaux-milliardaires-francais-entrent-au-classement-forbes-20210406]. Consulté le 5 juin 2021.

MARTIN Jean. Recherche bio-médicale : intérêts privés et intérêt public. Santé publique.89- Vol.13, 2001, p.89-93. [En ligne : https://www.cairn.info/revue-sante-publique-2001-1-page-89.htm]. Consulté le 10 février 2021.

Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. Deuxième Plan national pour la science ouverte. Généraliser la science ouverte en France 2021-2024. Juillet 2021. [En ligne : https://www.ouvrirlascience.fr/wp-content/uploads/2021/06/Deuxieme-Plan-National-Science-Ouverte_2021-2024.pdf] Consulté le 17 juillet 2021.

ROBIN Agnès, FRONTINI Francesca, CAILLOU Eliane, et al. Résultats de l'enquête CommonData : Pratiques de gestion des données scientifiques. 16 novembre 2020. [En ligne : http://www.mshsud.tv/spip.php?article1014]. Consulté le 10 février 2021.

ROUSSEAU-HANS Françoise, OLLENDORFF Christine et HARNAIS Vincent. Les pratiques de publications et d'accès ouvert des chercheurs français en 2019 : analyse de l'enquête Couperin 2019. 25 juin 2020. [En ligne : https://hal-cea.archives-ouvertes.fr/cea-02450324v2]. Consulté le 17 février 2021.

Set of common data elements for Rare Diseases Registration. [En ligne : https://eu-rd-platform.jrc.ec.europa.eu/sites/default/files/CDS/EU_RD_Platform_CDS_Final.pdf]. Consulté le 12 juin 2021.

Science ouverte : quels enjeux pour la psychologie ?

L'ouverture des données de la recherche sur les plateformes de science ouverte a un impact sur les usages qui en sont faits.
Dans le domaine de la psychologie, les enjeux sont multiples, de la publication des articles à la reproductibilité des expériences.

Guénola TARDY-JOUBERT

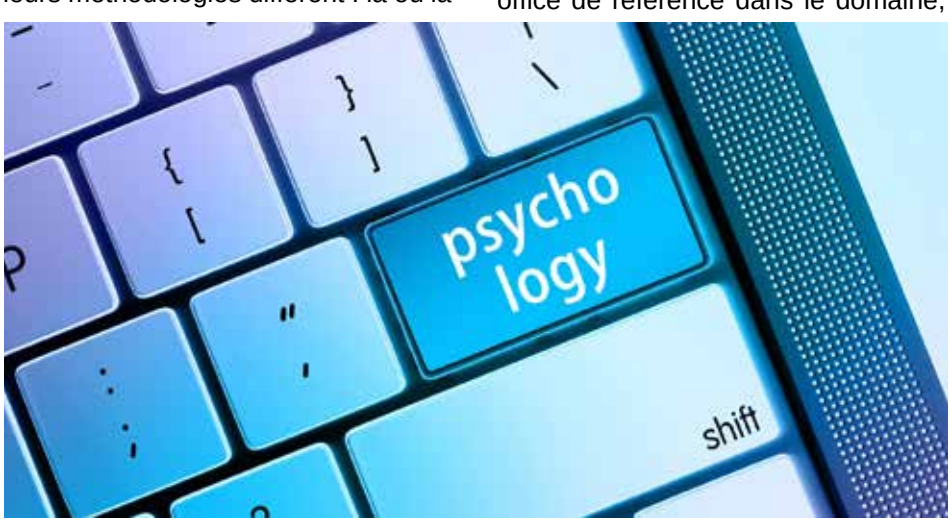
SPÉCIFICITÉS DE LA PSYCHOLOGIE

Une grande diversité intra-disciplinaire

La psychologie est une discipline relativement récente, puisqu'elle ne commence à être constituée comme science indépendante qu'à partir du XXème siècle, et particulièrement foisonnante : dans le même temps il s'agit d'un domaine complexe, qui comprend de nombreux sous-champs bien distincts et très spécialisés. Si ceux-ci ont tous en commun une volonté de comprendre les comportements humains, leurs objets comme leurs méthodologies diffèrent : là où la psychologie clinique va s'intéresser au vécu du patient et à l'étiologie de ses symptômes, la psychologie du travail peut élaborer des tests de personnalités, la neuropsychologie va analyser IRM et EEG, la psychologie du développement pratiquera des études longitudinales… Et ces différents champs de la psychologie n'ont pas forcément non plus les mêmes besoins – que cela soit en termes de pratiques, de méthodologie de recherche, etc.

Une discipline majoritairement anglophone

La psychologie américaine est particulièrement prégnante dans le domaine, et domine la discipline depuis ses débuts, au travers de nombreux courants et écoles de pensées.
La langue de publication majoritaire dans la discipline est l'anglais, aussi peu de revues ou publications dans les langues nationales permettent d'obtenir une véritable reconnaissance ; et jamais comparables aux revues anglophones. Dans le contexte actuel où il peut être difficile pour les chercheurs de distinguer les revues sérieuses des revues prédatrices, voire plus simplement peu scientifiques, l'APA (American Psychological Association), qui fait office de référence dans le domaine, propose régulièrement une liste des revues de référence [1]. Celle-ci reflète les orientations de la discipline, puisqu'à peine 3% des revues qui y sont citées sont francophones [2], et a soulevé de nombreuses protestations (Laurens, S., 2009), les indicateurs utilisés pour l'étalonner ignorant des revues nationalement reconnues.
Ceci explique certainement pourquoi il y a peu d'initiatives au niveau national pour la création de plateformes et portails destinées à la science ouverte, bien qu'un certain nombre d'initiatives européennes aient vu le jour ces dernières années [3].



Source : Pixabay

L'American Psychological Association

Ainsi que nous l'avons mentionné plus haut, l'APA (American Psychological Association) est une véritable institution dans le domaine, et propose aussi bien des références pour les normes appliquées au champ disciplinaire, que de nombreuses publications et services ; elle émet également des recommandations.
L'APA diffuse notamment via son site des ressources, destinées principalement aux documentalistes, auteurs, enseignants et étudiants ; ainsi qu'un « APA Databases and Electronic Resources » (notamment, la base de données PsycNet [4]). Cependant l'essentiel de ces documents ne sont pas accessibles librement ; le modèle de financement de l'APA se faisant par abonnement, généralement via des institutions, mais des solutions payantes sont également proposées aux particuliers. Cependant, l'APA s'engage également vers la science ouverte, et propose, notamment par un partenariat avec le Center for Open Science, une base de données, PsyArxiv [5] destinée à la publication de preprint, et encourage également aux dépôts de données de

lla recherche sur la plateforme OSF du *Center for Open Science,* ou sur des bases de données spécialisées [6], et renvoie également vers des sites permettant le dépôt de *datasets* [7].

L'association met également en avant l'ouverture des données de la recherche, la récompensant par des badges [8] décenés aux revues la mettant en oeuvre. Elle s'engage aussi en manifestant son soutien aux principes FAIR [9] et propose des recommandations et guides.

Les enjeux d'une science ouverte

Les enjeux de l'ouverture des données de la recherche

En 2011, un scandale a éclaté dans le domaine de la psychologie : Diederik Stapel, un éminent chercheur en psychologie sociale de l'Université de Tilburg aux Pays-Bas, a été reconnu coupable de fraude : depuis plusieurs années, il publiait des résultats falsifiés ; à l'insu du monde de la recherche y compris de ses co-auteurs et doctorants, auquel il ne communiquait pas ses données [10].

Cette affaire, qui a causé un véritable choc, a cependant le mérite de questionner les pratiques des chercheurs ; dans un champ disciplinaire traditionnellement relativement opaque, et mettre en avant la nécessité de diffuser non seulement les résultats, mais également l'ensemble des données de la recherche.

Afin de répondre à ces problématiques, plusieurs initiatives ont été mises en place dans la discipline.

D'une part, l'exigence par certaines revues à fort impact de la publication des jeux de données associées aux articles publiés dans leurs pages a un effet vertueux.

D'autre part, dans le but d'éviter, non pas les fraudes, mais les manipulations a posteriori des données non justifiées, certains sites ont mis en avant la possibilité de créer des « pré-enregistrements » [11] des expériences avant même la collecte des données : le chercheur peut ainsi « créer » virtuellement son expérience, élaborer et dater les hypothèses de travail. Cela permet de réduire les écarts entre ce qui est annoncé et ce qui sera effectivement mis en œuvre, ainsi que le risque de manipulation des statistiques a posteriori. La possibilité de mettre en œuvre un « peer review » avant même la collecte des données ouvre également des possibilités intéressantes : amélioration de la méthodologie, amélioration du design de l'expérience, ou certitude de pouvoir en publier les résultats quels qu'ils soient (Chambers, 2013) (Lantian, A., 2020).

Le "Reproducibility Project"

A l'instar d'autres sciences, la psychologie a connu ce que d'aucuns nomment une "crise de la reproductibilité" : une prise de conscience de la difficulté qu'il peut y avoir à répliquer les résultats obtenus dans les expériences.

C'est d'un questionnement à ce sujet qu'est né le "***Reproductibility project***" [12], mené par le psychologue Brian Nosek. Il s'agissait de tester la reproductibilité de la recherche en psychologie, en mettant en place des protocoles aussi proches que possible des expériences initiales, choisies dans des revues reconnues dans le domaine (le *Journal of Personality and Social Psychology, Psychological Science* et le *Journal of Experimental Psychology: Learning, Memory and Cognition*), en faisant appel à des chercheurs bénévoles. Les résultats, publiés à partir de 2015, ont montré que sur les 100 expériences testées, seulement 36 obtiennent des résultats statistiquement significatifs (alors que les publications faisaient état de résultats significatifs pour 97 sur 100). Ces divergences dans les résultats n'invalident pas nécessairement l'ensemble des recherches (Martone et al., 2018), mais soulignent la nécessité pour les chercheurs de pouvoir disposer de l'ensemble des données originales.

Source : Pixabay

Des oppositions à l'ouverture des données

Il existe cependant des voix qui s'interrogent, voire s'élèvent contre le partage ouvert des données de la recherche, pour des raisons très diverses. Outre l'obstacle du temps passé à préparer et publier les données, et le manque de connaissance des outils numériques permettant la publication ; certains y sont réticents par crainte de se voir "doubler" par d'autres chercheurs ; ou s'interrogent (parfois légitimement) sur le risque de réactions à leur endroit à la publication de leurs résultats. D'autres s'inquiètent quant à la possibilité de garantir la confidentialité de données sensibles, et des risques légaux comme éthiques qui découlent de leur partage.Enfin, il existe un réel questionnement sur l'utilité d'une telle pratique d'une part ; nombre de chercheurs jugeant peu probable que l'ensemble de leurs données soient réutilisées ou même simplement consultées par d'autres – voire s'interrogeant sur l'utilité qu'il peut y avoir à engager temps et ressources à reproduire des expériences passées, ou à réutiliser des données peut-être obsolètes (Nuijten et al., 2017) (Fiedler, K., Prager, J., 2018).

UN EXEMPLE DE TRAVAIL COLLABORATIF

La plateforme OSF (Open Science Framework)

La **plateforme OSF** (*Open Science Framework*) a été créée par le **COS** (*Center for Open Science*), qui est une organisation américaine à but non lucratif. Elle a été conçue comme une plateforme destinée à faciliter le travail collaboratif, l'échange entre les chercheurs, et les principes de la science ouverte. Elle permet d'y déposer tous les éléments permettant de documenter la recherche (hypothèses de recherche, protocoles expérimentaux, outils et modes d'emploi, résultats obtenus…), et en faciliter la réutilisation comme la réplication.

Les premiers travaux qui y sont initiés l'ont été dans le cadre du ***"Reproducibility Project"*** de Brian Nosek, qui a été suivi de travaux similaires sur la reproductibilité d'expériences de biologie sur le cancer [13]. C'est actuellement une plateforme multidisciplinaire, qui soutient autant que possible les principes FAIR (*Findable, Accessible, Interoperable, Reusable*) et la science ouverte (les licences sur la plateforme sont toutes de type *Creative Commons*).

Description de la plateforme

OSF est une plateforme collaborative *Open science*, qui propose plusieurs types de services à destination de la communauté des chercheurs, mais pas exclusivement puisqu'elle autorise 3 types d'inscription : avec identifiant ORCID, via une institution, ou une inscription libre, ce qui permet à tout un chacun de créer (ou collaborer à) des expériences.

Elle offre la possibilité de créer des expériences de façon individuelle comme collaboratives (il est possible d'inviter une équipe), de les élaborer en déterminant si l'on veut procéder à un pré-enregistrement, de proposer les différents fichiers et outils permettant sa mise en œuvre, d'expliquer l'analyse menée et comment répliquer les résultats (ou d'offrir des pistes pour sa réutilisation), et d'inclure les résultats obtenus, leur analyse, les éventuels liens. L'utilisateur peut choisir de partager avec une équipe ou ouvrir au public tout ou partie des données de la recherche, en choisissant quelle partie du travail il souhaite rendre publique.

La plateforme offre également la possibilité de se connecter avec différents outils (tels que *Dropbox, GoogleDrive*, et bien d'autres), ce qui permet une collaboration simplifiée et intuitive.

En 2017 elle s'adjoint *StudySwap* [14], une plateforme de partage de recherche, de réplication et d'échange avec d'autres chercheurs, qui permet notamment de signaler l'existence de données utilisables.

Conclusion : les points de vigilance pour l'ouverture de la science

De nombreuses initiatives voient le jour dans le domaine de la psychologie, bien qu'il reste encore un long chemin à parcourir pour une recherche plus ouverte et accessible à tous. Menées par la question de la vérifiabilité et reproductibilité des données, les interrogations actuelles, et leur mise en œuvre par divers acteurs reconnus dans la discipline tentent d'ébaucher des réponses au besoin de disposer de données de qualité et vérifiables – et particulièrement de jeux de données clairement décrits, aux utilisations explicitées et reproductibles, et réutilisables pour d'autres recherches. Les initiatives allant dans le sens d'un travail collaboratif et de l'ouverture de la science voient peu à peu le jour (on peut citer, par exemple, le "***Psychological Science Accelerator***" [15], qui cherche à fédérer un ensemble de laboratoires du monde entier), ce qui a un effet d'émulation. Un important travail reste cependant à accomplir, notamment d'information, mais aussi de formation de chercheurs ; et il semble important d'appuyer également la création de système européen, fussent-ils anglophones, afin de satisfaire aux exigences, en particulier légales, qui ont cours sur notre continent.

▪ **Guénola TARDY-JOUBERT**

Références:

[1] *https://www.apa.org/pubs/databases/psycinfo/coverage*
[2] *Celle-ci est reprise par l'HCERES https://www.hceres.fr/sites/default/files/media/downloads/Guide%20des%20produits%20de%20la%20recherche%20et%20des%20activite%CC%81s%20de%20recherche%20-%20Sous-domaines%20SHS%204%20-%20Discipline%20psychologie.pdf*
[3] *On peut citer notamment psychopen.eu et pub.psych.eu.*
[4] *https://www.apa.org/pubs/databases/psycnet*
[5] *https://psyarxiv.com*
[6] *https://www.re3data.org/*
[7] *https://www.apa.org/research/responsible/data-links*
[8] *https://www.apa.org/pubs/journals/resources/data-sharing*
[9] *https://www.apa.org/pubs/journals/resources/data-sharing-video*
[10] *https://www.apa.org/science/about/psa/2011/12/diederik-stapel*
[11] *Les sites osf.io ou aspredicted.org proposent notamment cette possibilité*
[12] *https://osf.io/ezcuj/wiki/home/*
[13] *https://osf.io/e81xl/wiki/home/*
[14] *https://osf.io/meetings/studyswap/*
[15] *https://psysciacc.org/*

Bibliographie

An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science. (2012). Perspectives on Psychological Science, 7(6), 657–660. https://doi.org/10.1177/1745691612462588

Open Science Collaboration, « Estimating the reproducibility of psychological science », Science, vol. CCCXLIX, n° 6251, 28 août 2015

Althaus, V. (2019). Le capitalisme à l'assaut des sciences humaines et sociales : l'exemple des revues payantes en psychologie. Zilsel, 2(2), 9-24. https://doi.org/10.3917/zil.006.0009

Chambers CD. (2013). Registered reports: a new publishing initiative at Cortex. Cortex. 2013 Mar;49(3):609-10. doi: 10.1016/j.cortex.2012.12.016. Epub 2012 Dec 26. PMID: 23347556.

Chartier, C. R., Riegelman, A., & McCarthy, R. J. (2018). StudySwap: A Platform for Interlab Replication, Collaboration, and Resource Exchange. Advances in Methods and Practices in Psychological Science, 574–579. https://doi.org/10.1177/2515245918808767

Crüwell, S., van Doorn, J., Etz, A., Makel, M. C., Moshontz, H., Niebaum, J. C., Schulte-Mecklenbeck, M. (2019). Seven easy steps to open science: An annotated reading list. Zeitschrift für Psychologie, 227(4), 237-248. http://dx.doi.org/10.1027/2151-2604/a000387

Friedler, K., & Prager, J. (2018). The regression trap and other pitfalls of replication science—Illustrated by the report of the Open Science Collaboration. Basic and Applied Social Psychology, 40(3), 115–124. https://doi.org/10.1080/01973533.2017.1421953

Garrett-Ruffin, S., Cowden Hindash, A., Kaczkurkin, A. N., Mears, R. P., Morales, S., Paul, K., Pavlov, Y. G., Keil, A., 2021, Open science in psychophysiology: An overview of challenges and emerging solutions, International Journal of Psychophysiology, Volume 162, Pages 69-78, https://doi.org/10.1016/j.ijpsycho.2021.02.005

Jelmer J. Zondergeld, Ron H.H. Scholten, Barbara M.I. Vreede, Roy S. Hessels, A.G. Pijl, Jacobine E. Buizer-Voskamp, Menno Rasch, Otto A. Lange, … Coosje L.S. Veldkamp (2020). FAIR, safe and high-quality data: The data infrastructure and accessibility of the YOUth cohort study. Developmental Cognitive Neuroscience, Volume 45, 100834, https://doi.org/10.1016/j.dcn.2020.100834

Lantian, A. (2020). Les pratiques de recherche ouvertes en psychologie. Psychologie Française ElsevierMasson, 66. 71-90. 10.1016/j.psfr.2020.09.001, hal-03161592

Laurens, S. (2009). L'étrange classement des revues de psychologie et le cas mystérieux du Bulletin de psychologie. Bulletin de psychologie, 1(1), 67-78. https://doi.org/10.3917/bupsy.499.0067

Les publications de psychologie en langue française. Journée du 3 décembre 2010: Conférence des publications de psychologie en langue française. Bulletin de psychologie, 1(1), 49-53. https://doi.org/10.3917/bupsy.511.0049

Marmion, J. (2017). Malaise dans la recherche. Dans : Jean-François Marmion éd., La psychologie aujourd'hui (pp. 129-132). Auxerre, France: Éditions Sciences Humaines. https://doi.org/10.3917/sh.marmi.2017.01.0129

Martone, M. E., Garcia-Castro, A., & VandenBos, G. R. (2018). Data sharing in psychology. The American psychologist, 73(2), 111–125. https://doi.org/10.1037/amp0000242

Miłkowski, M. & Hohol, M. & Hensel, W. (2018). Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. Journal of Computational Neuroscience. 1-10. 10.1007/s10827-018-0702-z

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., … Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network. Advances in Methods and Practices in Psychological Science, 501–515. https://doi.org/10.1177/2515245918797607

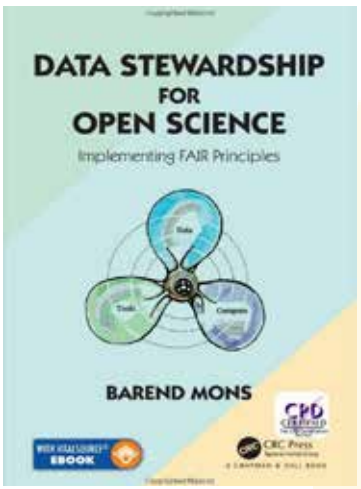
Nuijten, M., Borghuis, J., Veldkamp, C., Dominguez-Alvarez, L., van Assen, M., Wicherts, J.M. (2017). Journal Data Sharing Policies and Statistical Reporting Inconsistencies in Psychology. Collabra: Psychology 1 January 2017; 3 (1): 31. https://doi.org/10.1525/collabra.102

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An Introduction to Registered Replication Reports at Perspectives on Psychological Science. Perspectives on Psychological Science, 9(5), 552–555. https://doi.org/10.1177/1745691614543974

von Davier, A., Hao, J., Liu, L., Kyllonen, P., (2017). Interdisciplinary research agenda in support of assessment of collaborative problem solving: lessons learned from developing a Collaborative Science Assessment Prototype. Computers in Human Behavior, Volume 76, Pages 631-640, https://doi.org/10.1016/j.chb.2017.04.059

Wicherts, J.M., 2013. Science revolves around the data. Journal of Open Psychology Data, 1(1), p.e1. http://doi.org/10.5334/jopd.e1

BIBLIOGRAPHIE



Data Stewardship for Open Science
Implementing Fair Principles.
Mons, B. (2018).
CRC Press Inc.

Open Data
Ouverture, Exploitation, Valorisation
DesDonnées Publiques
Kober, V. (2014).
(Territorial Editions).

Libre accès et données de la recherche.
Quelle résonance dans les SIC ?
Études de communication, 1(52), 248.
(2019).

Ouvrir, partager, réutiliser
Regards critiques sur les données numériques.
In C. Mabi, J.-C. Plantin, & L. Monnoyer-Smith (Éds.),
(2017).
Éditions de la Maison des sciences de l'homme.

Éditorialisation des données de recherche
Le rôle des professionnels de l'information.
Schöpfel, J. (2020).
I2D - Information, données documents, n° 2(2), 82-84.

Archives ouvertes et publication scientifique
Comment mettre en place l'accès libre aux résultats de la recherche ?
Chanier, T. (2005).
Editions L'Harmattan.

GLOSSAIRE

D

Data Paper
Publication évaluée et validée par des pairs. Il s'agit d'une publication dont le descriptif du contenu se fait par jeux de métadonnées. Le *Data Paper* permet de garantir la qualité des informations et d'offrir la possibilité de les réutiliser grâce à une meilleure interprétation des données scientifiques qu'il contient.

Données à caractère personnel
Une information se rapportant à une personne physique identifiée et identifiable, directement ou indirectement relié à cette dernière. Les noms et prénoms sont des données directement identifiées mais un numéro de sécurité sociale ou une signature manuscrite sont des données qui peuvent servir à identifier une personne par un jeu de croisement de données. Le traitement des données à caractère personnel est encadré par le RGPD au niveau européen.

E

Entrepôt de données (EDD) ou Data warehouse (DWH)
Typologie de base de données spécialement conçue pour exécuter des analyses de données. À la différence d'une base de données relationnelle, l'entrepôt de données permet de « faire parler » les données à l'aide d'outils de type BI, SQL ou autre application. L'entrepôt de données, généralement un fichier OLAP, peut contenir plusieurs bases de données et il peut aller puiser de multiples informations et les croiser pour répondre à des requêtes et générer des rapports.

F

FAIR
Acronyme pour Findable, Accessible, Interoperable et Reusable. Le principe FAIR est la pierre angulaire de l'*Open Data*. En répondant aux principes évoqués plus haut, les données FAIR doivent être détaillées à l'aide de métadonnées et doivent avoir un moyen d'identification pérenne. Elles ne sont pas obligatoirement des données ouvertes (raisons juridiques ou besoin de consultation avant ouverture) mais elles doivent être récupérables. L'interopérabilité de ces données se fait par la sémantique et la syntaxe adaptée, leur format doit respecter les normes internationales ainsi que le vocabulaire informatique approprié. Enfin, la réutilisation des données FAIR doit être clairement énoncée et facilitée par l'usage de standards internationaux et des licences d'utilisation de type *Creative Commons*.

O

Open Data ou Données ouvertes
Données numériques dont l'accès, l'utilisation et la rediffusion sont libres. Ces données respectent une certaine structure et la licence ouverte associée garanti leur accès, leur réutilisation par tous sans restriction (juridique, financière). Pour que la donnée soit qualifiée d'ouverte elle doit respecter les 10 points suivants et doit être : complète, primaire, opportune, accessible, exploitable, non discriminatoire, non propriétaire, libre de droits, permanente et gratuite.

P

Plan de Gestion des Données ou Data Management Plan
Il s'agit d'un instrument de la pratique FAIR. Le PGD aide les chercheurs à gérer les données de leurs recherches dès le début afin que ces dernières soient décrites et organisées en vue de leur partage et de leur pérennité (accès et exploitation ultérieure). Le PDG peut varier selon les disciplines et le projet de recherche mais il reste dans tous les cas un outil évolutif qui s'appuie sur le cycle de vie des données (étapes de traitement).

www.didaktic.fr

UNIVERSITÉ PAUL VALÉRY MONTPELLIER 3

ITIC
INSTITUT DES TECHNOSCIENCES DE L'INFORMATION ET DE LA COMMUNICATION