ourouk.fr
CONSEIL
EN MANAGEMENT
DE L'INFORMATION

ANR-10-IDEX-0004-02

ETUDES COPIST (CATALOGUE D'OFFRES PARTAGEES D'IST)

Rapport de l'étude n° 2 : Une plateforme mutualisée pluridisciplinaire de stockage, de gestion, de signalement et de partage de données de recherche

Version finale du 24/08/2018

Référence : CNRS_COPIST_ETUDE 2_24/08/2018_VF

Cette étude résulte du travail du cabinet Ourouk.
Ses commanditaires ne sont pas engagés par les recommandations qu'elle contient.

5 RUE AMBROISE THOMAS
75009 PARIS
www.ourouk.fr

TÉL : 01 44 82 09 99
FAX : 01 44 82 72 70
MÉL : ourouk@ourouk.fr

SARL AU CAPITAL DE 358 800 €
RCS B 387472160 - APE 7022Z
AGRÉMENT FORMATION : 11753245175

SOMMAIRE

Introduction

Objet du document

Le présent document constitue la version finale du rapport de l'étude Copist n°2 « *Une plateforme mutualisée pluridisciplinaire de stockage, de gestion, de signalement et de partage de données de recherche : enjeux, conditions, faisabilité* ».

Il est complété par deux annexes, sur le développement et les usages des plateformes étudiées (annexe 2, page 52), et sur les éléments financiers relatifs à une plateforme de données (annexe 3, page 56). Il intègre en outre des commentaires post-étude de ses commanditaires (CPU, ADBU, COUPERIN, EPRIST et CNRS) dans le corps du rapport et en annexe 4, page 60.

Six études pour la mise en cohérence des projets numériques en IST de l'ESR

L'enquête nationale COPIST[1] sur les services d'information scientifique et technique a révélé une forte intention de partage dans les différents domaines de l'IST de la part de la centaine d'établissements de l'ESR qui y ont participé.

L'analyse conjointe des résultats de l'enquête par la CPU, l'ADBU, COUPERIN, EPRIST et le CNRS a abouti à la définition de 6 thèmes d'études destinées à préciser les attentes prioritaires des institutions de l'ESR.

Les thèmes retenus s'inscrivent tous dans la perspective d'une mutualisation de grands services d'IST à l'échelle nationale, à l'instar de la démarche qui a présidé au développement du projet ISTEX :

1. Des outils mutualisés de signalement et d'accès aux ressources documentaires électroniques pour l'ESR : enjeux, conditions, faisabilité
2. Une plateforme mutualisée pluridisciplinaire de stockage, de gestion, de signalement et de partage de données de recherche : enjeux, conditions, faisabilité
3. Des modèles innovants de publication et d'édition scientifiques publiques : conditions de développement
4. L'articulation des archives des établissements et de l'archive nationale pluridisciplinaire HAL : besoins et solutions
5. Des modèles économiques et de financement des services d'IST
6. Des actions de communication et de formation destinées à assurer une meilleure visibilité et appropriation des services d'IST par les communautés de recherche et les professionnels de l'information

Ces études de cadrage ont une visée opérationnelle. Leur objectif général est de définir la portée et le périmètre des évolutions de services existants et de nouveaux services répondant aux besoins et aux usages des différentes communautés de l'ESR, structurés et dotés d'un modèle économique pérenne.

[1] Rapport disponible sur www.cnrs.fr/dist/z-outils/documents/copist_rapport-analyse-conjointe_18-05-2017.pdf

Il s'agit, *in fine*, d'aboutir dans le consensus, et à partir d'un état des lieux initial des ressources et pratiques, à un Catalogue d'offres partagées d'Information scientifique et technique (COPIST) qui s'inscrit dans la stratégie du CNRS et de ses partenaires de l'ESR de « *Mieux partager l'IST pour mieux partager les connaissances* ».

Problématique de l'étude

Des communautés scientifiques ont, de longue date, organisé la gestion, le partage et la valorisation des données de recherche qu'elles produisent. La typologie de ces données est variée, que ce soit des mesures, des relevés, des résultats d'enquêtes ou d'expériences, des données textuelles, chiffrées ou photographiques. Ces données sont nécessaires à la validation des résultats scientifiques.

Lors d'un séminaire avec la Direction de l'Information Scientifique et Technique (DIST) du CNRS sur la donnée scientifique partageable, le 11 octobre 2017, le Conseil scientifique du CNRS a souligné que l'accès ouvert aux données et aux publications est essentiel et ouvre d'immenses perspectives : « *Ce développement d'un accès efficace et pertinent aux données de la recherche, respectueux des spécificités disciplinaires mais restant ouvert à la transdisciplinarité, suppose la mise en place ou le développement de structures IST dédiées à toutes les échelles territoriales, tant au niveau de l'organisme que de chaque Institut* »[2].

Mais toutes les communautés ne disposent pas de ce savoir-faire et certaines se trouvent dorénavant confrontées à la nécessité, voire à l'obligation, d'entreprendre cette démarche dans le cadre de leurs projets de recherche. Toutefois, leurs besoins étant inégaux et différenciés, des solutions adaptées doivent être imaginées comme cela a été fait, par exemple, pour les humanités numériques, avec les services développés par la TGIR Huma-Num[3], ou encore par le CCSD avec MédiHAL[4].

Les résultats de l'enquête nationale COPIST témoignent d'un besoin de nombreux établissements, les universités en particulier, de disposer pour leurs communautés les moins organisées dans ce domaine, d'une solution nationale d'archivage, de signalement et de partage de données de recherche, incluant les questions de droit et d'éthique. Il s'agit en quelque sorte d'un *« cloud de données scientifiques* » par défaut, pluridisciplinaire, indépendant des grands instruments, répondant à toutes les situations pour lesquelles il n'existe pas de solution spécifique.

Le verbatim suivant, issu de l'enquête COPIST, résume cette attente : *« Il serait urgent de développer une offre à destination des institutions pour le stockage et la conservation* [de données de recherche]*, articulée avec l'offre de service développée localement pour la diffusion. »* Or, comme l'ont constaté les correspondants Information Scientifique et Technique du CNRS (réunion des CorIST du 19 octobre 2017), pour les données de recherche, « *il existe 100 infrastructures en France, mais il n'y a pas de structure banalisée qui serait une infrastructure par défaut avec vocation à disparaitre quand toutes les disciplines seront pourvues* ».

[2] http://www.cnrs.fr/comitenational/cs/recommandations/23-24_nov_2017/Reco_Les-moyens-du-partage-des-donnees-scientifiques.pdf

[3] A propos d'un nouveau dispositif de stockage distribué au sein des MSH destiné à faciliter aux chercheurs le stockage, la sécurisation et la gestion de leurs données réputées « tièdes » voire « froides » : cf. Marchand, J., 2017. Stockage distribué sécurisé pour les sciences humaines et sociales. In: *JRES 2017, 14 novembre 2017, Nantes*. https://halshs.archives-ouvertes.fr/hal-01638113

[4] Archive ouverte pour images fixes, vidéos et sons https://medihal.archives-ouvertes.fr/

Ce besoin n'est pas spécifique à la France. Pour ne citer qu'un seul exemple, l'Australian National Data Service travaille depuis 2012 à la transformation de données scientifiques *« non gérées, déconnectées, invisibles et à usage unique* »[5] (en d'autres termes, les données de la longue traîne[6]) en collections structurées, gérées, connectées, identifiables et réutilisables. L'objectif est que les chercheurs australiens puissent facilement publier, découvrir, accéder et utiliser les données de la recherche qu'ils produisent[7].

L'enquête sur les archives ouvertes réalisée par COUPERIN en 2017 a révélé que trois établissements sur quatre n'ont pas défini de politique concernant les données de la recherche et que presque autant ne savent pas où leurs chercheurs déposent leurs jeux de données. Seulement 10% des universités disposent d'un système d'information recherche. De même, une enquête récente de l'URFIST de Rennes confirme, pour les SHS, l'absence de politiques en matière de données et des pratiques hétérogènes, chronophages, peu performantes et peu sécurisées, souvent avec des moyens et outils personnels voire privés[8].

Attendus de l'étude

L'objectif de l'étude est de définir le périmètre des services et les conditions de mise en œuvre d'une plateforme nationale mutualisée pluridisciplinaire dédiée à l'archivage, au signalement, au partage et à la diffusion de données de recherche, interopérable avec les initiatives locales (politiques de site), les prestations d'archivage pérenne proposées par le CINES et les principales plateformes internationales.

En tout premier lieu, le dispositif a vocation à répondre aux besoins de (petites) communautés souvent transdisciplinaires, sans réelles réflexions ou pratiques concernant la gestion des données, qui n'identifient pas de solutions, en France notamment, correspondant à leurs besoins et aux exigences qui leur sont faites d'exposer leurs données.

Ainsi, il s'agit d'évaluer la faisabilité d'un dispositif complémentaire (nous l'appellerons « *Data Alpha* » dans ce rapport) aux infrastructures existantes, pour faciliter la gestion globale des jeux de données à l'issue d'un projet ou d'une publication, sur le modèle, par exemple, des plateformes Zenodo (CERN, Suisse) ou EASY (DANS, Pays Bas).

[5] "unmanaged, disconnected, invisible and single-use"
[6] Cf. aussi les travaux du groupe d'intérêt « Long Tail Of Research Data » de la Research Data Alliance (RDA) https://rd-alliance.org/groups/long-tail-research-data-ig.html
[7] Treloar, A., Choudhury, G. S., Michener, W., 2012. Contrasting national research data strategies: Australia and the USA. In: Pryor, G. (Ed.), *Managing research data*. Facet Publishing, Croydon, pp. 173-203.
[8] Serres, A. et al., 2017. *Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2*. Université Rennes 2. https://hal.archives-ouvertes.fr/hal-01635186

Eléments méthodologiques

L'étude applique une démarche en deux temps :

- L'analyse de plateformes et de services de données comparables à l'international destinée à cerner le périmètre de ces dispositifs, leurs principales fonctionnalités, leur organisation et leurs enjeux. Néanmoins, l'objectif n'est pas d'importer des solutions qui fonctionnent dans d'autres pays, mais d'en identifier les traits caractéristiques et d'évaluer leur intérêt et faisabilité pour l'ESR français.
- Outre une synthèse des travaux réalisée dans le cadre de la BSN 10, des entretiens auprès de responsables de services et d'experts dans le domaine des données de la recherche au sein de l'ESR, ont été menés pour analyser le contexte et les opportunités d'un tel dispositif en France, mais aussi évaluer sa faisabilité et d'identifier les risques et verrous potentiels à sa réalisation.

Plateformes et entrepôts de données étudiés

Six critères de sélection

Le choix des dispositifs a été effectué à partir des six critères suivants :

- Positionnement national
- Dispositif mutualisé
- Pluridisciplinarité
- Indépendance par rapport à des équipements, procédures ou infrastructures spécifiques
- Interopérabilité avec les initiatives locales et les principales plateformes internationales
- Connexion avec les projets de recherche et les publications scientifiques

Huit dispositifs retenus

Huit dispositifs dans six pays, dont deux plateformes internationales, répondant à ces critères ont été étudiés :

Pays	Nom complet	Opérateur(s)	Services
International	Zenodo	CERN	Entrepôt
International	Figshare	Digital Science (Holtzbrinck)	Entrepôt
Pays Bas	EASY	DANS (KNAW)	Entrepôt
Allemagne	Research Data Repositorium RADAR	TIB / FIZ-K / KIT (DFG)	Plateforme (gestion, stockage)
Royaume Uni	Apollo University Data Repository	University of Cambridge	Entrepôt
République Tchèque	ASEP	Académie des Sciences	Entrepôt
Canada	Federated Research Data Repository FRDR	Canadian Association of Research Libraries CARL	Plateforme (gestion, stockage)
Etats-Unis	DataONE Dash	California Digital Library, University of California	Plateforme (gestion, stockage)

Certains dispositifs sont des entrepôts de données, les autres sont des plateformes de gestion et de stockage qui proposent (aussi) des services de gestion et d'analyse de données. Tous les dispositifs sont multidisciplinaires, même si EASY contient avant tout des jeux de données issus des sciences humaines et sociales.

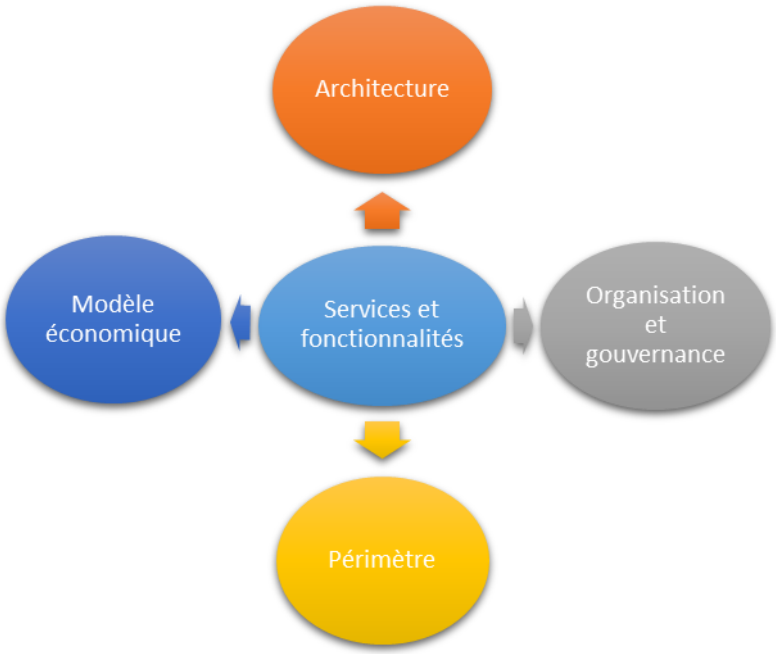
L'évaluation des dispositifs en ligne a été accompagnée par des entretiens avec des représentants des plateformes et entrepôts (cf. page 51).

Les résultats détaillés sont accessibles sur l'espace collaboratif (wiki) de l'étude[9].

Cinq axes d'analyse

A partir de plusieurs checklists et certifications[10], cinq axes d'analyse ont été retenus. L'objectif était de produire des éléments empiriques pour décrire le modèle d'un dispositif dédié au stockage, à la gestion, au signalement, au partage et à la diffusion de données de recherche.

Les cinq axes couvrent l'essentiel des fonctionnalités et de l'organisation des dispositifs de l'information scientifique :



Evaluation des conditions de réalisation : personnes interrogées

L'objectif de la deuxième étape de l'étude est l'évaluation de la faisabilité technique, politique et financière de la mise en place de la plateforme cible. Elle s'appuie sur une analyse stratégique des études de cas, suivie d'une analyse des risques identifiés sous forme de faiblesses et menaces d'un tel projet.

[9] http://copist2017data.pbworks.com

[10] Dont la *Matrix of use cases and functional requirements for research data repository platforms* de la RDA, le *Data Seal of Approval*, le *Core Trust Seal*, le certificat *DINI* et d'autres listes de critères, comme la liste de van Zeeland, H., Ringersma, J., 2017. The development of a research data policy at Wageningen University & Research: best practices as a framework. *LIBER Quaterly* 27 (1), 153-170. https://www.liberquarterly.eu/article/10.18352/lq.10215/

Cette démarche a mobilisé l'expertise des représentants de plusieurs acteurs au sein de l'ESR, aussi bien du côté des opérateurs que du côté des établissements et organismes producteurs et consommateurs de données.

Au total, 29 personnes ont été interrogées, parmi lesquelles des membres de la BSN10 et la personne ayant réalisé la cartographie des dispositifs services existants ou émergents, des représentants d'établissements ayant mis en œuvre des services d'appui à leurs communautés, des représentants d'établissements ayant exprimé des demandes dans le cadre de l'enquête COPIST, le directeur d'Huma-Num pour tirer des enseignements de l'expérience de la TGIR et des services qu'elle a développés, et des représentants de l'INIST-CNRS (cf. page 51).

Les professionnels de l'information constituent la plus grande part des personnes interrogées pour les principales raisons suivantes :

- Les contraintes budgétaires de l'étude, voire de planning, ne permettaient pas d'envisager un nombre significatif d'entretiens avec des chercheurs représentant les différentes disciplines susceptibles d'être concernées par un tel projet.
- Des enquêtes réalisées en France (Strasbourg, Rennes, Lille…) et dans d'autres pays ont montré que le besoin d'un entrepôt « par défaut » était plutôt exprimé par des « chercheurs isolés », producteurs de la « longue traîne des données ».
- Comme la Research Data Alliance (RDA) le souligne, les professionnels de l'information ont, avec les Directions des Systèmes d'Information et de la Recherche, une responsabilité particulière pour la construction des services adaptés à la longue traîne des données de la recherche[11].

Toutefois, nous pensons qu'il serait sans doute nécessaire de compléter, voire de confronter les enseignements de cette étude en interrogeant des chercheurs représentants de différentes disciplines et qui manipulent différents types de données, dans le cadre d'une étude complémentaire.

Précisions terminologiques

Nous utiliserons dans le rapport le terme de « dispositif » au sens large, comme outil (ou service d'outils) de médiation de l'information scientifique entre producteurs et utilisateurs.

Le terme de « plateforme » désignera des dispositifs informationnels « *aux fonctionnalités multiples* [dont] *l'accès à la documentation et aux publications scientifiques, services à valeur ajoutée de traitement de l'information (…)* »[12], soit un ensemble de logiciels, matériels, opérations et réseaux plus ou moins cohérent, avec en particulier des outils de découverte, de gestion des droits d'accès, de dépôt….

[11] "Libraries, IT services and research offices at the institution are asked to collaborate locally, nationally and globally in order to jointly build data libraries that support diverse data and research in the long-tail" (RDA Interest Group "Long tail of research data" 2016. *10 ways to support data diversity and the long tail* https://rd-alliance.org/group/long-tail-research-data-ig/post/please-contribute-10-ways-support-long-tail).

[12] Alain Bensoussan Avocats (2017) : *Analyse systémique de la loi pour une République numérique. Pour une Science ouverte et un droit des données scientifiques* V.1. Paris, CNRS.

Quant aux termes « archive » ou « entrepôt de données », nous appliquerons une définition large selon laquelle un tel « digital repository » possède quatre caractéristiques essentielles :

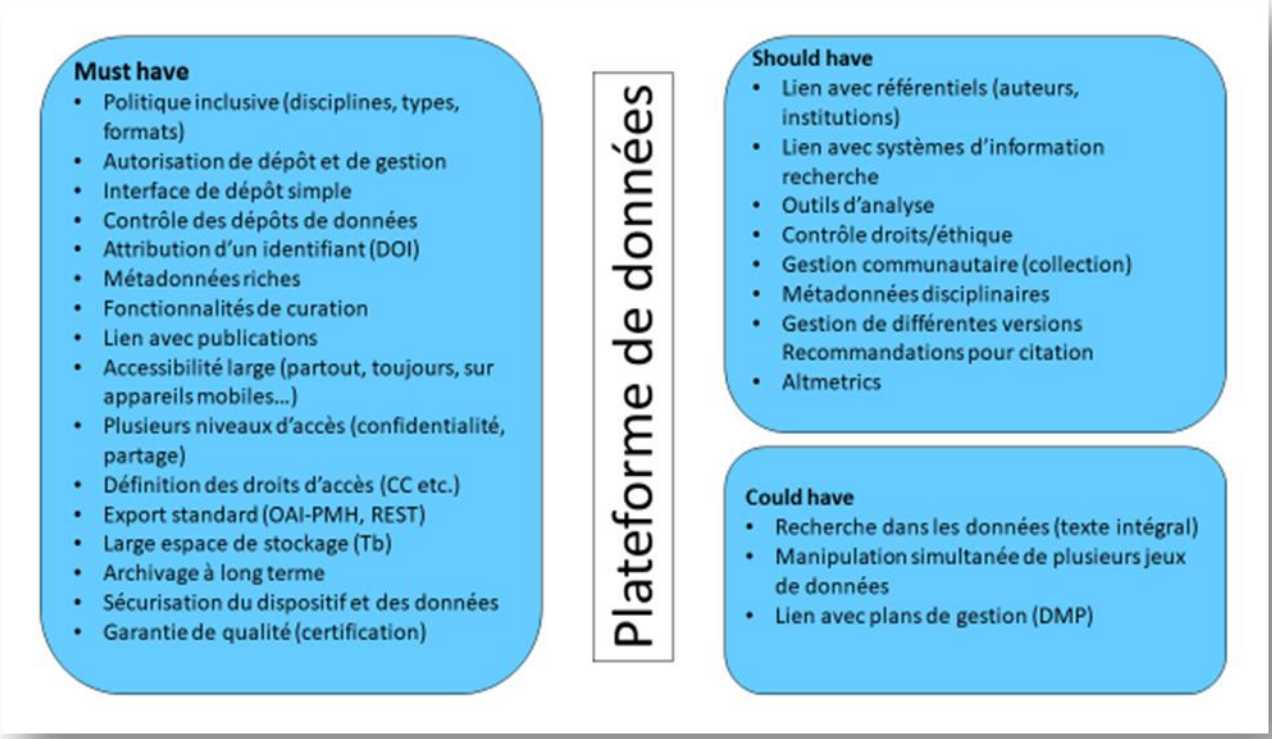
- Le contenu est déposé dans un entrepôt, que ce soit par le créateur de contenu, le propriétaire ou un tiers.
- L'architecture du dispositif gère le contenu ainsi que les métadonnées.
- Le dispositif offre un ensemble minimum de services de base, par ex. dépôt, export, recherche, contrôle d'accès.
- Le site doit être durable et fiable, bien supporté et bien géré[13].

[13] Heery, R., Anderson, S., 2005. *Digital repositories review*. Report. Joint Information Systems Committee, Bath.

Résumé

L'objectif de l'étude Copist n° 2, « *Une plateforme mutualisée pluridisciplinaire de stockage, de gestion, de signalement et de partage de données de recherche : enjeux, conditions, faisabilité* » est de définir le périmètre des services et les conditions de mise en œuvre d'une plateforme nationale mutualisée pluridisciplinaire dédiée à l'archivage, au signalement, au partage et à la diffusion de données de recherche, interopérable avec les initiatives locales (politiques de site), avec les prestations d'archivage pérenne proposées par le CINES et avec les principales plateformes internationales.

L'analyse de huit dispositifs de données permet de montrer les éléments essentiels de la plateforme cible dont certains sont indispensables pour qu'elle soit en conformité avec les principes FAIR. L'étude établit une hiérarchie de 27 services et fonctionnalités et décrit l'architecture, la gouvernance, le périmètre et le modèle économique des différents modèles.



Les entretiens réalisés révèlent des opportunités et une dynamique globalement favorable à la mise en place d'une telle plateforme, mais identifie aussi plusieurs verrous et risques qu'il convient d'anticiper.

Facteurs favorables	Facteurs défavorables
•Une prise de conscience de la nécessité d'agir •L'identification d'un faisceau de besoins sans réponse •La maîtrise des solutions techniques •La constat d'un paysage morcelé, incomplet •L'existence de modèles reconnus mais limités •Une politique volontaire en faveur de la Science ouverte	•Disparité des situations et besoins •La préférence d'une approche disciplinaire •Refus d'une solution imposée, rigide •Manque de ressources •Absence d'un opérateur "naturel", évident •Concurrence des solutions alternatives

Pour finir, l'étude formule neuf propositions pour renforcer la faisabilité d'une plateforme mutualisée pluridisciplinaire de stockage, de gestion, de signalement et de partage de données de recherche et pour réussir le lancement du projet.

- Pas de "couteau suisse"
- Opter pour une solution pragmatique
- Adopter une gestion agile et itératuve
- Prendre en compte les communautés et disciplines
- Définir une gouvernance partenariale
- Créer un environnement de service large, avec un modèle économique à trouver
- Soutenir des infrastructures opérationnelles
- Accompagner le projet par une politique d'incitation
- S'intégrer dans l'European Open Science Cloud (EOSC)

Le facteur temps sera crucial pour la réussite de ce projet : dans un environnement très dynamique, les chercheurs concernés et les professionnels de l'information attendent une réponse à court terme (2019).

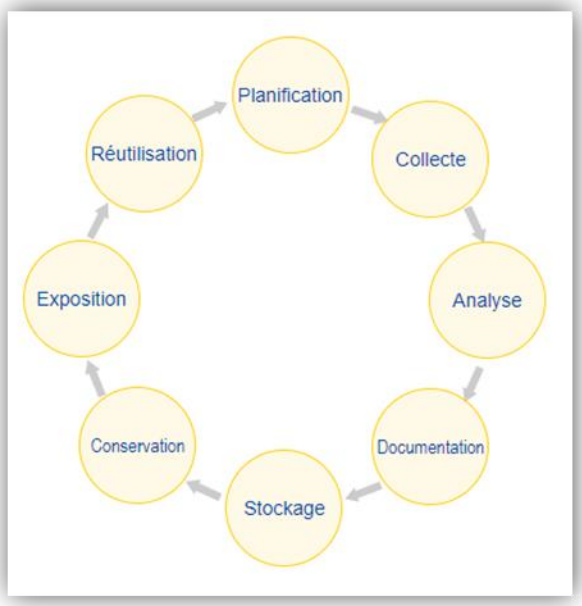
Partie 1 : Une plateforme FAIR, et davantage

Introduction

L'étude des huit dispositifs retenus permet de décrire des éléments essentiels de la plateforme « *Data Alpha* », sous forme d'un modèle idéal. Certains aspects sont indispensables et nécessaires, tandis que d'autres sont plutôt souhaitables ou facultatifs, selon la largeur des services offerts et la couverture des différentes étapes du cycle de vie des données de la recherche.

Rappelons qu'une partie des dispositifs étudiés sont des entrepôts de données, dont le but principal est de stocker, conserver et exposer les jeux de données, avec une documentation plus ou moins développée pour en faciliter la réutilisation. Les autres dispositifs sont des plateformes de gestion et de stockage qui intègrent les fonctionnalités d'un entrepôt, mais proposent aussi des services de gestion et d'analyse de données.

Fonctionnalités et cycle de vie des données[14]



La première partie de l'étude présente les éléments constitutifs d'une plateforme mutualisée pluridisciplinaire de stockage, de gestion, de signalement et de partage de données de recherche, à partir d'une analyse détaillée de différents dispositifs similaires.

Sur le plan international, la déclaration de l'European Open Science Cloud (EOSC) du 26 octobre 2017 préconise l'accès pour tous les chercheurs européens à un environnement de données de recherche ouvert par défaut, efficace et interdisciplinaire soutenu par les principes de données FAIR[15]. Fin 2017, les Pays-Bas, l'Allemagne et la France ont annoncé la création d'un bureau international d'appui et

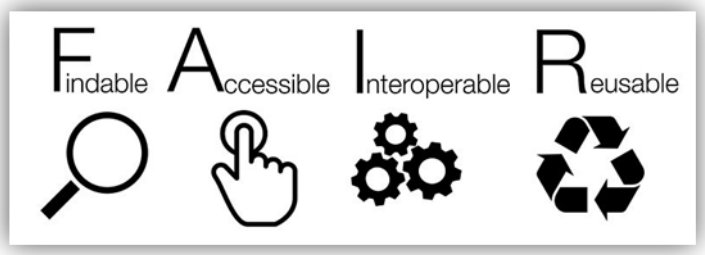
[14] Source : https://cat.opidor.fr
[15] http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

de coordination[16] de l'initiative GO FAIR,[17] pour aider les réseaux de mise en œuvre d'infrastructures numériques et pour soutenir la recherche et les communautés dans leurs efforts pour fournir des données et des services FAIR[18]. Pour cette raison, nous présenteront les différents aspects en deux temps : d'abord ceux directement concernés par l'initiative GO FAIR, puis les autres.

Les quatre principes FAIR

Le cadre conceptuel pour toute plateforme de données est aujourd'hui l'application des quatre principes du *fair data*[19]. L'étude des différents dispositifs confirme le souci des opérateurs de rendre les données trouvables, accessibles, interopérables et réutilisables, à des degrés plus ou moins avancés. En effet, il s'agit de principes pour le développement d'entrepôts et d'autres sites pour la gestion et le stockage de données de la recherche, afin de faciliter la communication entre machines et la création d'infrastructures scientifiques. Le facteur humain (compétences, attitudes…) n'est pas l'objet principal de ces *guiding principles* mais joue son rôle[20].

Les quatre principes du *FAIR data*[21]



Récemment, une équipe australienne a décrit comment une infrastructure nationale de données peut appliquer une démarche qualité (*Data Quality Strategy*) pour la mise en place d'une plateforme de données fiable, transdisciplinaire et performante[22]. La démarche s'appuie sur la standardisation des données et des services et inclut :

- le contrôle de la cohérence des structures de données,
- le contrôle de la conformité aux normes communautaires,
- et la vérification des fonctionnalités et des performances des dispositifs partagés.

L'approche paraît particulièrement intéressante pour l'interopérabilité de la plateforme et pour favoriser la réutilisation des données.

[16] https://nl.ambafrance.org/Ouverture-d-un-bureau-international-pour-l-initiative-GO-FAIR
[17] https://www.go-fair.org/
[18] http://www.enseignementsup-recherche.gouv.fr/cid124728/science-ouverte-la-france-rejoint-go-fair-en-tant-que-co-fondatrice.html
[19] Wilkinson, M. D.et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3, sdata201618+. http://dx.doi.org/10.1038/sdata.2016.18
[20] Cf. l'étude des pratiques, attitudes et attentes des directeurs d'unités du CNRS, Schöpfel, J., Ferrant, C., André, F., Fabre, R., 2018 (accepted). Research data management in the French National Research Center (CNRS). *Data Technologies and Applications* 52.
[21] Source : Wikimedia commons à commons.wikimedia.org
[22] *National Computational Infrastructure* (NCI), Evans, B.et al. 2017. A data quality strategy to enable FAIR, programmatic access across large, diverse data collections for high performance data analysis. *Informatics* 4 (4), 45+. http://dx.doi.org/10.3390/informatics4040045

Findable : les données doivent être (re)trouvables

La plateforme attribue un identifiant pérenne à chaque jeu de données, des *digital object identifiers* (DOI) en partenariat avec DataCite, comme EASY, Zenodo ou la plateforme canadienne, ou des *handle* (comme la plateforme tchèque). Aussi, pour chaque jeu de données, les métadonnées contiennent au moins un élément d'identification avec une URL exploitable basée sur un identifiant pérenne.

Le format des métadonnées est riche pour faciliter la description et la découverte des données. Le format est compatible avec le Dublin Core et avec les recommandations DataCite. La plateforme permet la création de liens avec d'autres jeux de données, avec des publications (DataONE Dash propose un moyen rapide de décrire les relations entre les données et les publications) ou des logiciels (par exemple utilisés pour nettoyer les données, calculer les résultats…). On peut ajouter des métadonnées bibliographiques. L'adaptation du format des métadonnées à des besoins et standards disciplinaires est possible sous forme d'un projet (partenariat). Les métadonnées sont librement accessibles sur la plateforme.

Accessible : les données doivent être accessibles et dans la mesure du possible en libre accès

Les données sont accessibles à partir de n'importe quel système d'exploitation (Windows, Linux, Mac OS) et peuvent être consultées à tout moment et de n'importe où. De plus, les données sont accessibles depuis n'importe quel terminal (smartphone, tablette), comme le proposent Zenodo ou Figshare.

L'accès libre n'est pas obligatoire pour toutes les données. Néanmoins, dans tous les cas de figure, les conditions d'accès sont clairement affichées (licence ouverte, durée d'embargo…), les groupes d'utilisateurs autorisés sont précisés en toute transparence, et l'identifiant permet leur récupération via un protocole standard de communication.

La plateforme peut conserver en toute sécurité les données selon quatre niveaux de confidentialité de l'information :

- Ouvert (à tous)
- Interne (disponible sur demande ou pour une communauté restreinte)
- Confidentiel (« closed access data », sans cryptage)
- Secret

La solution de stockage permet le partage de fichiers de données.

En particulier, la plateforme permet le moissonnage des métadonnées (OAI-PMH ; API ouverte, cf. REST). Figshare, par exemple, propose en option une interface avec ses propres API pour des collections ou communautés.

Interoperable : les données doivent être interopérables

La plateforme applique des normes pour assurer l'interopérabilité avec d'autres dispositifs. Ainsi, Zenodo respecte plusieurs normes internationales telles que DOI, JSON Schema, Memento, OAI-PMH, ORCID, OpenAIRE, REST, XML et le schéma de métadonnées de DataCite. DataONE Dash peut s'intégrer à n'importe quelle plateforme via SWORD, OAI-PMH ou le protocole standard ResourceSync (ANSI/NISO). Ceci concerne aussi la terminologie.

Reusable : les données doivent être réutilisables

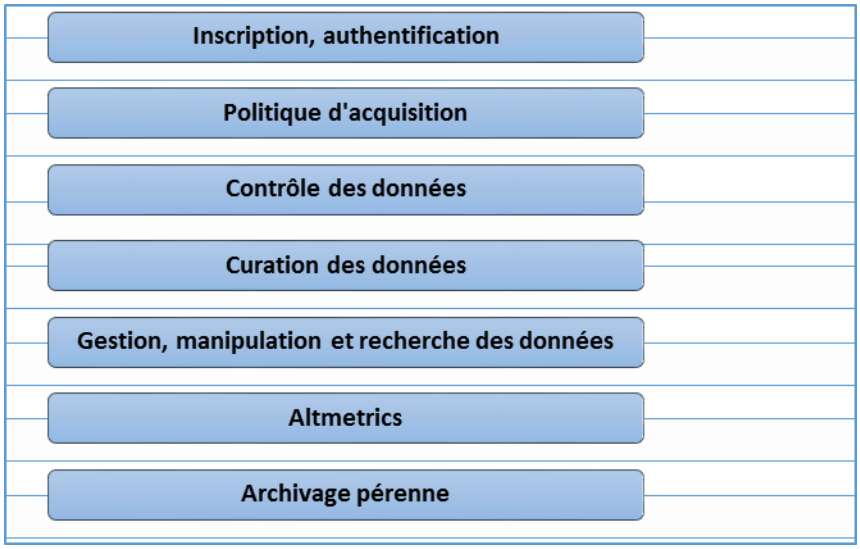
L'application des standards communs (normes…) contribue également à la réutilisation des données, tout comme leur description précise (métadonnées riches, documentation) et la diffusion sous licence ouverte, suivant des droits d'utilisation explicites.

Les droits et obligations des déposants et des opérateurs sont clairement énoncés, y compris les implications juridiques (transfert de droits, octroi de licences). La plateforme affiche la situation juridique des dépôts dans ses métadonnées. Elle propose plusieurs licences ouvertes, notamment les *Creative Commons* dont la licence domaine public CC0.

Autres caractéristiques

A part les fonctionnalités liées à la FAIRness d'un service de données, l'étude des dispositifs a dégagé sept autres domaines clés, utiles voire indispensables pour assurer la qualité d'une telle plateforme. Nous allons décrire les grandes lignes de ces fonctionnalités, à partir des exemples des dispositifs étudiés.

D'autres services et fonctionnalités d'une plateforme de données



Inscription, authentification

Le serveur définit clairement qui est autorisé à déposer et à gérer des données. Le site vérifie l'identité des déposants suivant une procédure d'authentification, qui fait le lien avec un répertoire national ou institutionnel de chercheurs et/ou d'institutions ou projets (Renater ou autre). Le système gère différents niveaux de droits d'accès (rôles), en particulier :

- utilisateur (chercheur dépositaire),
- administrateur,
- et conservateur (*curator*) de communauté (ou collection).

Grace à un identifiant auteur (ORCID ou autre), la plateforme peut aussi faire le lien entre les dépôts de données et les profils de chercheurs.

Ainsi, l'entrepôt de données de Cambridge est lié à leur système d'information recherche local (Symplectic). La plateforme canadienne FRDR externalise l'authentification à un partenaire et à un réseau d'institutions locales fédérées. DataONE Dash gère les accès via un compte institutionnel (Shibboleth) ou un compte Google.

L'authentification peut également inclure des projets ou des partenariats, et l'administrateur peut donner à d'autres individus (par exemple, ne faisant pas partie d'un annuaire) certains rôles avec des droits : EASY identifie les membres de l'ES néerlandais et les organismes de recherche via leur compte SURF. De plus, EASY gère des comptes liés à des partenariats ou à des projets (archéologie, Greynet...).

Zenodo permet une authentification flexible via ORCID, un compte GitHub ou une simple identification par email/mot de passe, sans aucun contrôle. Figshare gère des comptes individuels et institutionnels. Pour l'authentification institutionnelle, Figshare utilise les infrastructures d'authentification locales pour les identifiants de connexion institutionnels ; en parallèle, en tant qu'utilisateur individuel, tout le monde peut utiliser le service après l'enregistrement d'une adresse électronique (messagerie).

Politique d'acquisition

La plateforme mutualisée et multidisciplinaire met en œuvre une politique d'acquisition large et inclusive. Cette approche doit être clairement affichée. Les principaux critères de sélection sont de deux sortes :

- L'affiliation de l'auteur de données et/ou du dépositaire (authentification). La plateforme s'adresse avant tout aux communautés scientifiques, c'est-à-dire aux membres de l'enseignement supérieur et des organismes de recherche. Néanmoins, son positionnement peut inclure un public plus large, la société civile (ONG, science citoyenne, chercheurs amateurs…) ou les entreprises (recherche industrielle…).
- Le volume des données. Une limitation de la taille des fichiers déposés peut être envisagée selon plusieurs niveaux et en gérant les exceptions. Certaines plateformes facturent la prestation de stockage et d'archivage, en fonction de l'espace nécessaire, de l'usager et/ou de la durée de la conservation. Des exemples :
    - Canada : la solution permet de stocker jusqu'à 3 000 fichiers ou 300 Go par compte.
    - ASEP : l'entrepôt de l'Académie des Sciences à Prague accepte des jeux de données de 2 à 20 Go.
    - Figshare : l'espace est limité à 5 Go par fichier ou 20 Go par compte privé. Pour les institutions, Figshare propose un stockage gratuit de 100 To qui peut être alloué aux groupes et aux particuliers.
    - Cambridge : l'entrepôt accepte les fichiers de données individuels jusqu'à 2 Go et les jeux de données complets jusqu'à 20 Go. Les dépôts de jeux de données plus volumineux (> 20 Go) sont facturés (4 £ par Go).

La sélection de contenu (domaine de recherche, instruments scientifiques, typologie etc.) est gérée au niveau de la collection ou de la communauté (cf. Figshare ou EASY).

Comme l'entrepôt de Cambridge, elle accepte des logiciels (pour l'accès aux jeux de données, mais aussi pour leur analyse) ou gère les liens vers ces logiciels. En revanche, elle n'accepte pas nécessairement le dépôt de publications ou d'autres documents, contrairement à l'entrepôt de Cambridge ou Zenodo. Un lien suffit.

En principe, tous les formats de données sont acceptés. Alternativement, la plateforme indique clairement les formats préférés ou exclusivement acceptés[23].

Contrôle des données

D'une manière générale, les opérateurs de plateforme mettent tout en œuvre pour déplacer la responsabilité du dépôt et de l'utilisation des données vers les chercheurs et leurs institutions, en tant qu'auteur (dépositaire) et/ou usager.

La plateforme contrôle systématiquement les aspects formels (techniques) des données (exhaustivité, lisibilité des fichiers...). Elle ne contrôle pas la qualité des données ; cependant elle permet ce contrôle, soit au niveau local d'un établissement ou d'une institution (par un *data librarian* comme au Canada), soit par le responsable d'une collection (par un *community/collection manager* comme Zenodo).

Idem pour la conformité avec des normes ou règles éthiques ou disciplinaires : au lieu d'imposer ou d'exercer un tel contrôle, la plateforme exige l'engagement du chercheur que les jeux de données déposés sont créés, conservés, consultés et utilisés conformément aux normes disciplinaires et éthiques. Le contrôle peut également être formalisé en amont, au niveau local et/ou communautaire. En aval, la plateforme informe sur l'utilisation responsable des données sensibles (conditions d'utilisation).

A titre d'illustration, la plateforme canadienne repose sur un contrôle local de la conformité éthique (par un *data librarian* ou un comité d'éthique). L'entrepôt de Cambridge informe sur la responsabilité légale (FAQ) et exige le respect des directives institutionnelles. Figshare ne contrôle pas si les données déposées sont « sensibles » mais s'engage à les supprimer s'il y a un problème.

Le cryptage des données confidentielles ou sensibles pose un problème. La plateforme RADAR propose aux institutions une option sur demande pour le dépôt et la préservation à long terme des jeux de données cryptés, sans garantie pour leur décryptage. En d'autres termes, la clé de cryptage reste au niveau des institutions.

Curation des données

Pour favoriser le dépôt par les chercheurs eux-mêmes, la procédure de dépôt est (relativement) facile et rapide, avec des exigences minimales. Ainsi, DataONE Dash propose une indexation rapide pour les « scientifiques pressés » en seulement trois étapes, avec peu de champs obligatoires.

Pour les utilisateurs institutionnels, Zenodo et Figshare distribuent le contenu tel qu'il a été déposé, sans curation. Néanmoins, pour augmenter *findability* et *reusability* du dispositif, les plateformes proposent souvent une « curation de base », par exemple, par une vérification basique et par la possibilité de mettre à jour ou d'ajouter des métadonnées. De même, une documentation minimale des méthodes et techniques

[23] Comme .pdf .txt .sgm(l) .xml .jpg .tif .wav .shp (et autres formats ESRI) .csx .tab .nc (et .cdf) ou formats de fichiers ouverts.

utilisées pour collecter et analyser les données est obligatoire, tandis qu'une documentation étendue peut être ajoutée ultérieurement.

Gestion, manipulation et recherche des données

Sur le modèle de Figshare, la plateforme permet le dépôt de plusieurs versions du même jeu de données et fournira un système de contrôle de version. Ce *versioning* peut être déclenché par plusieurs actions, telle que :

- Modification du titre
- Ajout de nouveaux fichiers
- Suppression de fichiers
- Remplacement des fichiers
- Suppression de la confidentialité des fichiers
- Remplacement d'un lien associé à un élément

D'une manière très différente et plus simple, l'entrepôt de Cambridge attribue à chaque nouvelle version son propre identifiant et laisse le soin au dépositaire d'établir le lien.

Certaines plateformes intègrent des outils d'analyse et de manipulation de données (cf. plus loin, à propos de l'architecture du dispositif). C'est le cas pour EASY (avec la couche DataverseNL) et pour RADAR.

La plateforme propose des fonctionnalités de recherche simple et avancée uniquement pour les métadonnées. La recherche dans les jeux de données eux-mêmes n'est pas une priorité. Certaines plateformes y travaillent mais considèrent cette fonctionnalité plutôt comme une prestation des dispositifs disciplinaires, difficile à réaliser pour une plateforme multidisciplinaire et inclusive.

Altmetrics

La plateforme publie quelques « scores d'impact » pour chaque jeu de données, notamment le nombre de téléchargements. DataONE Dash, par exemple, affiche les téléchargements et vues, Cambridge permet un suivi des visites par mois, pays et ville.

La plateforme propose également des connexions vers les médias sociaux (comme Twitter, G +, ...) ou des outils scientifiques (outils de gestion de référence...), comme Zenodo par exemple qui propose l'exportation des références sous différents formats (MarcXML, DC, JSON, DataCite...).

L'intérêt est double : pour l'auteur et son établissement, produire des indicateurs d'impact pour chaque jeu de données ; et pour l'opérateur de la plateforme, contribuer à l'évaluation du retour sur investissement.

Archivage pérenne

La plateforme garantit l'archivage à long terme. L'archivage fait l'objet d'une certification. Le temps de conservation est d'au moins 10 ans. La conservation à long terme peut faire l'objet d'un contrat. L'archivage pérenne peut être externalisé.

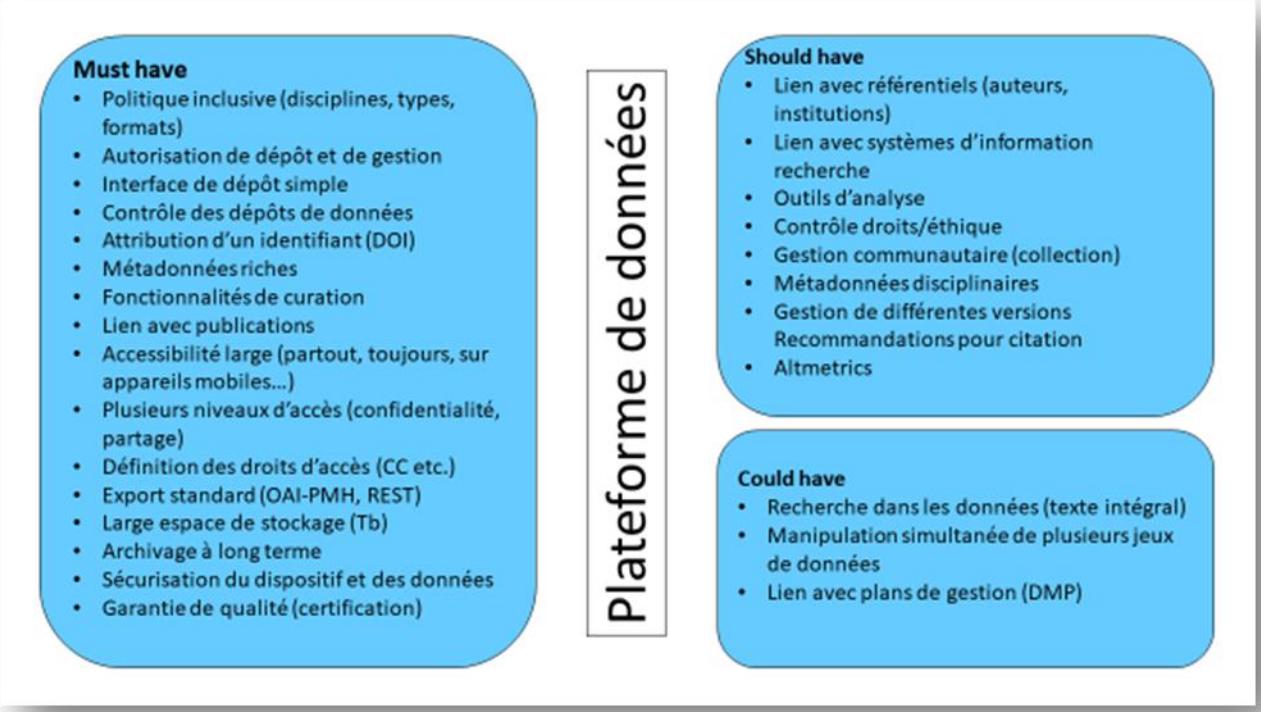
Quelques cas de figure :

- RADAR : s'il y a un accord de partenariat, la plateforme allemande garantit l'archivage pérenne pour au moins 25 ans.
- Figshare a signé des accords avec des éditeurs qui contiennent une garantie d'archivage pour 10 ans et propose une vérification de l'obsolescence du format des données.
- Zenodo et EASY garantissent l'archivage par leurs propres moyens.
- La plateforme canadienne assure la conservation avec des partenaires reconnus.

Une hiérarchie de fonctions et services

A partir des plateformes opérationnelles et en cours de montage, il est possible de tenter un classement des fonctionnalités et services d'une plateforme modèle en trois catégories, selon la méthode MoSCoW[24].

Priorisation des fonctionnalités et services



Il est certain qu'une telle hiérarchisation doit être ajustée aux priorités et contraintes du projet.

[24] Must have, Should have, Could have, Won't have (cette dernière catégorie ne s'applique pas ici).

Architecture

L'analyse des plateformes révèle surtout deux approches, une solution centralisée (Figshare, Zenodo, EASY) et une architecture fédérée ou distribuée (« network », Canada)[25]. Dans certains cas, la solution centralisée sépare les services de création, d'analyse, de traitement…, de la fonction de conservation et d'archivage (« bridge »), comme c'est le cas de la nouvelle architecture de la plateforme EASY avec, pour la gestion des données, la « couche » DataverseNL.

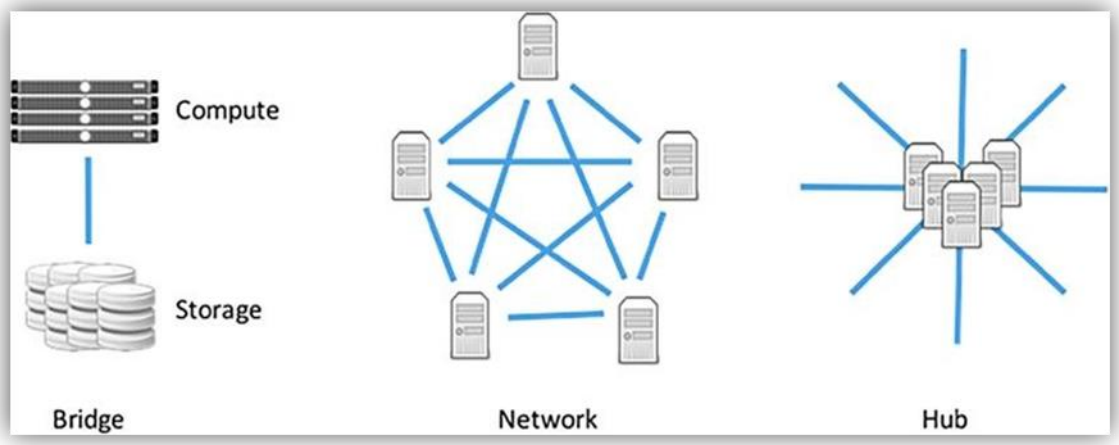
Nous n'avons pas trouvé de plateforme du type « hub », un modèle de service qui extrait continuellement des données en direct de plusieurs sources potentielles, effectue le traitement nécessaire, puis diffuse les informations traitées à un nombre potentiellement important de consommateurs de données.

D'une manière basique, un dispositif de données doit intégrer :

- Une solution de stockage temporaire (limitée, jusqu'à 5-10 ans) permettant la gestion des données de recherche (research data management au sens strict : dépôt, analyse, mise à jour, gestion des versions…).
- Une plateforme pour la conservation à plus long terme (plus de 10 ans), avec des fonctionnalités de diffusion et partage, des liens vers d'autres jeux de données ou de publications.
- Eventuellement une solution de back-up pour la conservation à long terme (*dark archive*).

L'idée qui soutient l'architecture de la plateforme RADAR est l'interconnexion des données avec les outils et technologies du web sémantique (RDF knowledge graphs, suivant le modèle de DBpedia), à l'opposé des silos de données[26].

Trois types de dispositifs[27]



[25] Cf. Goldstein, S., 2017. *The evolving landscape of federated research data infrastructures*. Knowledge Exchange, Bristol. http://www.knowledge-exchange.info/news/articles/29-11-2017

[26] Sören Auer, un des protagonistes allemands du web sémantique, vient d'être nommé à la tête de la TIB Hannover, membre du projet RADAR.

[27] Source : Xie, Z., Fox, E. A., 2017. Advancing library cyberinfrastructure for big data sharing and reuse. *Information Services & Use* 37 (3), 319-323 (p.321). http://dx.doi.org/10.3233/isu-170853

Ajoutons deux autres aspects, non spécifiques aux plateformes de données mais suffisamment significatifs pour être mis en avant par les opérateurs :

- **Le recours aux logiciels *open source*.** La California Digital Library (DataONE Dash), le CERN (Invenio/Zenodo), Cambridge (DSpace), DANS (FEDORA, DataverseNL), pour ne citer qu'eux, ont développé leurs services de données avec des logiciels *open source*. Pour le projet RADAR, le Fachinformationszentrum Karlsruhe (FIZ-K) a adapté une solution propriétaire, appelé « eSciDoc » qui avait été développé à partir de FEDORA (*open source*). « eSciDoc next generation », le système de RADAR n'est donc pas *open source* mais peut être diffusé à la demande, d'après les responsables du projet.
- **Le focus sur la sécurité.** Pour des raisons réglementaires (données à caractère personnel, données sensibles, données confidentielles…), mais aussi pour gagner la confiance des institutions et des chercheurs (*trustworthiness*), les opérateurs concentrent leurs efforts pour garantir la sécurité du dispositif et des données déposées. Cela implique une communication détaillée sur les mesures prises (sauvegardes quotidiennes des données et métadonnées, stockage distribué, contrôle et vérification des fichiers…), et cela signifie aussi, pour certains (dont Figshare), un investissement ciblé pour obtenir une certification[28]. Le système canadien repose sur des mesures de sécurité fédérées.

Organisation et gouvernance

Dans un souci de transparence et d'efficacité, l'opérateur formule une politique qui contient :

- La définition des droits et obligations du prestataire de services.
- La description du service et des déclarations détaillant pour qui et dans quelles conditions il est fourni.
- La définition des droits et obligations des auteurs ou contributeurs, y compris pour l'autorisation d'une mise à jour des données déposées.
- Une description des types de contenus publiés via le service, ainsi que des exigences relatives au contenu et à la qualité technique des données.

Certains opérateurs, comme l'Académie des Sciences de la République Tchèque, vont plus loin, avec des actions de formation, d'information, de sensibilisation et de communication à destination des chercheurs, gestionnaires, informaticiens, bibliothécaires…

Au niveau du pilotage et de la coordination, la gouvernance politique (pilotage) et opérationnelle (gestion) sont clairement séparées :

- La plateforme dispose d'un comité de pilotage (conseil d'administration) représentant l'autorité politique et administrative.
- Elle dispose également d'un comité consultatif (conseil scientifique) représentant les principaux acteurs et partenaires (bibliothèques, universités, organismes de recherche, organismes de financement, …), y compris des experts internationaux.

[28] ISO 27001 sécurité des SI, Core Trust Seal, Data Seal of Approval etc.

De même, les responsabilités techniques et administratives sont définies et affichées : qui héberge le dispositif ? qui exploite le dispositif ?

Hébergeur et opérateur (responsable d'exploitation) ne sont pas nécessairement identiques. Une solution du type *Software as a Service* (SaaS) est proposée par la plateforme RADAR. Dans le même temps, le projet RADAR distingue très clairement la fonction technique et développement, y compris la gestion des contrats (opérateur FIZ Karlsruhe), et la fonction marketing et conseil/assistance (TIB Hannover).

Modèle économique

Dans la mesure où « *Data Alpha* » est positionné au sein des infrastructures de la recherche publique, comme la plupart des dispositifs étudiés, son modèle économique dépendra nécessairement d'un financement public. Il faut distinguer le financement du budget de fonctionnement opérationnel (« durable ») et l'investissement pour tout futur développement.

L'étude des plateformes montre qu'il n'y a pas de modèle unique et que le flux des recettes est souvent diversifié, avec au moins cinq options :

1. **Projet.** Il s'agit d'un financement limité dans le temps, sur appel d'offres ou négocié, dans le cadre d'un projet, d'après un cahier des charges plus ou moins précis. L'objectif peut aussi bien être la mise en place d'un nouveau dispositif (RADAR, Zenodo ou DataONE Dash), l'amélioration des performances ou le développement de nouveaux services. L'origine des financements sur projet est variable, aussi bien publique (administrations, organismes publics, agences de moyens, programmes de recherche ou d'infrastructures…) que privée, comme la Fondation Alfred P. Sloan.

2. **Abonnement.** L'utilisation de la plateforme (administration, dépôt, stockage, conservation...) par une institution ou un particulier nécessite l'acquittement d'une cotisation annuelle (Figshare). RADAR annonce un tarif de 500 € pour un espace de stockage jusqu'à 100 Go et frais administratifs, pour une durée de 6 mois.

3. **Services payants.** Certains services sont facturés, forfaitairement ou en fonction de l'activité réelle (taille de stockage ou de conservation comme pour la plateforme de l'Université de Cambridge, nombre d'utilisateurs autorisés...). RADAR par exemple facture la garantie d'un archivage de 25 ans minimum par Go. D'autres exemples : le *Archaeology Data Service* (ADS, Université de York) gère les données de l'English Heritage et Historic Scotland moyennant des frais, et le *European Bioinformatics Institute* (EBI, Wellcome Trust Genome Campus) a développé une offre de service pour les données des entreprises et de la recherche privée. Dans ces cas, l'objectif de générer des revenus est lié à la volonté de s'assurer que le service reste entièrement accessible[29].

[29] Collins, E., 2012. The national data centres. In: Pryor, G. (Ed.), *Managing research data*. Facet Publishing, Croydon, pp. 151-172.

4. **Partenariats stratégiques.** Plusieurs plateformes ont conclu des partenariats avec d'autres organismes pour le développement de services, la création d'infrastructures ou la préservation à long terme, avec une mutualisation des dépenses.

5. **Activité commerciale.** Plusieurs plateformes (et pas seulement les plateformes privées) vendent certaines prestations à des clients privées (éditeurs, sociétés commerciales, entreprises…), en dehors de leur public cible de l'ESR public. Plusieurs services sont concernés, surtout l'espace de stockage et de conservation, l'utilisation du logiciel en tant que service, la vente du logiciel sous licence. Le système Figshare peut être concédé sous licence à des éditeurs ou des institutions, afin de fournir des fonctionnalités et des services plus nombreux et de meilleure qualité, comme par exemple le contrôle de la qualité des données et métadonnées.

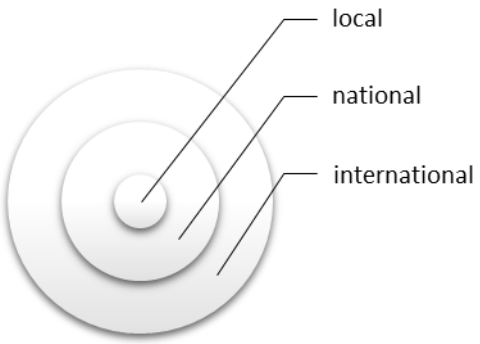
Figshare fournit des services payants (édition, conservation, serveur preprint) à plus de 50 clients : organismes, établissements, éditeurs, agences de moyens, entreprises. L'opérateur de Figshare peut vendre le workflow de la plateforme sans stocker les données sur leur propre serveur (solution de stockage locale). Il ne reçoit pas de subvention et affirme ne jamais avoir limité l'accès aux contenus et ne pas vouloir générer de revenus à partir des données déposées sur la plateforme.

Une telle activité commerciale n'est guère possible si la mission de la plateforme est clairement définie comme « non lucrative », liée au financement public comme pour RADAR. Mais même dans ce cas, la restriction peut s'avérer transitoire et la situation peut évoluer, pour diversifier les recettes et assurer un financement durable ; les opérateurs de RADAR ont déjà reçu des demandes de la part d'organismes danois et néerlandais et n'excluent pas une approche plus »commerciale » à l'avenir. Car le principal défi de tous les modèles économiques est la durabilité au-delà de la période du projet, avec en plus le besoin d'une équipe R & D, réactive et flexible, capable de répondre aux appels à projet.

Périmètre

En fonction de son positionnement et rattachement institutionnel, chaque plateforme a un périmètre caractéristique qu'on peut décrire à trois niveaux, local, national et international. A partir des différents cas de figures, on peut formuler plusieurs bonnes pratiques pour une plateforme mutualisée.

Périmètres d'une plateforme



La plateforme doit être entièrement intégrée dans le paysage national des services de données existants. Elle doit également être capable de communiquer avec les systèmes de recherche locaux. L'approche adéquate sera un dispositif distribué (« federated research data infrastructure »[30]).

Les opérateurs de RADAR insistent sur un point précis : qu'ils ne se positionnent pas en concurrence avec d'autres projets nationaux, en particulier avec celui d'un cahier de laboratoire électronique (ELN) financé par la DFG et porté par l'un des partenaires de RADAR, le KIT.

Le projet américain de la California Digital Library montre la possibilité d'une toute autre forme d'intégration : il ouvre la plateforme aux données de la société civile, comme « open public data repository », pour la science citoyenne, en dernier ressort.

Si la plateforme conclut des accords avec des institutions et établissements, elle doit créer et entretenir un réseau de correspondants. Parmi les dispositifs étudiés, seul le FRDR canadien dispose d'un réseau de contacts (*local data librarians*) dans les établissements membres du dispositif fédéré.

Sur le modèle de RADAR ou EASY, elle devrait s'insérer dans le paysage international de projets, réseaux et organisations, pour différentes raisons (diversification des sources de financement, partenariats, développement mutualisé, échange d'expériences…).

En option, la plateforme pourrait également coopérer avec une ou plusieurs revues de données (*data journals*), afin de faciliter l'exposition et l'impact des données. Deux exemples :

- L'opérateur DANS a conclu un partenariat avec l'éditeur néerlandais Brill pour alimenter le *Research Data Journal for the Humanities and Social Sciences* à partir de l'entrepôt EASY (le directeur de DANS est co-éditeur de la revue).
- DataONE Dash peut éditer des *data papers* génériques en PDF, à partir des métadonnées.

[30] Cf. Goldstein, S., 2017. *The evolving landscape of federated research data infrastructures*. Knowledge Exchange, Bristol. http://www.knowledge-exchange.info/news/articles/29-11-2017

Partie 2 : Un environnement dynamique

Introduction

La réalisation du projet allemand RADAR a pris quatre ans et, durant cette période, la nature du projet a profondément changé : l'entrepôt de données du premier cahier des charges, proposant des fonctionnalités de conservation et d'exposition, est devenu une plateforme plus riche et plus complexe. Le projet français n'a pas la même temporalité : tous les experts conviennent qu'il y a urgence.

En fait, ce constat d'urgence implique deux conséquences : pour le dire simplement, soit « Data Alpha » est mis en chantier maintenant pour être opérationnel en 2019, soit le projet n'a probablement plus de sens. Pour des raisons juridiques et réglementaires, mais aussi pour satisfaire les exigences des grands programmes scientifiques et des éditeurs, les chercheurs et leurs institutions ont besoin d'une solution maintenant, pour la conservation et la diffusion de leurs données.

L'absence d'une solution nationale incitera la communauté scientifique à chercher des alternatives, en développant des dispositifs locaux, en adoptant des infrastructures internationales et/ou en achetant des produits commerciaux. Les enjeux sont multiples : la qualité des outils et services, les coûts (investissements, achats, licences…), le contrôle des données, leur sécurité, et la perte potentielle de compétences et d'expertise en cas d'externalisation de la gestion des données.

Commentaire des commanditaires du rapport :

La phase de structuration des données de recherche de l'ESR s'inscrit à l'échelle mondiale, amenant de grands groupes à se positionner pour offrir leurs services. Toutefois, l'absence d'alternative aux propositions de ces nouveaux opérateurs risque de conforter leur place et, à l'instar des revues, voir la propriété intellectuelle des travaux et les usages des chercheurs migrer vers eux. C'est un enjeu de souveraineté scientifique. Les données appartiennent aux établissements financeurs et les chercheurs ne devraient pas avoir le droit de céder la propriété intellectuelle de ces données.

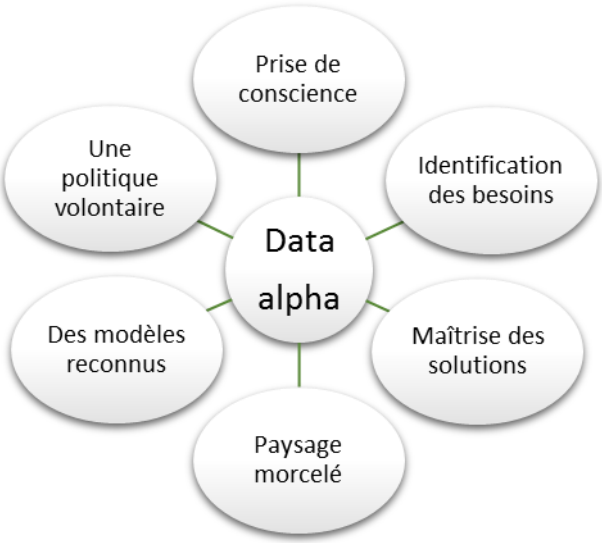
A partir de nos entretiens et études, nous allons d'abord dégager quelques facteurs favorables à la réalisation d'une plateforme mutualisée pluridisciplinaire de stockage, de gestion, de signalement et de partage de données de recherche, avant d'identifier les principaux verrous et risques du projet. Cependant il faut tenir compte du dynamisme de la situation. Les conditions changent rapidement, et l'analyse n'a qu'une portée limitée.

Mais cela est vrai aussi pour la solution, et c'est la deuxième conséquence : le constat d'urgence doit admettre qu'une solution, quelle qu'elle soit, ne pourra jamais être que transitoire (pas provisoire !) et qu'elle doit avant tout rester flexible et ouverte aux développements à venir. Ce qui, paradoxalement, devrait faciliter la gestion et la réalisation du projet en limitant les coûts afférents.

Une dynamique favorable et des opportunités

Actuellement, plusieurs variables convergent en faveur du lancement d'un projet « Data Alpha ». Nous avons identifié six facteurs dynamiques dont le potentiel facilite la prise de décision et la mise en chantier d'une future plateforme nationale.

Points forts et opportunités



Une prise de conscience de la nécessité d'agir

Le sujet est d'actualité et suscite de l'intérêt. Au sein des universités, la question des données de la recherche « *excite tout le monde en ce moment qui brasse plein d'idées pas toujours très pertinentes* ». De leur côté, les organismes de recherche (se) posent également la question, sans toujours aller plus loin. Il y a au moins trois raisons pour cet intérêt :

- **Les programmes scientifiques.** Les exigences des financeurs (Commission Européenne etc.) incitent les chercheurs à demander un endroit pour déposer leurs données avec un DOI. « *Avec le programme H2020 et le Open Data Pilot, les chercheurs découvrent leurs besoins* ». Sans doute, l'ANR, sur le modèle d'autres agences, suivra le même chemin et conditionnera tôt ou tard le financement de la recherche à l'existence d'un plan de gestion des données et, dans la mesure du possible, d'une mise à disposition des résultats.
- **Les conditions des éditeurs.** De plus en plus souvent, les éditeurs scientifiques demandent aujourd'hui, au moins pendant la période de l'évaluation d'un manuscrit, l'adresse ou l'identifiant pour accéder aux résultats décrits. Parfois, la publication d'un article est conditionnée par l'accessibilité aux données, afin d'en assurer la vérification, la réplication et la réutilisation.
- **La protection des données.** Le renforcement des droits des personnes, y compris pour la recherche impliquant la personne humaine dans le domaine de la santé, demande une attention particulière par rapport à la sécurité des systèmes d'information et des données. Les solutions ad hoc et l'environnement informatique des établissements ne sont pas (plus) toujours à la hauteur des nouvelles contraintes et menaces.

C'est avant tout ce dernier aspect qui oblige les chercheurs et leurs organisations à chercher des solutions pour le stockage et la gestion de leurs données dans un environnement sécurisé. « *Leur problème est surtout la sécurité et les aspects juridiques, les données personnelles, les données sensibles ou confidentielles, notamment quand il s'agit de grands volumes* ».

Google Drive n'est pas une option mais reste fortement utilisé. Formulé par le responsable d'une mission recherche au sein d'un SCD, cela donne : « *En matière de diffusion, le cas des données dites à caractère personnel pose un problème particulier. La spécificité de ces données c'est qu'elles peuvent faire l'objet d'une dérogation en matière de diffusion uniquement si elles sont réutilisées à des fins de recherche scientifique ou archivistique. En général, comme aucune solution n'est facile à mettre en œuvre ces données ne sont pas diffusées et rarement archivées, ce qui est un problème pour les chercheurs pour lesquels c'est le type de données principal* ».

Il s'agit de « *données exigeantes* » en matière de services associés, dont les solutions locales (institutionnelles) paraissent difficiles, voire irréalistes. Le projet d'un RENATER Drive est considéré comme allant dans le bons sens, sans toutefois répondre à toutes les questions.

Un autre argument en faveur d'une solution mutualisée a été avancé par le premier directeur de la TGIR Huma-Num, Marc Renneville, devant l'Assemblée Nationale en 2016[31] : « *(…) je voudrais (…) souligner combien le développement de structure permettant une gestion mutualisée des données peut aussi avoir un effet bénéfique à l'évolution de nos savoirs et au décloisonnement de nos disciplines. L'incidence de la diffusion de ces solutions de gestion co-élaborée avec les chercheurs, c'est qu'elle stimule le travail collectif et nous permet de nous concentrer sur l'élaboration de nouveaux outils de publication, de visualisation et de diffusion de l'information scientifique.* »

L'identification d'un faisceau de besoins sans réponse

Sur le terrain, cette prise de conscience se traduit par la description d'un ensemble de besoins qui à ce jour restent trop souvent sans réponse adéquate. Plusieurs enquêtes[32] et études ont contribué à reconnaître l'existence de « *chercheurs sans outils* » (par exemple pour faire du TDM en sciences humaines), de « *petites communautés* » (notamment transdisciplinaires, sans réelles réflexions ou pratiques concernant la gestion des données) et d'une « *longue traîne de données* ».

Certains résultats de ces études sont paradoxaux. D'une part, les chercheurs sont globalement assez favorables au partage de leurs données, avec une préférence pour des entrepôts multidisciplinaires (nationaux ou internationaux) et des plateformes spécialisées[33]. Mais ils connaissent assez mal les infrastructures nationales qu'ils utilisent (trop) peu.

[31] Renneville, M., 2016. Sciences technologiques et sciences humaines et sociales : les enjeux d'une gestion mutualisée des données de la recherche. Note pour l'audition OPECST (office parlementaire d'évaluation des choix scientifiques et technologique). In: *Assemblée nationale. Séance du 21 janvier 2016 consacrée aux sciences humaines et sciences technologiques*, Paris. https://halshs.archives-ouvertes.fr/halshs-01452949

[32] Comme à Strasbourg, Lille et Rennes, cf. Serres et al. op.cit.

[33] Prost, H., Schöpfel, J., 2015. *Les données de la recherche en SHS. Une enquête à l'Université de Lille 3. Rapport final.* Université de Lille 3, Villeneuve d'Ascq. http://hal.univ-lille3.fr/hal-01198379v1 Voir aussi Serres et al. op.cit. ; Schöpfel et al. op.cit.

Ceci étant, leurs demandes portent avant tout sur l'environnement de travail performant, proposant des espaces de travail collaboratif et de stockage sécurisés, afin de protéger les données scientifiques (« *données chaudes* »), ainsi que sur l'archivage pérenne des résultats, une fois le projet de recherche terminé (« *données froides* »).

« Data Alpha » apporterait des réponses à plusieurs besoins, dont notamment :

- un espace de stockage et de conservation,
- la création de métadonnées,
- l'attribution d'un identifiant pérenne (DOI),
- l'export vers d'autres plateformes,
- l'ouverture (partage, publication) des données,
- la gestion de la confidentialité de certaines données (accès restreint ou sur demande, avec authentification et engagement de respecter les conditions d'une réutilisation),
- une gestion communautaire,
- la connexion avec les SI des établissements et organismes,
- l'interconnexion des réservoirs de données (« décloisonner des silos de données », « qu'on puisse faire communiquer les données »),

  Commentaire des commanditaires du rapport :

  L'Inist et Huma-Num pourraient travailler ensemble pour couvrir l'ensemble des champs disciplinaires et associer les compétences en lien avec le CINES, notamment.

- la mise à disposition d'outils « *très concrets* » pour la gestion et l'analyse des données.

Ajoutons qu'une telle plateforme apporterait une réponse aussi à trois besoins exprimés par les professionnels de l'information dans les SCD et laboratoires : le besoin de pouvoir orienter « *leurs chercheurs* » vers « *une solution par défaut* », en absence d'une option adéquate (disciplinaire, spécifique…) ; le besoin de pouvoir remplir la fonction de médiation des infrastructures nationales[34] ; et le besoin de s'approprier des outils de gestion car en ce moment « *on n'a que des outils externes* ».

La maîtrise des solutions techniques

L'étude des plateformes et la comparaison avec les besoins des chercheurs et professionnels confirme un fait qui devrait s'avérer propice à la réalisation d'un projet national : les solutions techniques d'un tel dispositif semblent globalement maîtrisées. Il ne s'agit d'inventer une solution novatrice ou originale, mais de construire un dispositif approprié à partir d'éléments opérationnels.

Concrètement, le développement de « Data Alpha » peut faire appel à :

[34]Cf. l'expérience du correspondant Huma-Num à Nice http://adbu.fr/demi-journee-detude-adburesearchdata-presentations-des-intervenants/

- Des modèles connus, réputés et utilisés (Zenodo, Figshare, EASY, Dataverse, Huma-Num).
- Des architectures éprouvées (EASY avec DataverseNL, la plateforme RADAR).
- Des logiciels non propriétaires (Zenodo, Dataverse mais aussi, dans une certaine mesure, RADAR), avec leurs communautés de développeurs et opérateurs.
- Des standards et normes acceptés (pour les métadonnées, les protocoles d'échange, la sécurité d'un SI).
- Des critères explicites et acceptés pour l'évaluation du produit (principes FAIR, sécurité, certifications).

Nos entretiens ont révélé une certaine lucidité sur un point précis : « *la question n'est pas forcément technique* », « *on connaît les systèmes* », « *on sait ce qu'il faut faire* », « *on a les éléments nécessaires pour construire une offre cohérente, lisible, sérieuse (trustworthy), sécurisée et facile à prendre en main* » et « *au pire, il y a des solutions sur le marché* ». Quant à l'archivage pérenne, « *il y a le CINES* », partenaire incontournable pour toute question relative à la conservation des données de l'ESR français. La technique n'est pas le problème majeur.

C'est d'autant plus vrai selon les mots de Peter Hinssen, l'auteur du livre « The New Normal » et expert de l'économie numérique, le consommateur (ici : chercheur) ne veut pas obligatoirement la meilleure solution mais un produit « just good enough », fiable et rapide[35]. Récemment, la Confederation of Open Access Repositories (COAR) a fait remarquer que la tendance générale des entrepôts de publications mais aussi de données allait vers davantage de « inclusiveness and diversity », et non vers la spécialisation[36]. Un autre argument qui joue en faveur d'une plateforme mutualisée.

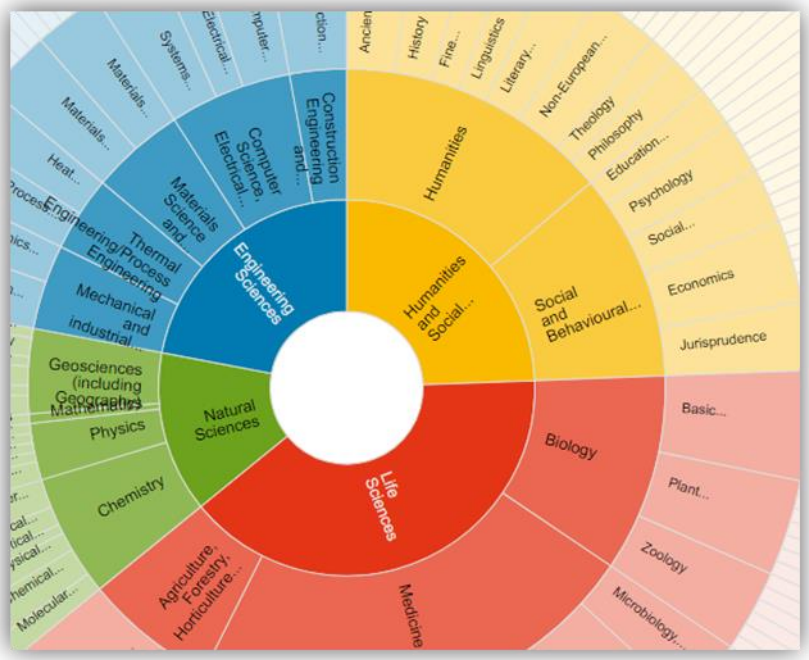
Le constat d'un paysage morcelé, incomplet

La tendance vers une approche inclusive et la diversité des contenus est confortée par le constat déjà évoqué de l'existence de « *chercheurs sans outils* ». Interrogés, nos interlocuteurs soulignent sans exception l'intérêt pour des outils, des dispositifs et des infrastructures disciplinaires, proposant des formats, standards et procédures spécifiques. Force est de constater, cependant, que la réalité est loin de satisfaire ce besoin.

Une analyse rapide des sites répertoriés par le CNRS[37] révèle une toute autre situation, une sorte de « *patchwork* » de services de données, plateformes, entrepôts et autre outils confondus. Certaines disciplines ont une offre de services plus ou moins développée, au moins pour certaines thématiques et instruments, y compris au niveau national (sciences humaines et sociales, sciences de l'univers, biologie, physique, chimie), tandis que d'autres ne disposent pas d'outils de dépôt ou de gestion. Ici, l'entrepôt par défaut est le plus souvent Zenodo, la plateforme du CERN avec une vocation européenne.

[35] Hinssen, P., 2010. *The new normal: explore the limits of the digital world*. Across Technology, Gent.
[36] COAR, 2017. *Next generation repositories. Behaviours and technical recommendations*. COAR Next Generation Repositories Working Group, Göttingen. https://www.coar-repositories.org/activities/advocacy-leadership/working-group-next-generation-repositories/
[37] Cat OPIDoR, wiki des services dédiés aux données de la recherche https://cat.opidor.fr

Quelques disciplines du répertoire re3data[38]



Ce constat n'est pas spécifiquement français et n'est pas le reflet d'une cartographie incomplète du paysage français. La comparaison avec le répertoire international re3data et ses plus de 2 000 entrepôts et bases de données disciplinaires et institutionnelles confirme ce constat d'un « *patchwork* », à l'échelle internationale[39]. Malgré une offre dynamique et malgré un nombre croissant de plateformes et autres dispositifs, il reste beaucoup de lacunes. Trouver des entrepôts adaptés pour certaines disciplines et thématiques est difficile, et il existe toujours une longue traîne de données, d'équipes et de sujets sans services. Une plateforme mutualisée, multidisciplinaire apporterait une réponse, au moins pendant une période transitoire, le temps que chaque communauté s'organise davantage.

L'existence de modèles reconnus mais limités

L'intérêt de « *Data Alpha* » se trouve conforté par un cinquième facteur, quelque peu paradoxal. Plus haut, nous avons souligné l'existence de modèles opérationnels et reconnus comme variable de faisabilité. Néanmoins, chaque réalisation a ses limites, et l'usage (ou non-usage) de ses différents dispositifs fait naître et renforce la demande explicite d'une alternative.

Plusieurs interlocuteurs s'interrogent sur les garanties et mesures de sécurité des entrepôts et font remarquer que certains « *ne répondent clairement pas à ces enjeux* ». D'autres outils ne s'adressent pas aux chercheurs individuels, demandent beaucoup (trop) d'effort et de temps pour s'en servir ou sont limités à quelques appels à projets, ce qui restreint le périmètre.

[38] Registry of research data repositories https://www.re3data.org
[39] Voir aussi Kindling, M. et al., 2017. The landscape of research data repositories in 2015: A re3data analysis. *D-Lib Magazine* 23 (3/4). http://www.dlib.org/dlib/march17/kindling/03kindling.html

Une autre limitation est disciplinaire : Dataverse de l'Université Harvard par exemple a beaucoup de points forts (interopérabilité, faible coût, attribution de DOI, expérimentation, exposition) mais un inconvénient : le système s'adresse avant tout aux sciences sociales[40].

Un argument revient régulièrement : le fait que certaines plateformes soient hébergées et administrées à l'étranger. Certains souhaitent une solution publique française pour l'indépendance nationale en matière de données, d'autres observent une « *crainte larvée de confier ses données à des plateformes étrangères* ».

Cet argument n'enlève rien à la reconnaissance de la qualité des sites étrangers : « *Si on doit jouer au jeu des modèles étrangers, le UK data service[41] me paraitrait un bon exemple (…)* ». Mais cette reconnaissance est suivie d'une mise en garde, notamment par rapport à la sécurité des données : « *Un service national serait aussi plus facile à faire identifier par nos CIL comme une solution adéquate, ce qui est un enjeu non négligeable* ».

Face à ces limitations, une plateforme nationale pourrait proposer une solution, si quatre conditions sont respectées : la garantie d'une sécurité suffisante, une politique d'acquisition large, transdisciplinaire (inclusive), la prise en compte des besoins des chercheurs, et l'hébergement en France, sous contrôle public.

Une politique volontaire en faveur de la Science ouverte

Le dernier facteur favorable au projet d'une plateforme nationale est d'ordre politique. Il s'agit d'une part de l'initiative Open Science du MESRI, coordonnée par Marin Dacos. « *Il faut construire une politique de données autour des établissements (…) Les données appartiennent aux établissements et pas aux chercheurs (qui ne peuvent donc pas en disposer et les céder).* [Marin Dacos] *suggère, sur la proposition d'Henri Verdier, d'avoir un administrateur des données dans les établissements. La loi rend obligatoire la création d'un guichet qui saurait où se trouvent les données.* »[42]

Cette initiative fait écho et prolonge la politique de l'ouverture des données de la France et des engagements pris au plan international dans le cadre de l'Open Government Partnership, thème Science and Technology.

A condition de répondre aux besoins, d'apporter une contribution aux engagements pris et de s'insérer dans la politique nationale et internationale, une plateforme nationale de données trouvera un environnement favorable.

Au plan européen, l'espace d'infrastructures et de recherche est actuellement marqué par la mise en place de l'European Open Science Cloud (EOSC) et par la concertation étroite entre plusieurs pays leaders, afin de soutenir « *implementation networks and support research and e-infrastructure communities in their endeavours to provide FAIR data and services* (…) »[43].

[40] Cf. le retour d'expérience par Sophie Forcadell, Sciences Po Paris, 21 novembre 2017 http://adbu.fr/demi-journee-detude-adburesearchdata-presentations-des-intervenants/
[41] http://www.data-archive.ac.uk/
[42] Réunion des CorIST du 19 octobre 2017
[43] https://nl.ambafrance.org/Ouverture-d-un-bureau-international-pour-l-initiative-GO-FAIR

D'autres initiatives vont dans le même sens et offrent un cadre favorable : des opportunités de partenariats et de financement, comme GO FAIR ou la Open Science initiative d'OpenAIRE,[44] le développement d'un dispositif « Open Science as a Service » (oSaaS) « on top of the OpenAIRE infrastructure » qui implique, au niveau national, la mise en place d'e-infrastructures horizontales[45].

Des verrous et des risques

Cependant, les entretiens ont également permis d'identifier un certain nombre de verrous et problèmes potentiels qu'il convient d'anticiper. Régulièrement, nos interlocuteurs, après avoir confirmé le besoin des chercheurs et exprimé leur intérêt ont émis quelques réserves, comme par exemple qu'il ne fallait pas sous-estimer « *les risques et obstacles, notamment le manque de ressources humaines au sein de l'ESR (correspondants locaux) et l'absence de mobilisation de la part des chercheurs à qui il faudrait faire comprendre l'avantage* ».

Commentaire des commanditaires du rapport :

De plus, comme le montrent les analyses de l'alliance RDA, les barrières de partage des données sont plus souvent d'ordre social que technique.

Quels sont les risques liés à la mise en place d'une plateforme nationale mutualisée pluridisciplinaire dédiée au stockage, à la gestion, au signalement, au partage et à la diffusion de données de recherche ? Notre analyse aura identifié six faiblesses (facteurs internes) et obstacles (facteurs externes) d'un tel projet.

Verrous et obstacles

Disparité des situations et besoins

La demande d'une plateforme mutualisée et pluridisciplinaire pour les données de la recherche n'est pas portée par une communauté, un métier ou un lobby bien

[44] https://blogs.openaire.eu/?p=2602
[45] Cf. la présentation du pilote EOSC par Simone Sacchi https://zenodo.org/record/1045812#.WjP6ft_ibIU

identifié. Au contraire, elle reflète une grande disparité des situations, pratiques et besoins.

D'une certaine manière, la demande d'une plateforme correspond à une sorte de plus grand dénominateur commun des intérêts et choix multiples d'un ensemble hétérogène de communautés scientifiques et professionnelles.

La préférence d'une approche disciplinaire

Les répertoires re3data et Cat OPIDoR exposent une évidence : les plateformes et outils privilégiés sont des émanations de communautés scientifiques, construits autour d'un grand instrument, d'une infrastructure, d'une méthode, d'une discipline ou thématique. De cette évidence à dire que toute solution doit être disciplinaire, il n'y a qu'un pas.

L'observation revient régulièrement : « *La solution pour déposer les données doit être disciplinaire, il faut mobiliser les compétences disciplinaires, identifier et soutenir les plateformes disciplinaires, au niveau national et international* ». L'argument met l'accent sur les particularités disciplinaires, communautaires et institutionnelles.

Un outil comme Zenodo ne suffirait pas, car pas assez disciplinaire, notamment par rapport aux métadonnées et à la documentation : « *Une structure générique est un cul-de-sac car non adaptée aux besoins de réutilisation* ».

Un deuxième argument est lié : l'exigence de maîtriser l'ensemble des étapes du cycle de vie des données. Autrement dit, « *il faut développer une (des) solution(s) plus large(s) qu'un simple entrepôt* », avec des outils de production, d'analyse etc. qui seront nécessairement spécifiques à chaque communauté : « *On a besoin de dominer notre pipeline* ». Une telle approche, aussi justifiée soit elle dans l'absolu, risque de constituer un verrou au lancement d'une plateforme générique.

Un troisième point a été avancé : « *La longue traîne en France, c'est surtout en SHS. Or, il existe déjà une infrastructure SHS. Il suffit peut-être de développer une interface utilisateur et d'augmenter les capacités du TGIR ?* »

De nouveau, l'argument mêle du vrai et du faux. Du vrai, puisqu'il convient effectivement de renforcer les infrastructures existantes. Du faux, puisque de toute évidence, la question d'un entrepôt par défaut ne se pose pas uniquement dans les SHS. Les contenus et communautés de Zenodo le prouvent, si besoin.

Citons trois exemples de besoins qui seront difficiles à prendre en charge par une plateforme générique :

- Le contrôle des usages des données, notamment pour des données sensibles, protégées (comme en médecine).
- Le traitement et l'extraction des données par l'opérateur lui-même (comme pour les centres de données cliniques).
- Le dépôt et la conservation de données complexes (« données 3D »).

Refus d'une solution imposée, rigide

Une autre remarque revient souvent quand on pose la question de l'intérêt d'un nouveau dispositif : il ne faut surtout pas une solution imposée, rigide, (trop) centralisée, sans articulation avec les communautés et les outils sur le terrain. Tous

les interlocuteurs insistent sur la nécessaire souplesse et adaptabilité d'une plateforme nationale. Cette souplesse va bien plus loin qu'une simple personnalisation, dans la mesure où elle inclut les aspects de responsabilité, de contrôle et de gouvernance : « *Toutefois, si une solution nationale devait émerger, (il serait souhaitable) qu'elle se traduise par une plateforme technique, avec les services ad hoc, où chaque institution a la totale maîtrise d'ouvrage de son espace* ».

Commentaire des commanditaires du rapport :

Il est nécessaire de s'appuyer sur l'existant, et non de créer de nouvelles infrastructures (y compris à l'étranger). Il faut avoir un impact sur les services et s'appuyer sur des initiatives soutenues par des communautés dynamiques.

Cette demande est particulièrement forte là où l'établissement poursuit une politique d'archive ouverte institutionnelle locale, indépendante mais interconnectée avec l'infrastructure nationale (Strasbourg, Bordeaux, Lille…). On peut distinguer plusieurs craintes :

- Une perte de contrôle sur « ses » données (description, gestion, préservation…).
- Un manque de souplesse et de facilité d'usage, y compris pour le dépôt des données.
- Des problèmes avec l'articulation de la plateforme avec les systèmes locaux (workflows).
- L'absence d'une gestion communautaire, ou le manque de droits d'administration.
- La crainte d'une relation « prestataire de service – client » avec l'opérateur de la plateforme, au lieu d'une relation de partenariat.

Ces craintes se nourrissent d'expériences du passé. Après avoir insisté sur la nécessité d'une installation nationale et de son articulation avec le système d'information local, l'un des interlocuteurs a ajouté que « *l'absence de souplesse, une trop grande rigidité de gestion et l'absence d'une gestion communautaire sont les principaux obstacles pour la réussite d'un tel projet* ».

Un autre interlocuteur pose le fonctionnement en SaaS (instance locale, téléchargement du logiciel) comme condition nécessaire pour l'acceptation d'une nouvelle plateforme. A partir d'une expérience satisfaisante avec un autre logiciel en mode SaaS, il met l'accent sur la marge de personnalisation, sur la responsabilité, mais aussi sur la facilité de mise en œuvre (solution clé en main), la souplesse de gestion, la maintenance et du développement en ligne.

Ce refus d'une solution (trop) centralisée, qui constitue un risque pour le projet, n'est pas uniquement conditionné par le contexte français. A partir du retour d'expérience dans plusieurs pays, le rapport récent sur les « next generation repositories », de la Confederation of Open Access Repositories (COAR), met en garde contre des solutions trop loin des communautés et établissements. Il préconise une gouvernance partagée et distribuée des infrastructures de la recherche, ainsi qu'un travail en réseau où les responsabilités sont réparties, non centralisées[46].

[46] COAR op.cit.

Manque de ressources

La plupart des interlocuteurs, d'une manière ou d'une autre, ont attiré l'attention sur le manque de moyens comme risque potentiel. Ils évoquent deux aspects : le manque de disponibilité des personnels dans les SCD et les unités de recherche pour développer et faire fonctionner de nouveaux services de support à la recherche ; et le manque de moyens financiers pour le développement et la maintenance dans de nouveaux outils, plateformes etc.

« *Peu de temps* », « *pas d'argent* » : d'après les entretiens, il s'agit davantage d'expériences faites sur le terrain, plutôt qu'une analyse précise des besoins en matière de ressources humaines et financières pour développer, installer, faire fonctionner et évoluer une nouvelle plateforme de données. Les interlocuteurs opposent au projet un constat global (« *pas de modèle économique* », « *pas de sources de financement* », « *pas assez de monde dans les services* »), sans argument précis sur la réalité des besoins et du manque de ressources. Les professionnels voient les seuls modèles de services de données opérationnels à l'étranger, avec des moyens plus importants (3-7 ETP), mais la réflexion sur les contours d'un tel service, en termes de postes, de profils et de compétences, n'est qu'à son début en France[47].

Le constat d'un manque de moyens est parfois associé au regret de l'absence de décision politique dans le domaine de la gestion des données de la recherche.

Absence d'un opérateur « naturel », évident

Certes, tous les interlocuteurs évoquent un certain nombre de parties prenantes et partenaires obligés pour la mise en place d'une plateforme mutualisée, pour différentes raisons, notamment le CINES pour l'archivage pérenne, Huma-Num pour les données SHS, le CCSD et l'ABES pour les identifiants et liens vers les publications, ou encore l'INIST pour le DOI, la formation et les plans de gestion.

Mais aux yeux des différents experts et communautés, il n'y a parmi ces opérateurs aucun maître d'ouvrage ou « leader de projet » qui s'imposerait, par ses missions ou son positionnement, ou ses compétences, ou ses services, ou tout à la fois.

L'absence d'un opérateur évident peut handicaper le lancement d'un projet de cette envergure ; en revanche et assez paradoxalement, cette absence augmente les degrés de liberté et peut faciliter une gestion partenariale, une approche consortiale ou bien, une solution externalisée.

L'idée a été avancée de s'appuyer sur un groupe « RDA France »[48] pour fédérer les acteurs.

Ajoutons que certains interlocuteurs ont évoqué un autre problème qui paraît lié à cette absence de leadership : un manque de contact et de communication, sur le terrain, entre les chercheurs, les informaticiens, les documentalistes des organismes de recherche et les bibliothécaires des bibliothèques universitaires.

[47] Cf. la journée d'étude de l'ADBU du 21 novembre 2017, avec la présentation des équipes de données des universités de Bristol, Cambridge et Utrecht http://adbu.fr/demi-journee-detude-adburesearchdata-presentations-des-intervenants/
[48] Research Data Alliance cf. https://www.rd-alliance.org/groups/rda-france

Concurrence des solutions alternatives

Il suffit de consulter les répertoires Cat OPIDoR et re3data pour s'en convaincre : il y a déjà des alternatives crédibles et fiables, au niveau international, aussi bien privé que public. Les entretiens confirment ce constat : même s'il n'y a pas de projet concurrent en France, un nombre croissant de chercheurs connaissent et utilisent des entrepôts comme Zenodo, Figshare ou DRYAD. L'un des experts l'a formulé très clairement : « *Si cette plateforme n'est pas facile à prendre en main, si le dépôt n'est pas simple, et si elle n'est pas interopérable avec les autres systèmes d'information recherche, les chercheurs et les établissements vont se tourner vers d'autres solutions, en particulier vers PURE d'Elsevier, plus souple, plus facile, interopérable* ».

Partie 3 : Neuf propositions pour réussir « Data Alpha »

Introduction

Comment profiter de la dynamique favorable de la situation et, conjointement, comment, réduire ou prévenir les risques ? Voici neuf propositions pour renforcer la faisabilité d'une plateforme mutualisée pluridisciplinaire de stockage, de gestion, de signalement et de partage de données de recherche et pour réussir le lancement de « *Data Alpha* ».

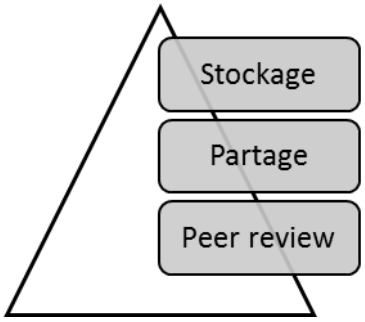
1/ Pas de « *couteau suisse* »

Les résultats de l'enquête COPIST ont révélé le besoin au sein d'établissements et d'organismes de l'ESR de disposer d'une solution nationale d'archivage, de signalement et de partage de données de recherche - une sorte de « *cloud de données scientifiques* » par défaut, mutualisé, pluridisciplinaire, indépendant des grands instruments, pour toutes les situations où l'ESR ne propose pas de solution spécifique.

Notre étude confirme ce besoin. Force est de constater, cependant, qu'il n'est pas attendu une « *solution miracle* », une sorte de « *couteau suisse* » offrant une réponse à toutes les questions, toutes les demandes et tous les besoins actuels et à venir. Le succès grandissant de Zenodo, voire même de Figshare, témoigne, si besoin, qu'il suffit de mettre en place un service numérique « just good enough » (Peter Hinssen), crédible, fiable, proposant le strict minimum des prestations et fonctionnalités attendues.

Par ailleurs, la critique que les experts formulent à l'encontre de Zenodo (« *seulement cinq métadonnées suffisent pour obtenir un DOI, très insuffisant notamment pour assurer la réutilisation* ») servira de cadre pour orienter la définition de ce strict minimum exigible. D'une manière générale, il s'agira avant tout de trois propositions de valeur incontournables, le stockage des données, leur partage, et leur mise à disposition pour des besoins de publication (peer review).

Trois propositions de valeur



L'étude préalable des dispositifs a permis de dégager les principales caractéristiques et de hiérarchiser les fonctionnalités essentielles d'une nouvelle plateforme de données.

Sur cette base, il convient de prioriser la liste des services attendus autour de quelques prestations cœur, *a minima* :

- La conservation des données de la recherche.
- La création de métadonnées, assez (suffisamment) riches, s'appuyant sur des référentiels.
- L'attribution d'un DOI.
- Le lien vers les publications, notamment pour et pendant le processus d'évaluation (peer review).
- Une interface de programmation (TDM, visualisation, data papers etc.).
- Un fournisseur OAI-PMH (interopérabilité avec d'autres dispositifs).

Avec la formule de l'un de nos interlocuteurs, on peut synthétiser cette approche comme « *une solution légère : définir le strict minimum (FAIR), avec une couche exploitation* ».

Six prestations cœur

Si l'ensemble des demandes et besoins ne sera pas satisfaits, il faudra par contre rester vigilant pour que la solution soit en conformité avec plusieurs principes, en particulier :

- Couvrir l'ensemble des types et domaines des données de la recherche.
- Assurer la sécurité des données.
- Simplifier le dépôt des données.
- Faire le lien entre la gestion des données (« chaudes ») et leur conservation à long terme (« froides »).
- Allouer un espace adapté aux besoins réels.
- Proposer l'option d'une instance locale.

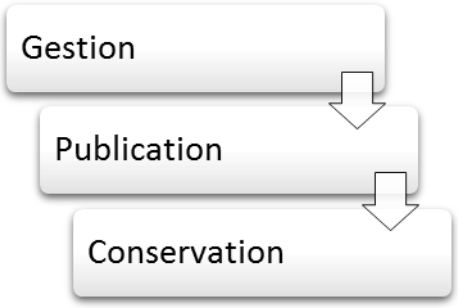
L'offre de service de la plateforme devrait inclure la proposition d'un accompagnement personnel et d'un programme de formation.

En revanche, pour les données sensibles (ex.: recherche biomédicale) et pour l'archivage pérenne, l'opérateur de plateforme fera appel au CINES et à d'autres dispositifs au sein des CHU, par exemple.

2/ Opter pour une solution pragmatique

« *Given the choice, we favour the simpler approach* ». Ce pragmatisme prôné par la COAR[49] pour le développement de nouvelles archives ouvertes pourrait servir de devise pour le projet d'une plateforme mutualisée de données. Au lieu de proposer un développement spécifique et coûteux, il est préférable d'opter pour des solutions opérationnelles, des technologies éprouvées et des standards reconnus.

Dispositif en trois couches (type « bridge »)



Cela peut se traduire par plusieurs choix :

- Adopter une architecture similaire à d'autres grands entrepôts et plateformes, tels que RADAR ou EASY, en trois couches, proposant des solutions différentes pour la gestion (« *données chaudes* »), la publication (« *données tièdes* ») et la conservation (« *données froides* »).
- Implémenter des systèmes opérationnels, en particulier Dataverse pour la gestion des données ; négocier éventuellement l'utilisation de Zenodo pour la couche publication.
- Préférer des solutions standards aux choix spécifiques (métadonnées etc.).
- Externaliser certaines prestations, en particulier la conservation pérenne. « *Il faut tenir compte du paysage riche en France et ne pas réinventer le fil à couper le beurre*. » Cela veut dire avant tout, ne pas réinventer le CINES pour l'archivage pérenne. Aussi, au moment du dépôt, le recours au service de validation de formats du CINES[50] peut s'avérer utile.
- Viser les « *niches* » : identifier les sites existants et répondre aux besoins niches qui n'ont pas de solution à ce jour - des communautés qui sont peu ou pas organisées en la matière (« *chercheurs isolés* »), et ne disposent pas de services de données satisfaisants.
- Utiliser les référentiels opérationnels.
- Ne pas viser le long terme, se contenter d'une solution transitoire, évolutive.

[49] COAR op.cit. "Where possible, we choose technologies, solutions and paradigms which are already widely deployed (…) standard Web technologies, adjusting existing software and systems (…) adoption of widely recognized conventions and standards"
[50] FACILE https://facile.cines.fr/

Ce pragmatisme est d'autant plus conseillé qu'il renforcera l'interopérabilité du dispositif et contribuera aux principes FAIR.

3/ Adopter une gestion agile et itérative

Face à l'évolution des pratiques scientifiques et compte tenu des exigences de la part des agences de financement et des éditeurs, il y a urgence. Sans doute, le temps manquera pour (re)faire une analyse des besoins et de l'existant, pour constituer un cahier des charges consensuel, pour lancer un appel d'offres et pour sélectionner un prestataire chargé du développement et de l'installation d'un nouveau service de données.

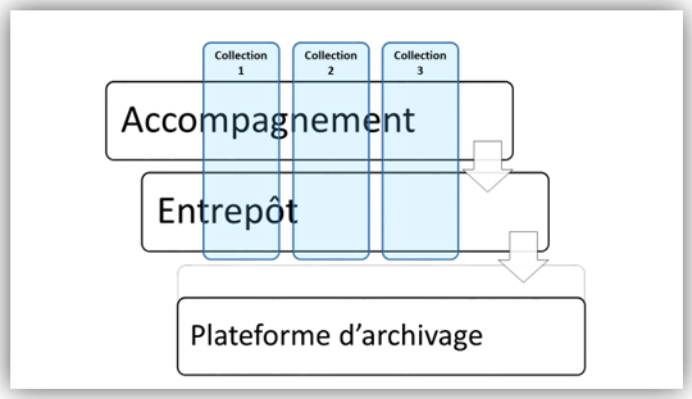
La méthodologie de choix sera sans doute une gestion agile et itérative du projet, qui permettra de répondre rapidement (d'ici 2019 ?) et d'une manière visible à la demande, en évitant l'effet tunnel, en impliquant les différentes parties prenantes et en assurant assez de flexibilité au développement, pour adapter le dispositif à l'évolution de l'environnement, si besoin.

4/ Prendre en compte les communautés et les disciplines

L'une des conditions essentielles de la nouvelle plateforme sera sa capacité à répondre, jusqu'à un certain degré, aux besoins spécifiques de certaines disciplines et communautés. Sur le terrain des SCD, instituts et laboratoires, elle sera jugée par rapport à ce critère : son succès et son usage dépendront de la réalisation de cet objectif. Cette prise en compte se traduira par plusieurs actions :

- L'implication des représentants de certaines communautés et disciplines dès le démarrage du projet, dans le cadre d'une gestion agile, non pas comme « clients/utilisateurs » mais comme partenaires.
- La mise en place d'une option de collections pour les établissements, laboratoires, instituts, sections etc. qui le demandent ; le périmètre de ces collections sera à déterminer, mais il inclura sans doute les deux couches de gestion et de publication (exposition, diffusion) des données.

Collections communautaires et disciplinaires



- La mise en place d'une gestion au niveau des collections, avec des responsabilités et droits étendus.
- Le paramétrage disciplinaire des formats et standards des métadonnées.

- La possibilité d'ajouter des outils disciplinaires au système de gestion.

5/ Définir une gouvernance partenariale

Nos interlocuteurs ne veulent pas de relation prestataire-client. Pas de « *mutualisation sous forme d'un 'au-dessus' des autres* ». Le partenariat est une condition pour le lancement du projet mais plus encore, pour la gouvernance de la future plateforme. Ceci veut dire : partenariat avec les établissements, instituts et organismes ; mais partenariat aussi avec les différents métiers, car d'après l'observation d'un des experts, il faudra d'une « *équipe intégrée* » sur place, regroupant informaticiens, chercheurs, documentalistes et bibliothécaires, garant d'un partenariat durable et de la valeur ajoutée du service. « *La question n'est pas forcément technique.* » L'essentiel est la qualité des données, la fiabilité et la crédibilité du dispositif ; « *la confiance (trust) est l'ingrédient de base* ».

Il faudrait voir dans quelle mesure il faudrait associer d'autres corps de métier et structures, comme par exemple les pôles de calcul scientifique et les ingénieurs de recherche chargés du traitement des données dans les laboratoires.

Quelle forme pour la gouvernance ? Sans doute l'articulation d'un conseil de direction et d'un conseil scientifique, intégrant des experts externes. Mais dans tous les cas de figure, il faudra s'assurer que la responsabilité publique reste entière et intacte, et que la mise en place d'une instance locale est possible : cette instance devra avoir une réelle autorité et responsabilité, non seulement pour assurer la visibilité et le contrôle des accès et contenus, mais aussi, pour enrichir les métadonnées, le cas échéant.

A l'échelle du territoire, il faudra organiser et animer le réseau de ces représentants des communautés, établissements… Au niveau local, il faudra désigner une personne ressource sur le campus, dans l'institut…

6/ Créer un environnement de service large avec un modèle économique à trouver

Comme évoqué plus haut, il n'est pas envisageable que la plateforme fasse tout et réponde à tous les besoins. Néanmoins, elle doit permettre l'articulation ultérieure avec des outils existants ou à développer, pour la gestion, production, analyse…, mais également pour leur valorisation. Pour les communautés SHS, Huma-Num paraît un bon modèle, car les services correctement conçus (expérience utilisateur) sont appréciés, notamment par les communautés les moins techniques. Un autre modèle pourrait être ISTEX avec plusieurs cercles de services autour des fonctionnalités et services cœur de la plateforme, dont certains seraient de caractère expérimental.

Outre les services déjà évoqués (data papers, outils de visualisation, plans de gestion…), ajoutons ici une articulation possible avec les cahiers de manipulation (ou de laboratoire) : une piste à explorer. Cet environnement de service large inclura l'offre de formation, d'information et de veille sur la gestion des données de la recherche.

Commentaire des commanditaires du rapport :

Parallèlement, il serait utile d'interroger les chercheurs – notamment les plus desservis en matière de gestion de données – pour recueillir leurs témoignages et attentes.

Tous les services ne sont pas nécessairement gratuits. Plusieurs exemples ont été mentionnés dans les entretiens : l'exemple du CINES (« *certains services seraient facturés, en fonction de leur valeur ajoutée, sur le modèle du CINES pour l'archivage pérenne* »), le cas du Webservice-Energy de Mines-Paritech (« *les laboratoires seraient prêts à payer un certain niveau de service, pour certaines catégories de données* ») ou celui du mésocentre HPC@LR, dont le modèle économique s'appuie sur une tarification variable en fonction du nombre d'heures de calcul et/ou du volume de stockage demandés.

La Research Data Alliance a publié deux rapports sur le financement et les modèles économiques d'entrepôts et d'autres services de données[51].

Commentaire des commanditaires du rapport :

Une analyse approfondie de l'offre commerciale émergente est nécessaire.

7/ Soutenir des infrastructures opérationnelles

Sur un point, les experts sont unanimes : le Ministère, les organismes et les établissements doivent continuer et même renforcer leur soutien aux infrastructures, entrepôts, plateformes et outils opérationnels de données. « *Il ne faut pas déshabiller Pierre pour habiller Paul*. » Ce soutien est attendu pour plusieurs services, dont Huma-Num et Progedo.

Un tel soutien facilitera également l'interconnexion avec d'autres services et dispositifs, dans un paysage où les rôles seraient clairement distribués. Le soutien pourrait être conditionné par la réalisation d'un audit selon les principes FAIR et d'autres critères de certification reconnus (CoreTrustSeal).

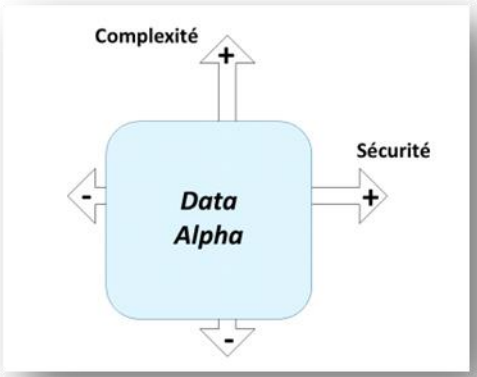
Cette répartition de la gestion et de la conservation des données de la recherche pourrait s'articuler le long d'un continuum de complexité, avec des données moins complexes (1D, 2D) pour « *Data Alpha* » et des données plus complexes (3D) pour des dispositifs disciplinaires et spécialisés.

De même, la plateforme devrait assurer un niveau de sécurité suffisamment élevé pour satisfaire les critères de la CNIL et du nouveau Règlement général sur la protection des données, sans pour autant prendre en charge tous les besoins en matière de sécurité (données secrètes, données patient…).

[51] Benedict, K., et al. 2015. *Sustainable business models for brokering middleware to support research interoperability*. Research Data Alliance (RDA). https://www.rd-alliance.org/group/brokering-ig-brokering-governance-wg/outcomes/sustainable-business-models-brokering-middleware
Dillo, I., et al. 2016. *Income streams for data repositories*. Research Data Alliance (RDA). http://dx.doi.org/10.5281/zenodo.46693

D'une manière schématique, on pourrait positionner le futur dispositif sur ces deux dimensions – complexité et sécurité des données - de la manière suivante :

Positionnement de la plateforme



8/ Accompagner le projet par une politique d'incitation

Le projet d'une plateforme nationale n'a pas de sens sans accompagnement institutionnel et politique. Comme suggèrent plusieurs interlocuteurs, il faudrait « *lancer le projet par le Ministère comme enjeu national* », « *mobiliser la CPU* », « *réveiller les instituts* », « *faire le lien avec les chercheurs* ».

Commentaire des commanditaires du rapport :

Selon l'Inra, il est possible de développer une offre de services qui aiderait les institutions ou les unités de recherche à mettre en place leurs entrepôts de données (en aidant par exemple à déployer le projet Dataverse). En revanche, l'idée d'une plateforme nationale est difficilement concevable : *« Nous n'avons pas la structure DANS ou ANDS correspondante, pas plus que les compétences ni les capacités de stockage ».* Les commanditaires de l'étude estiment qu'il revient aux institutions de piloter le système car les données leur appartiennent. Un fort engagement institutionnel est nécessaire et légitime pour accompagner la prise en compte du changement des pratiques des chercheurs sur le travail des données.

Les chercheurs sont globalement favorables à l'idée de partager les résultats de leur travail, dans la mesure du possible. Néanmoins, pour encourager la conservation et la mise à disposition sur des plateformes adéquates (= FAIR), il faudrait renforcer la prise en compte de la gestion des données de la recherche dans les outils de l'évaluation des structures et des chercheurs, dans les procédures des appels à projets et dans les protocoles éthiques.

Le projet d'une plateforme doit être accompagné d'une offre de formation (« *éducation du peuple* ») sur les plans de gestion, les licences etc. afin de développer une culture de données au sein des communautés scientifiques.

Aussi, afin de développer cette « *culture de données* » et renforcer la qualité des services, un expert imagine une certification des sites à plusieurs niveaux (A, B, C), par les communautés elles-mêmes, sur le modèle européen de la certification CLARIN.

9/ S'intégrer dans l'European Open Science Cloud (EOSC)

La question de la souveraineté nationale en matière de données a été posée plusieurs fois, sans qu'il y ait une réponse unanime. La tendance générale, cependant, est de dire que « *la souveraineté numérique ne devrait pas s'arrêter aux frontières* ».

Comme l'un des experts interrogé le dit, insister sur la souveraineté nationale est une mauvaise idée, non adaptée aux questions des données de la recherche et à la réalité des infrastructures européennes de la recherche. « *Il vaut mieux investir pour améliorer Zenodo que de développer une plateforme nationale dont le fonctionnement et la maintenance vont coûter cher*. » Un autre ajoute : « *Zenodo existe, pas besoin de refaire l'étude* ».

L'idée est un « *nœud national à l'échelle européenne* », soit disciplinaire, soit institutionnel. Une plateforme mutualisée fera sens si elle réalise une économie d'échelle et si elle est complémentaire et interopérable avec le paysage des infrastructures de données.

Indépendamment du choix de la solution, il faudra s'assurer dès le démarrage du projet de sa compatibilité avec les nouveaux projets et orientations dans l'espace de recherche européen. Ceci concerne en particulier la nouvelle initiative pour un cloud européen, l'EOSC.

Le facteur temps

La situation est là. Les besoins ont été exprimés. Il existe des projets qui peuvent et doivent mûrir ; il leur faut du temps pour fédérer tous les acteurs autour d'un objectif, pour monter un cahier des charges, pour trouver un prestataire et réunir les ressources nécessaires.

Le projet d'une plateforme mutualisée pluridisciplinaire de stockage, de gestion, de signalement et de partage de données de recherche n'est pas de ce genre. De l'avis des experts et intéressés, il faut agir maintenant, et vite. « *Il faut des pistes à réaliser à court terme* ». Vite, cela veut dire, lancement de projet maintenant, avec une première réalisation en 2019.

L'alternative (ou la variante) est toute tracée. Pour citer la célèbre formule du Président du Conseil Henri Queuille, « *il n'est pas de problème dont une absence de solution ne finisse par venir à bout* ».

Si l'ESR ne fait rien, ou s'il le fait trop tard, il restera trois options :

- Des solutions disciplinaires et locales. La conséquence sur le terrain serait probablement un amoindrissent de la qualité et des compétences. La conséquence pour la gouvernance serait moins de pilotage, moins de contrôle. Les coûts seront certainement supérieurs.
- Une augmentation des dépôts sur Zenodo (au moins transitoirement, cf. la collection INRA), entraînant des problèmes dus en particulier aux métadonnées relativement pauvres.
- Un recours aux solutions commerciales d'Elsevier, de Digital Science (Holtzbrinck) ou d'autres[52].

[52] Les universités suisses externalisent une partie de la gestion de leurs données à une fondation de droit privé (SWITCH) qui propose un service d'archivage « in the cloud », SWITCHdrive https://www.switch.ch/drive/

Annexes

Annexe 1 : Personnes interrogées

Représentants des plateformes et entrepôts analysés

California Digital Library	Stephen Abrams, directeur adjoint de l'UC Curation Center (UC3)
Académie des Sciences Prague	Zdena Chmelarova, responsable de l'archive des données
DANS	Peter Doorn, directeur
Figshare	Elwin Gardeur, directeur des ventes Europe
Université Cambridge	Danny Kingsley, directeur adjoint de la BU
TIB Hannover	Angelina Kraft, responsable du projet RADAR
CERN	Lars Holm Nielsen, responsable du service Zenodo

Représentants de l'ESR

ABES	David Aymonin, directeur
CCSD	Christine Berthaud, directrice
CDS	Françoise Genova, chercheuse, ancienne directrice, RDA France
CNAM-INTD	Violaine Rébouillat, doctorante, chargée d'étude BSN10
DIST-CNRS	Francis André, chargé de mission « données de recherche »
INIST	Paolo Laï, Bernard Sampité, Alain Zasadzinski, Jean-Michel Parret, Jean-François Nominé, Claire François, Florian Mazur
INRA	Odile Hologne, directrice déléguée à l'IST
INRIA	Laurent Romary, chercheur, directeur DARIAH
INSERM	Marc Cuggia, PRU, responsable de l'Unité Fouille de données du CHU de Rennes, co-cordinateur du réseau des centres de données cliniques du Grand Ouest
IRSTEA	Emmanuelle Jannès-Ober, responsable IST
MESRI	Marin Dacos, directeur du CLEO, conseiller scientifique pour la science ouverte auprès du DGRI du MESRI
Renater	Patrick Donath et Xavier Misseri (Renater)
Sciences Po Paris	François Cavalier, directeur de la bibliothèque
TGIR Huma-Num	Stéphane Pouyllau, directeur technique
Université de Bordeaux	Jérôme Poumeyrol, responsable du soutien et des services à la recherche à la Direction de la documentation, chef de projet AOI, ADBU
Université de Bordeaux Montaigne	Julien Baudry, responsable des services aux chercheurs, SCD
Université de Lille Sciences et Technologies	Marie-Madeleine Géroudet, chef de projet AOI, service support à la recherche, SCD ; Romain Féret
Université de Nice	Mathieu Saby, chargé des données de la recherche, SCD

Université de Paris Descartes	Aurore Cartier, service support à la recherche
Université de Paris Diderot	Christophe Pérales, directeur SCD, président ADBU
Université de Strasbourg	Adeline Rege, chef de projet AOC, SCD, chercheuse associée à l'EA3400
URFIST Rennes	Alexandre Serres, enseignant-chercheur, directeur

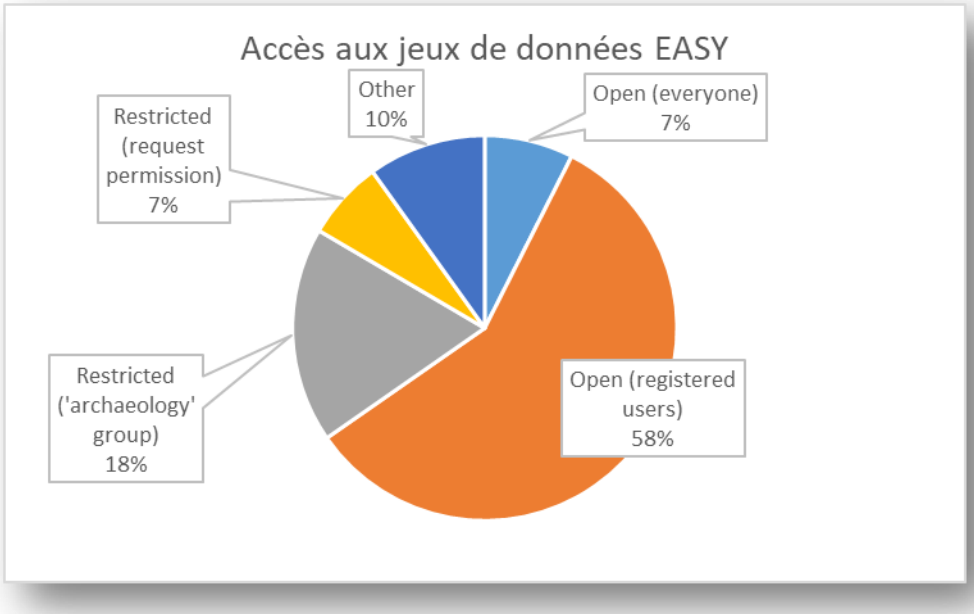
Annexe 2 : Développement et usage des plateformes étudiées

EASY (DANS, Pays Bas)

EASY contient 42 207 jeux de données (janvier 2018) inégalement répartis entre les sciences sociales (11%) et les sciences humaines (89%), avec très peu de données dans d'autres domaines.

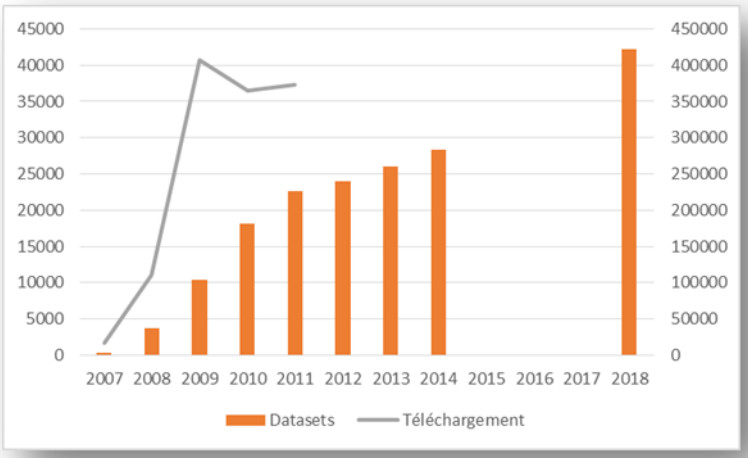
Behavioural and educational sciences	1327	3%
Economics and Business Administration	234	1%
Humanities	37679	89%
Interdisciplinary sciences	3919	9%
Law and public administration	814	2%
Life sciences, medicine and health care	493	1%
Science and technology	146	0%
Social sciences	4672	11%

Seulement 7% des jeux de données sont librement accessibles, sans restriction. 58% sont accessibles sur inscription.



EASY a été lancé en 2007. D'après les chiffres incomplets des rapports d'activité et du site de DANS, la croissance annuelle des dépôts a été d'environ 10%.

Nous ne disposons pas de statistiques d'utilisation récentes. Pour 2011, DANS rapporte 373 553 téléchargements, pour plus de 50 000 visiteurs du site.



Nous disposons également de quelques chiffres sur le développement des Dataverse dans les établissements néerlandais. En janvier 2018, le site DataverseNL[53] contenait des données de 11 universités dont quatre universités des sciences appliquées (4TU), et d'organismes de recherche, représentant 519 jeux de données et 2 242 fichiers, répartis dans 297 instances disciplinaires et institutionnelles :

Etablissement	Nb Dataverse	Nb datasets	Nb data files
4TU	45	40	216
Amsterdam	9	5	261
Groningen	32	34	95
Leiden	6	0	6
Maastricht	54	56	168
Rotterdam	15	3	4
Tilburg	8	22	504
Utrecht	45	81	362
NIOO KNAW	36	25	69
DANS	4	6	9
Tilburg Workspace	26	234	530
Autres	17	13	18
Total	**297**	**519**	**2242**

Chaque établissement a créé des instances en fonction de ses besoins, avec un nombre médian de 26 instances par structure. Il s'agit pour 33% de serveurs Dataverse institutionnels, 18% ont été créés par des équipes scientifiques, 8% par des projets de recherche et 4% par des chercheurs individuels.

Les chiffres pour les jeux et fichiers de données ne sont pas élevés, mais il s'agit d'un dispositif assez récent. 43% des jeux de données sont issus des sciences sociales, 12% proviennent des sciences humaines et des arts. Les sciences médicales et sciences de la vie représentent 13%, les sciences de la terre et de l'environnement 9%.

Le site indique un total de 7 379 téléchargements pour l'ensemble des Dataverse (3 téléchargements par fichier en moyenne).

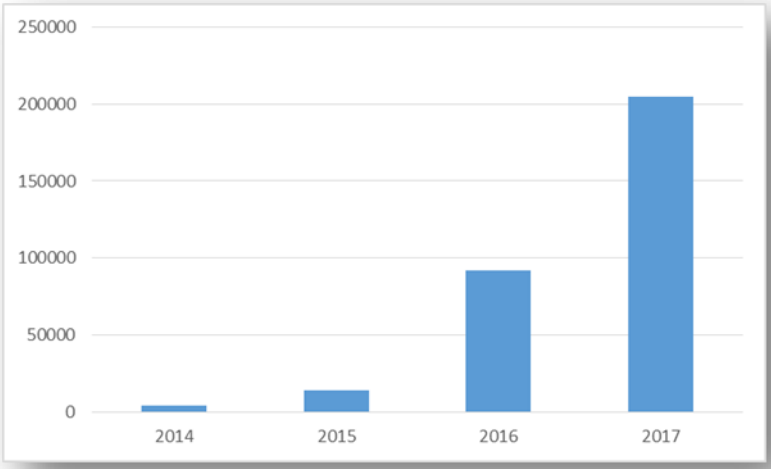
Zenodo (CERN, Suisse)

En janvier 2018, Zenodo indiquait :

[53] https://dataverse.nl/

- 341 923 dépôts
- 22 405 jeux de données
- Presque 40 000 comptes d'utilisateurs

L'évolution des dépôts annuels (non cumulés) depuis 2014 montre une très nette augmentation en 2017 :



Figshare

Nous disposons de peu d'informations précises concernant Figshare :

- 2,5 millions de fichiers (dépôts)
- 500 000 jeux de données (y compris vidéo que Figshare considère comme « research output » et « research data »)
- 15 000 citations
- Plusieurs millions de téléchargements
- Plusieurs centaines de milliers d'utilisateurs enregistrés partout dans le monde

CSIRO Data Access Portal (DAP)

L'analyse de l'évolution des téléchargements du portail australien[54] révèlent deux aspects :

- Le premier est classique. L'usage des données représente une distribution du type Pareto, avec quelques dépôts fortement demandés et une longue traîne de données peu téléchargées comme le montre le schéma suivant.

[54] Sanderson, T., et al., 2017. *Understanding and unlocking the value of public research data. OzNome social architecture report*. CSIRO, Melbourne.

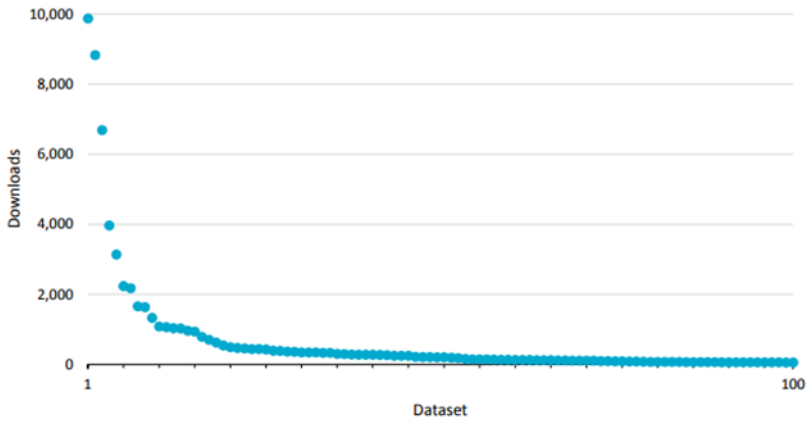


Figure 17: DAP datasets by download frequency (top 100)

- Le second, en revanche, est plus inattendu. Il n'y a, en effet, qu'une faible corrélation entre l'âge des données et le nombre de téléchargements, du moins pour les premières années :

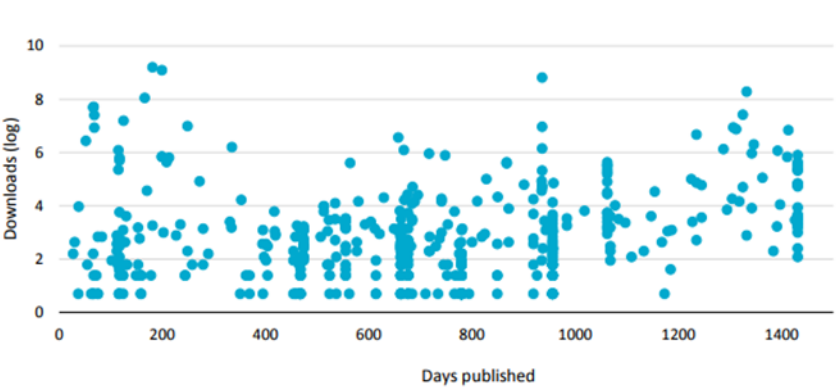


Figure 21: Download count (log adjusted) by length of time published (public data collections)

Pour certaines données, l'étude révèle même une augmentation de la demande avec le temps. La même observation a été faite par un projet allemand, en particulier pour des données dans le domaine des sciences de la terre (données géologiques, sismiques etc.)[55].

[55] http://www.forschungsdaten.org/index.php/Radieschen

Annexe 3 : Eléments financiers relatifs à une plateforme de données

Le coût de production

Un rapport allemand a chiffré en 2011 les coûts annuels d'une plateforme ou d'un centre de données à 3,5 million €[56] (par exemple, pour GESIS ou DFD). A partir de ces chiffres, la commission à l'origine de ce rapport a extrapolé un besoin de 700 million € par an pour maintenir un dispositif de données dans chaque établissement de l'ESR allemand (200 plateformes en tout). Ce budget annuel s'ajouterait aux dépenses de développement et de lancement, estimées à 500 million € sur une période de 4-5 ans.

En gros, cette commission a estimé les dépenses liées à la gestion des données à 5-10% du budget recherche ; elle considère ces dépenses pour les services de données comme une nécessité, pas comme une option.

D'après une étude du JISC, les coûts de la préservation des données de la recherche se répartissent ainsi : 55% acquisition des données (avec promotion), 31% services d'accès, 15% archivage numérique. L'essentiel correspond à des coûts fixes (ressources humaines, licences…) indépendants du nombre ou du volume des données déposées[57]. Ces coûts diminuent progressivement. L'étude ajoute que les services d'accès (31% des coûts) sont souvent des services à valeur ajoutée et pourraient faire l'objet d'une facturation.

Cette estimation est confortée par une analyse allemande de plusieurs plateformes et dispositifs allemands, anglais et américains[58]. Selon cette étude financée par la DFG, plus de la moitié du personnel des centres de données sont affectés à l'acquisition et « ingestion » des données, les autres personnes se partageant entre la curation, la conservation et la diffusion.

La même étude estime les coûts de système et d'équipement (hardware, maintenance etc.) d'un centre de données entre 20 000 et 200 000 € par an, en fonction du volume des données traitées (chiffres pour l'Allemagne, 2012). L'étude observe qu'il faudrait renouveler l'équipement tous les 4-5 ans.

Un deuxième rapport du même projet propose un modèle pour calculer le coût par prestation et activité.[59] A partir d'une étude de cas, il construit un modèle pour une analyse comptable au plus fin, par heure/homme. Il n'est pas important de présenter

[56] Kommission Zukunft der Informationsinfrastruktur, 2011. *Gesamtkonzept für die Informationsinfrastruktur in Deutschland.* Gemeinsame Wissenschaftskonferenz des Bundes und der Länder, Berlin. https://www.leibniz-gemeinschaft.de/fileadmin/user_upload/downloads/Infrastruktur/KII_Gesamtkonzept.pdf

[57] *Keeping Research Data Safe Factshee*t https://beagrie.com/static/resource/KRDS_Factsheet_0711.pdf

[58] Rathmann, T., 2013. *Kostenverteilung und Risiken*. DFG Projekt RADIESCHEN Rahmenbedingungen einer disziplinübergreifenden Forschungsdateninfrastruktur. http://gfzpublic.gfz-potsdam.de/pubman/item/escidoc:117203:2/component/escidoc:117202/ProjektRadieschen_Kostenverteilung_und_Risiken.pdf

[59] Rathmann, T., 2013. *Preise, Kosten und Domänen*. DFG Projekt RADIESCHEN Rahmenbedingungen einer disziplinübergreifenden Forschungsdateninfrastruktur. http://gfzpublic.gfz-potsdam.de/pubman/item/escidoc:117052:3/component/escidoc:117199/ProjektRadieschen_PreiseKosten_und_Domaenen.pdf

le modèle en détail ici (cf. copie du tableau ci-après), puisque les différents éléments évoluent avec l'évolution des technologies et procédures.

Une transposition en France impliquerait nécessairement une nouvelle analyse comptable, au sein d'un service de données en France.

	Pro Auftrag	Pro Experiment	Pro Experiment bei gleichen Datenstrukturen
Daten- und Metadatenarchivierung (Ingest)			
Information und Beratung	4		
Projekt-Spezifikation (Festlegung Daten-umfang, Formate, Datenorganisation, Speicherstrategie, Weg der Daten zum WDCC, Data-Policy, Zugriffsbedingungen)	2		
Erstellen eines Konzeptes (Metadaten-umfang, Preprocessing, Zeitplan) und Kostenabschätzung	4		
Erfassen, Einfüllen und Qualitätskontrolle der Metadaten	10	5	3
Aufsetzen Datentransfer und Einfüllen der Daten	7	1	1
Qualitätskontrolle der Daten einschl. Prüfung der Konsistenz von Metadaten und Daten		10	4
Freischaltung und Abschluss-Report	6		
insgesamt	33	16	8
10 Jahre Speicherung inklusive Pflege			
Aktualisierung der Metadaten	10	10	8
Pflege der Datensätze innerhalb der Datenbank		10	5
Anpassung der Zugriffsberechtigungen	8	2	2
Laufende Anpassung an die DKRZ-Infrastruktur	10	5	3
insgesamt	28	27	18

Tabelle 2: Personalaufwand am WDCC für Teilprozesse in Personenstunden (Luthardt, 2010)

L'essentiel est ailleurs. La comparaison des coûts par commande, expérimentation et expérimentations similaires révèle un fort potentiel d'une économie d'échelle, estimée à au moins 50% : si le volume augmente de 100%, les dépenses augmenteront seulement de 50%. Il s'agit d'un argument fort en faveur de la mutualisation[60].

Le JISC a par ailleurs estimé qu'il faudrait au minimum 1 ETP par unité de recherche ou institut pour la gestion des données. Pour la commission allemande, cela correspond à environ 5% des ressources humaines dans un laboratoire.

Mutualisation

Ceci étant, la mutualisation peut changer la situation. Le lancement du projet RADAR a été financé par la DFG à hauteur de 800 000 €. Chaque établissement membre du projet a ajouté des ressources propres, avant tout des postes (au total, 4 ETP dans cinq établissements).

[60] Cf. aussi Beagrie, N., Lavoie, B., Woollard, M., 2010. *Keeping research data safe 2: Final report*. JISC, Bristol. http://repository.essex.ac.uk/2147/

Aux Pays Bas, l'ESR a mutualisé la plateforme de gestion DataverseNL. Les dépenses fixes (administration, serveurs, maintenance…) sont prises en charge par l'opérateur national, DANS, tandis que les frais variables (dépôt et stockage des données, en fonction de l'espace réellement occupé) sont à la charge des établissements membres du réseau. Le montant annuel moyen de leurs dépenses s'élevait en 2014 à 3 950 €.

Une étude du JISC présente le cas d'un établissement ayant pu réduire les frais de gestion par l'externalisation de l'archivage numérique vers un entrepôt central (mutualisé) d'environ 41%[61].

Risques et perte de données

La reproductibilité des résultats scientifiques est souvent limitée. Plusieurs études de 2011 et 2012 dans le domaine de la médecine estiment que seulement 11-25% des études sont reproductibles, faute de données[62] ; en d'autres mots, à cause de la perte des données ou parce que les données ne sont pas (plus) accessibles.

Une analyse de publications en morphologie révèle une baisse progressive de la disponibilité des données sur une période de vingt ans (1991-2011). Deux ans après la publication, la plupart des données sont encore disponibles pour d'autres chercheurs. Mais 20 ans plus tard, 80% des données ont été perdues, avec une baisse moyenne de 17% par an[63].

Beaucoup de revues prestigieuses, notamment dans les sciences de la vie, demandent aux auteurs de rendre leurs données accessibles. Or, une étude de 2011 montre qu'en réalité, les données sont accessibles pour seulement 9% des articles[64].

Un projet allemand a essayé de chiffrer les coûts de la perte de données[65] à partir des dépenses liées à la production, les dépenses liées à la récupération ou remplacement des données perdues, à la perte de valeur réelle dû au temps perdu et à la perte de valeur potentielle si les données demandées n'ont pas pu être récupérées ou substituées.

Le rapport résume le risque ainsi : « *Si on ne s'occupe pas de l'archivage des données de recherche, on court un risque élevé de perte de données. Même si les données sont encore quelque part sur des supports de données, la recherche des données, la migration (récupération) et la reconstitution deviennent coûteuses. Quand les données sont perdues, les conséquences peuvent être beaucoup plus chères. S'il faut reproduire les données perdues, les dépenses correspondent souvent aux coûts de la recherche – sans parler du temps perdu. Si les données ne sont pas retrouvées ou reconstituées, il y a perte de valeur potentielle, et les usagers doivent supporter le coût de la démarche infructueuse.* » (Rathmann 2013, p.31-32, trad.).

[61] *Keeping Research Data Safe Factsheet* https://beagrie.com/static/resource/KRDS_Factsheet_0711.pdf
[62] Begley, C. G., Ellis, L. M., Mar. 2012. Raise standards for preclinical cancer research. *Nature* 483 (7391), 531-533. Prinz, F., et al. 2011. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* 10 (9), 712.
[63] Gonzalez, A., Peres-Neto, P. R., 2015. Act to staunch loss of research data. *Nature* 520 (7548), 436. Vines, T. H., et al. 2014. The availability of research data declines rapidly with article age. *Current Biology* 24 (1), 94-97.
[64] Alsheikh-Ali, et al. 2011. Public availability of published research data in High-Impact journals. *PLOS ONE* 6 (9), e24357+.
[65] Rathmann 2013 loc.cit.

Le rapport évoque un exemple de l'information scientifique : si les données issues d'un projet de numérisation sont perdues ou de mauvaise qualité, la valeur perdue s'élève au budget du projet, à condition que les items numérisés soient toujours disponibles.

Un article sur la protection des données sensibles (à caractère personnel, patients…) fait état d'un risque élevé, dû aux erreurs logicielles (bugs) et humaines, vols, pertes de support. Ainsi, en peu d'années, des millions de données ont été perdues au Royaume-Uni[66] : « *Between 24 April 2009 and 29 October 2009 alone, the loss of 878 513 records by 35 organizations is listed. Included in the list is the University of Manchester, after a member of staff emailed an attachment to 469 students with data on 1700 people including information on student disabilities; and Imperial College when six laptops were stolen resulting the loss of medical data containing confidential information on 6000 patients.* »[67]

Le Digitale Bewaring Project aux Pays Bas a estimé la production d'un batch de 1 000 jeux de données (avec métadonnées et documentation) à 333 €. S'il fallait réparer ce batch dix ans plus tard, dû à des métadonnées manquantes, perdues ou incomplètes, l'action de récupération coûterait 10 000 €. En d'autres mots, produire et maintenir des métadonnées riches génère un bénéfice indirect (ou économie) de 97%[68].

Retour sur investissement

Début 2017, l'Australian National Data Service (ANDS) a mis en ligne seize projets exemplaires pour démontrer l'impact sociétal de la publication des données de la recherche des établissements et organismes australiens[69]. Peu d'études contiennent des chiffres pour décrire les retombés (populations concernées etc.) et seulement deux évoquent des enjeux financiers, dans le domaine de l'agriculture. Pour un projet (EverGraze project boosts Australian farming, Charles Sturt University and partners), le retour sur investissement de la mise à disposition des résultats est estimé à 9:1, avec un bénéfice global de $306 million.

Un rapport du CSIRO[70] essaie de calculer le retour sur investissement de son Data Access Portal (DAP) comme rapport entre le coût de téléchargement (198 $) et la valeur d'un jeu de données pour les utilisateurs et clients (recherche, secteur public, entreprises) de la plateforme (4 802 $). En d'autres termes, le rapport est établi entre les frais de fonctionnement (13,9 million $ pour six ans) et la valeur économique créée (estimée à 67 million $ par an) : « *While this estimate is highly uncertain, the fact that it exceeds the annual costs by nearly two orders of magnitude suggests that the DAP provides an excellent return on public investment* » (loc.cit., p.9-10).

[66] Aldridge, J., et al. 2010. The problem of proliferation: Guidelines for improving the security of qualitative data in a digital age. *Research Ethics* 6 (1), 3-9.
[67] *UK Privacy Debacles* https://wiki.openrightsgroup.org/wiki/UK_Privacy_Debacles
[68] Cf. aussi https://beagrie.com/krds.php
[69] ANDS, 2017. *#dataimpact Stories about the real-life impact of Australian research data.* http://www.ands.org.au/news-and-events/latest-news/news/new-data-impact-book-published
[70] Sanderson, T., et al., 2017. *Understanding and unlocking the value of public research data. OzNome social architecture report*. CSIRO, Melbourne. https://publications.csiro.au/rpr/download?pid=csiro:EP168075&dsid=DS2

Un exemple de cette étude sur la valeur créée par l'utilisation des jeux de données du portail du CSIRO (p.54) :

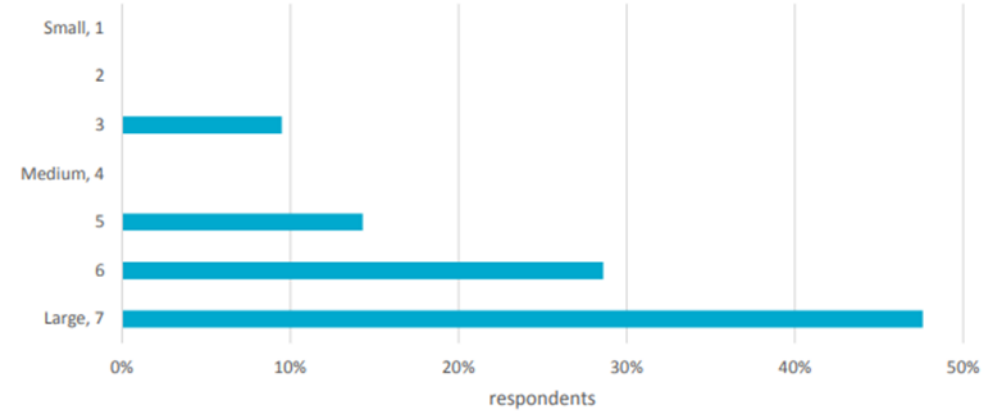


Figure 28: On a scale from 1 (small) to 7 (large), given your current uses for this data (research, product development, etc.), what is the worth (value) of this data to your current work efforts?

Les utilisateurs interrogés ont estimé cette valeur à plusieurs milliers de dollars : « *Responses ranged from a low of $5,000, to $10,000, $40,000, several estimates of $100,000, through to values of millions of dollars.* »

L'étude allemande déjà citée a tenté de faire la synthèse de plusieurs études de retour sur investissement[71] et arrive à un rapport moyen de 1:5-7. En d'autres termes, pour un budget d'investissement ou de fonctionnement de 100 €, un centre de données créé jusqu'à 5 à 7 plus de valeur, c'est-à-dire 500-700 €. Ce retour sur investissement peut être supérieur, mais aussi inférieur, en fonction des domaines, de la nature des données et de l'usage.

[71] Rathmann 2013, loc.cit.

Annexe 4 : Commentaires post-étude des commanditaires

Introduction

Après relecture de l'étude, ses commanditaires ont recensé des bonnes pratiques et des actions susceptibles d'être menées dans un cadre institutionnel et dans des instances extérieures. Il s'agit d'éléments qui sont de près ou de loin abordés par l'étude, mais qu'ils jugent essentiels et veulent préciser.

In fine, les commanditaires jugent très utile de compléter et de confronter les enseignements tirés de cette étude en interrogeant des chercheurs de disciplines différentes qui manipulent des types de données multiples. Cette enquête comparative pourrait faire l'objet d'une étude complémentaire.

Des actions à inscrire dans un cadre institutionnel

- Prendre en compte les besoins d'espace de stockage (DSI) et suivre les initiatives des universités travaillant au développement de leurs entrepôts.
- Faire le lien entre les publications (HAL) et les entrepôts existants. Approfondir l'outil SCHOLIX que HAL a utilisé pour lier les données aux publications : sur 3 années (2016, 2017, 2018) HAL a réussi à trouver 2% de liens avec les jeux de données.
- Suivre les évolutions des produits commerciaux tels que ExLibris qui a développé l'outil Esploro à partir de la solution Alma, liée au SGBM (le système collecte automatiquement les métadonnées et remplit l'archive institutionnelle et le chercheur n'a plus qu'à déposer le fichier. Il existe un dispositif pour faire le lien avec les données de la recherche via un workflow mis en place).
- Avoir une représentation (Couperin, Abes) dans le collège des données de la recherche au Comité pour la Science ouverte et s'intégrer à l'EOSC (European Open Science Cloud).
- Apprendre à optimiser les données, au-delà des données accompagnant les publications, et inciter les institutions à s'emparer de la question. Examiner en particulier l'exemple de l'Université de Bordeaux qui cible ses politiques d'accompagnement sur les chercheurs et les équipes lauréates d'appels à projets. Examiner plus largement les politiques d'accompagnement dans les laboratoires initiant des propositions de formations.
- S'intéresser aux communications et échanges du réseau informel de data-librarians provenant d'une douzaine d'établissements.
- Suivre et synthétiser l'enquête lancée à l'initiative de TU Delft, University of Cambridge, EPFL et University of Illinois, qui porte sur les pratiques et attentes des chercheurs.
- Monter une journée d'étude et d'ateliers à l'Inist, sur le modèle de la journée du 17 mai 2018, qui portera sur les pratiques des universités en matière de dépôt des données. A l'Université de Bordeaux, par exemple, ce sont le signalement, l'appui sur les outils existants et une gouvernance partagée qui sont favorisés, au détriment d'une création de dépôt. L'USPC a le projet de créer un datacenter. A l'Université Grenoble-Alpes, un datacenter existe et parallèlement s'est créé le

Grenoble Alpes Data Institute. Cette journée servirait à créer des liens entre ces structures et à les rapprocher du GRICAD et des SCD.

Actions à inscrire dans des instances extérieures (CoSO, etc.)

- Avoir un fort engagement institutionnel et auprès des instances extérieures actives. Le ministère, les organismes et les établissements doivent renforcer leur soutien aux infrastructures, entrepôts, plateformes et outils opérationnels de données afin de faciliter l'interconnexion avec d'autres services et dispositifs. De fait, les rôles et missions de chacun s'inscriraient dans un contexte dynamique et seraient clairement distribués.
- Lancer un projet de plateforme expérimentale pour répondre à un besoin temporaire et qui pourra couvrir les besoins des communautés non desservies par les plateformes existantes.
- Définir progressivement la gouvernance partenariale à mettre en place, le modèle économique le plus adapté tout en permettant éventuellement de faire émerger un opérateur.
- S'intégrer dans l'European Open Science Cloud (EOSC) et participer à l'initiative GO-FAIR.
- Mettre en place une plateforme nationale pour venir en soutien aux établissements qui n'en possèdent pas et qui n'ont pas les moyens de mettre en place ce genre de plateforme.
- Analyser les éléments de concurrence opensource, entre Zenodo et Figshare, et de concurrence commerciale à travers PURE par exemple.

Autres commentaires

- La plateforme nationale est d'un usage difficile en matière de gouvernance. Certaines communautés fonctionnent déjà sur ces modes et ont déjà des pratiques et des infra.
- Les changements de pratiques impliquent du personnel d'accompagnement et de soutien ainsi que des infrastructures d'accueil, pour permettre la curation, la diffusion et le partage.