Synthèse des journées d'étude RIP Data : Quelle sélection, conservation et suppression des données de recherche ?

Christine Hadrossek

Synthèse des journées d'étude RIP Data : Quelle sélection, conservation et suppression des données de recherche ?

https://rip-data2024.sciencesconf.org/

Groupe de travail sur les Données de la Recherche de la Mission pour les Initiatives Transverses et Interdisciplinaires du CNRS

Christine Hadrossek[1]

Les 1er et 2 octobre 2024, trois demi-journées d'études, organisées par le Groupe de travail DoReMITI[2] (Données de REcherche de la Mission pour les Initiatives Transverses et Interdisciplinaires) ont rassemblé plus d'une centaine de participants à la Maison Méditerranéenne des Sciences de l'Homme à Aix-en Provence pour parler de la sélection, conservation et suppression des données de recherche.

Ces trois demi-journées d'échange et de discussion avaient pour objectif de témoigner des pratiques d'archivage et de préservation des données de recherche dans nos laboratoires.

Animées par Christine Hadrossek, chargée des données de la recherche à la DDOR et Amandine Hénon, cheffe de projet InDoRES au Muséum national d'Histoire naturelle, les journées RIP Data ont débuté par la présentation d'un programme structuré autour de 4 grandes parties :

- La première partie était dédiée à l'archivage des données de la recherche et avait l'ambition au travers d'une présentation générale et théorique[3] et de retours d'expériences d'archiviste[4]

[1] CNRS, DDOR - Direction des Données Ouvertes de la Recherche

[2] DOReMITI est un groupe de travail inter-réseau constitué en 2016, à l'occasion des rencontres des réseaux professionnels du CNRS. Soutenu par la MITI (Mission pour les Initiatives Transverses et Interdisciplinaires du CNRS) il rassemble plusieurs réseaux (devlog, Renatis, Calcul, Ecobio etc, autour d'un objet d'étude commun : les données de recherche. A l'affut de toute nouvelle problématique en lien avec le cycle de vie des données, le groupe s'attache en particulier à développer une approche transversale, à croiser des regards, des expériences et expertises métiers et à sensibiliser les communautés professionnelles à la gestion des données.

[3] Marie-Laure Bachelerie-Gouverneur (DSI – CNRS), L'archivage des données de la recherche.

[4] Stéphy Lefoll (LAAS-CNRS), Du sauvetage à l'archivage de la production documentaire d'un groupe de recherche : l'exemple du département Robotique du LAAS.
Aurélie Montagne Borras (MSH Monde), Archivistique et bonne gestion de l'information. Quelle pratique à la MSH Mondes ?
Océane Valencia (Sorbonne Université), L'archivage des codes et logiciels.

de mieux définir cette notion parfois mal comprise et mal utilisée. Une vision de ce qui se construit à l'échelle nationale au sein de la Mission des Archives du Ministère et au Service des archives de France a également pu être partagée[5].

- La seconde partie était consacrée à des démarches de préservation des données au sein d'infrastructures ou d'organisations telles que le CCIN2P3[6], Data Terra[7], le Museum d'Histoire Naturelle[8] ou l'Université d'Aix Marseille[9]. Une table ronde, introduite par une présentation de la certification CoreTrust Seal[10] a également réunis les administrateurs de 3 entrepôts[11] (INDORES, GBIF, Nakala) pour aborder cette question à plus petite échelle.
- La troisième partie avait pour objectif de discuter la question du coût, des besoins et moyens nécessaires au stockage et à la conservation des données. Le coût environnemental d'abord, pour prendre la mesure de l'impact environnemental du traitement des données de recherche[12] mais aussi le coût financier, économique. Pour finir nous avons discuté du déploiement d'offres de services[13].
- La dernière partie, enfin, proposait de dresser, à l'occasion d'une table ronde finale un panorama des pratiques en cours dans les communautés pour décider de la sélection des données à conserver.[14]

Ce document a pour objectif de synthétiser les principaux éléments ou questionnements soulevés à l'occasion des nombreuses présentations qui ont nourri nos débats et des riches échanges qui ont suivi.

De ces journées, nous retenons 7 principaux points d'attention débattus par les participants.

- Les archives, l'archivage et l'élimination contrôlée des données
- Les archivistes face à la gestion des données de recherche
- Les dispositifs en place pour la sélection et l'archivage et l'enjeu d'une stratégie archivistique

---

Véronique Ginouvès et Emilie Groshens (MMSH), Le Plan de gestion de données rétrospectif, un outil pour l'accès aux archives de la recherche sur le long terme : documenter les usages archivistiques du présent pour garantir les (ré)usages du futur.

[5] Cyprien Henry (MESR), Livre blanc Mission des archives du MESR.
Dominique Naud (Service interministériel des Archives de France), Le rôle du Service interministériel des Archives de France.

[6] Yonny Cardenas (CNRS-CCIN2P3), Data Future - un projet pilote d'infrastructure pour la préservation des données scientifiques sur le long terme à l'IN2P3/CNRS

[7] Joël Sudre (UAR Data Terra), La préservation des données au sein de Data Terra/Gaïa Data

[8] Cécile Callou (CNRS-MNHN), Présentation d'une démarche de préservation initiée au MNHN.

[9] Fabien Borget (amU), Déploiement d'une offre de service données sur le site d'Aix Marseille.

[10] Olivier Rouchon (CNRS-DDOR), Certification Core Trust Seal pour les entrepôts de données et gestion des risques

[11] Amandine Hénon (CNRS-MNHN), InDores - Sophie Pamerlon (GBIF France- UMS Patrinat), GBIF - Hélène Jouguet (Huma-Num), Nakala

[12] Alexis Arnaud (GRICAD), Préserver les données dans un monde contraint
Guillaume Levavasseur (IPSL), Expérience de la communauté du climat

[13] Jean-François Perrin (ESRF), Retour d'expérience ESRF
Denis Veynante (CNRS-DDOR), Initiatives en cours autour du stockage de données au CNRS - Philippe Prat (CINES), Archivage - retour d'expérience et étude des coûts
Henri Valeins (CNRS-CRMSB), Archiver les données des cahiers de laboratoire

[14] Julien Peloton (IJCLAB, Orsay), Communauté astro-physique - Françoise Genova (CDS Strasbourg), Communauté astronomie - Denis Veynante (CNRS-DDOR), Communauté mécanique des fluides - Stéphane Renault (CNRS-LAMPEA), Communauté archéologie - Isabelle Charpentier (Réseau des Zones Ateliers), Communauté écologie et environnement - Pierre Susbielle (GIPSA Lab), Communauté robotique

- Les difficultés et facteurs de réussite dans la mise en place d'un écosystème pour le partage et la préservation des donnés
- La gestion des risques et préservation des données dans les entrepôts, les bonnes pratiques et les opportunités
- Les défis et solutions pour une production de données responsable dans un monde contraint
- Les pratiques et stratégies pour la sélection des données de recherche

1. Les archives, l'archivage et l'élimination contrôlée des données

L'archive est un terme particulièrement connoté dans l'esprit collectif, compris dans 90% des cas comme un ensemble de papiers poussiéreux pour historiens. Qu'en est-il vraiment ?

C'est pour commencer, nous explique Marie-Laure Bachèlerie Gouverneur, responsable du pôle national de conservation de données et documents au CNRS, quelque chose qui s'inscrit dans le temps et qui aujourd'hui, à l'ère du numérique, se présente sous une multitude de supports et formats différents qu'il faut pouvoir préserver sur la durée. C'est aussi un terme défini par la Loi, dans le code du patrimoine (article L 211-1) : « Les archives sont l'ensemble des documents, y compris les données, quelle que soit leur date, leur lieu de conservation, leur forme et leur support, produits ou reçus par toute personne physique ou morale et par tout service ou organisme public ou privé dans l'exercice de leur activité ».
Ici, tous les termes sont importants et significatifs et d'après cette définition, il faut bien comprendre que tout document (document bureautique, messagerie électronique professionnelle, photographie numérique…) est une archive dès sa conception !

Ainsi, nous produisons, dans le cadre de nos activités de recherche des archives publiques. Ceci, loin d'être anodin nous engage fortement à un devoir de conservation à plus ou moins long terme. Peu de gens le savent, mais ces archives publiques, destinées à la mémoire collective, se doivent d'être maintenues dans un institut public et ne peuvent être détruites sous peine de sanctions sans le visa du service interministériel des archives de France. Leur destruction est donc encadrée par la loi et l'élimination ne se fait pas au hasard, mais selon un tableau de gestion. Le bordereau d'élimination, généré pour valider une destruction de document est ici un élément important pour justifier de la traçabilité des informations et les raisons de la destruction.

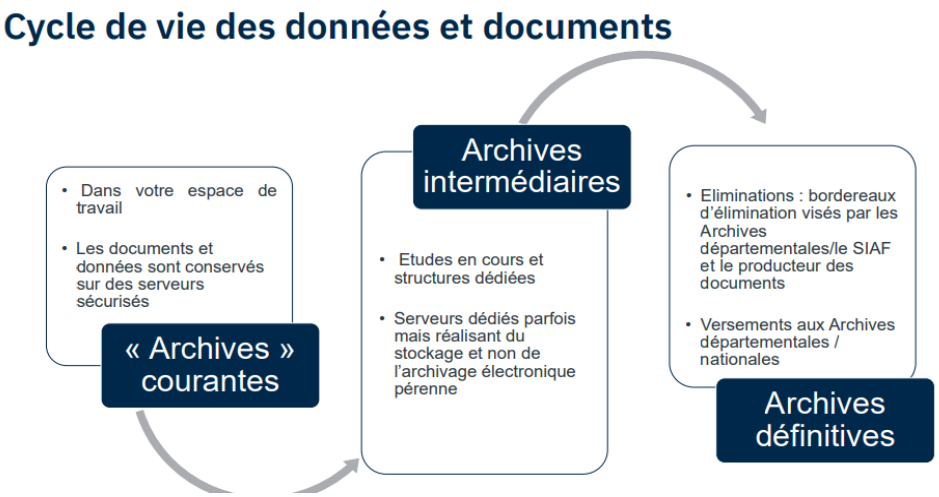
Cette disposition légale, si elle garantit le devoir de conservation, cause néanmoins, comme en témoigne Yonny Cardenas pour le Centre de Calcul de l'IN2P3, une difficulté majeure, celle de ne pouvoir justifier, alors qu'elles surchargent inutilement les serveurs, l'élimination de données volumineuses, « orphelines », inexploitables et intraçables, du fait de l'absence totale d'information sur leur contenu. Le Centre se trouve donc en situation de stocker des données mortes dans une proportion difficile à établir sans pouvoir juridiquement les supprimer ou les fouiller !

Quelques précisions ensuite pour mieux comprendre les différentes étapes d'archivage possible de nos données de recherche. En parallèle du cycle de vie de la donnée, co-existe pour les archivistes, un cycle de vie des données et documents. Il s'organise autour de 3 étapes :

- Les archives courantes : cette première étape correspond à la phase de production et manipulation au quotidien de données stockées sur des espaces de travail, et n'est donc pas « naturellement » considérée comme une étape d'archivage par les producteurs de données, elle l'est en revanche du point de vue de l'archiviste. Pour faciliter la gestion de l'information

et la transition vers l'archivage intermédiaire et définitif, il est important dès cette étape et très en amont d'adopter de bonnes pratiques de nommage et d'organisation.

- Les archives intermédiaires : cette seconde étape est une phase de transition au cours de laquelle le scientifique n'a pas un usage constant de l'information, les informations sont archivées afin de répondre à des besoins de suivi de projet face à des enjeux juridiques par exemple. Les données sont stockées sur des serveurs dédiés.
- Les archives définitives : cette étape ultime correspond au versement des données aux archives départementales ou nationales. On parle ici d'archives définitives ou historiques et il s'agit d'assurer, pour ces données, une préservation et un archivage à long terme ou ad vitam aeternam.



Source : Marie-Laure Bachelerie-Gouverneur, L'archivage de la recherche en théorie et en pratique, 2024

Cet archivage numérique présente de nombreux enjeux historiques, financiers, commerciaux, éthiques ou légaux et la gestion des archives est fortement corrélée à la gestion des espaces de stockage, elle facilite l'accessibilité au document réalisant ainsi un gain de temps précieux pour la recherche. Il est essentiel dans ces conditions de réfléchir le plus tôt possible à la préservation des données et, en fonction des enjeux à leurs délais de conservation.

2. Les archivistes face à la gestion des données de recherche

La mission des archivistes en établissement consiste à gérer les archives courantes et intermédiaires jusqu'à leur transfert aux Archives Nationales ou départementales pour leur conservation définitive ou jusqu'à leur destruction après tri. Stéphy Le Foll, archiviste au département robotique du LAAS, souligne l'urgence à gérer, sauver et rendre accessibles les archives scientifiques alors qu'elles sont généralement retrouvées à la faveur de travaux de rénovation de locaux et à la suite du départ à la retraite ou du décès d'un chercheur.

Les retours d'expérience des archivistes en laboratoire, font état de la difficulté d'opérer un tri dans la massification des informations collectées, lorsqu'elles ne sont pas sauvagement éliminées au mépris du code du patrimoine. Les archivistes soulignent à cet égard l'importance de gérer très en amont la documentation des données qui s'avèrent souvent très incomplètes ou inexploitables du fait de formats obsolètes. Véronique Ginouvès et Emilie Groshens, archivistes à la MMSH, mettent en garde contre la fragilité des outils ou plateforme de gestion qui disparaissent et se renouvellent constamment occasionnant un risque réel de perte de données. Ce paysage numérique en évolution impose dans la pratique une façon solide de décrire et d'organiser les documents avec une mention précise des acteurs et du contexte de production. La rédaction de plans de gestion de données

rétrospectifs qui permettent de lire et de comprendre le récit de ces données ainsi que l'ensemble des évolutions qui ont façonné leur production et leur catalogage est pour elles le moyen de faciliter la réutilisation des données sur le long terme.

De la même manière, Océane Valencia, responsable du service Archive Sorbonne Université recommande, pour l'archivage des codes et logiciels l'usage d'environnements logiciels comme Jupyter Notebook, RStudio, OrgMode ou MakDoc pour documenter les projets de recherche et assurer leur reproductibilité. Selon elle, la principale difficulté reste de récupérer les morceaux de code éparpillés dans les PC qui n'ont pas été déposés sur des forges logicielles. Seul le dépôt sur les plateformes de Software Heritage, du CINES, de la BNF ou des Archives Nationales permettrons toutefois de garantir la conservation et l'accès pérenne aux codes et logiciels et la constitution d'un patrimoine informatique.

Aurélie Montagne Borras, responsable du pôle archive à la MSH Monde à Nanterre, insiste quant à elle sur l'importance de la collaboration avec les producteurs de données pour appliquer, dans le respect du fonds, les principes archivistiques fondamentaux qui permettent de maintenir les documents dans leur contexte de création et de faciliter la constitution de sources historiques exploitables, fiables et authentiques.

Cette collaboration nécessaire à la gestion des archives et des données de recherche pose néanmoins quelques difficultés énoncées dans le livre blanc présenté en avant-première par Cyprien Henri, conservateur du patrimoine et responsable de la Mission des archives et du patrimoine culturel au MESRI.
La production des données, la science ouverte, le patrimoine sont vus par la gouvernance des établissements de plus en plus comme des enjeux forts, mais la gestion des données de recherche qui sont des archives publiques semble, selon la section Aurore, association d'archivistes, échapper de plus en plus à la profession.

Les archivistes, généralement identifiés comme en charge des archives administratives sont rarement intégrés aux laboratoires de recherche qui considèrent les données davantage comme des objets scientifiques que comme des archives de recherche.

Très peu nombreux et inégalement répartis sur le territoire, peu d'archivistes ont été en mesure d'intégrer les cellules d'aide à la recherche des institutions comme Humathèque, Progedo ou de l'écosystème Recherche Data Gouv, principalement investis par les professionnels de l'IST et les personnels de bibliothèques qui se sont ainsi saisi des archives de la recherche.

Face à ces évolutions, et le sentiment d'exclusion, assez largement ressenti par les archivistes, un groupe de travail informel composé d'archivistes en établissement et piloté par la Mission archives du ministère de l'Enseignement supérieur et de la Recherche s'est constitué pour rédiger un livre blanc des archives de la recherche.

Ce livre, qui a vocation à être relu de façon très large et par les « non archivistes » (chercheurs, informaticiens, professionnels de l'IST) a pour but de clarifier le rôle et l'apport de chacun dans la gestion des données de recherche et de réaffirmer l'apport métier spécifique des archivistes. Il a également pour objet de permettre à tous de mieux appréhender les différentes stratégies qu'il reste à concevoir pour mettre en œuvre l'archivage des données de recherche.

Avec en parallèle la volonté aussi de réformer la commission « archive scientifique » du conseil supérieur des archives (instance de conseil pour donner des orientations au SIAF), le livre blanc

souhaite relancer une démarche réflexive prospective sur le sujet et proposer une brique de formalisation vers une politique interministérielle de la donnée.

3. Les dispositifs en place pour la sélection et l'archivage et l'enjeu d'une stratégie archivistique

Les données de recherche ne peuvent pas être détruites de manière aléatoire, nous l'avons vu, mais en réalité seule une petite partie des données produites est destinée à être préservée et archivée sans limitation de durée. Ce sont les Archives Nationales et départementales qui en ont la charge.

Le Service Interministériel des Archives de France joue dans cette affaire un rôle central puisqu'il définit la politique de gestion des archives publiques et assure la coordination des services. Ses missions touchent aussi bien à la conservation physique et numérique des archives qu'à l'accès et à la valorisation de ces documents, tout en assurant la conformité avec les règles légales et scientifiques.

Au sujet de la collecte d'archives numérique, Dominique Naud, Cheffe du bureau de l'expertise numérique et de la conservation durable au SIAF a présenté une courbe de croissance moyenne annuelle de 30%, accentuée depuis 2020. Le volume des archives numérique est en expansion et il est clair que ce constat impose un passage à l'échelle pour accueillir les données produites dans les établissements de recherche. Elle assure que l'archivage numérique, déjà présent dans les objectifs cadre 2020-2024 est reconduit pour les années à venir et sera articulé avec les questions de transition écologique et de cybersécurité. Dans quelle mesure les services d'archives sont-ils toutefois en capacité d'accueillir les données de recherche ?

La courbe de croissance présentée (limitée à une échelle maximum de 60 terra octets) pose un problème d'ordre de grandeur en considération du volume de données produites dans nos établissements de recherche. Un accélérateur de particules comme le LHC du CERN est en capacité de produire 1 peta octet par seconde. Est-il envisageable, dans ce cas de figure, de conserver les données de recherche massives ? Ici peut-être plus qu'ailleurs, la question de la sélection des données à archiver s'impose comme une évidence.

Selon Cyprien Henri, toutes les données de recherche n'ont pas vocation à être archivées de manière définitive, leurs finalités pouvant varier. L'archivage définitif bien que relevant d'une obligation de moyens pour l'État, nécessite avant tout la mise en place d'une politique de patrimonialisation.

Dans certains cas, un « archivage en Y » peut aussi constituer une solution adaptée, il s'agit d'une méthode de gestion qui combine l'archivage d'un jeu de données complet pour un usage primaire d'exploitation scientifique (une réutilisation) et un archivage sélectif, voire restreint pour des objectifs de préservation historique à long terme.

Ainsi, il n'est pas nécessaire de reverser tous les pétaoctets de données générées. Il reste cependant crucial de définir les éléments à préserver pour l'histoire et de développer des outils adaptés aux différents usages des données, qui peuvent parfois coexister dans le temps.

Actuellement, les outils archivistiques, conçus pour des besoins administratifs, nécessitent une évolution pour répondre aux spécificités scientifiques. Par exemple, la « durée d'utilité administrative » appliquée aux archives intermédiaires ne correspond pas aux données de recherche pour lesquelles la notion de « durée d'exploitation » conviendrait mieux. Cette durée d'exploitation peut néanmoins s'avérer extrêmement longue pour certaines données en astronomie collectées il y a plus d'un siècle et toujours en usage !

Aussi, pour correspondre aux réelles capacités d'archivage définitif, une stratégie doit être élaborée afin de déterminer ce qu'il convient de conserver dans un but patrimonial et pour quel usage.

Pour assurer l'archivage numérique et accompagner le versement des données aux archives, des moyens et des solutions ont été développé par le SIAF qui propose entre autres :

- Le programme VITAM ((Valeurs Immatérielles Transmises aux Archives pour Mémoire) associé à un arsenal logiciel qui existe sous forme de back-office et de services. VITAM est une solution d'archivage numérique pérenne, sécurisée et interopérable, capable de gérer les archives électroniques à long terme, tout en garantissant leur intégrité, leur accessibilité et leur confidentialité.
- Un Dispositifs d'Accompagnement des Missions pour l'Archivage Numérique (DIAMAN) vise dans le même temps à soutenir financièrement les administrations dans leurs projets d'archivage numérique.
- Des outils comme Archifiltre, Octave, ReSIP ou PASTIS soutenus par l'état, téléchargeables, ouverts et documentés peuvent être utilisés pour simplifier l'archivage, l'analyse et la conformité aux normes de conservation.
- Le service interministériel crée par ailleurs des études et des partenariats sur les formats, métadonnées, stockage et support notamment avec l'association Aristote et la cellule nationale de veille sur les formats.

Le CINES (Centre Informatique National de l'Enseignement Supérieur), acteur central pour les institutions académiques, propose également des solutions d'archivage à long terme et accompagne les services versants dans la mise en œuvre de processus métier et qualité pour prétendre notamment à des certifications.

Comment faciliter les versements d'archives ?

Une collaboration entre entrepôts de données et SAE pourrait offrir un modèle de gestion de données de recherche plus complet et durable. Des questions persistent néanmoins sur les liens potentiels entre le développement des Systèmes d'Archivage Electronique (SAE) comme VITAM et les données de recherche déposées dans les entrepôts de données. En effet, les entrepôts ne sont pas conçus pour être des SAE, car leur objectif n'est pas la conservation à long terme, tandis que les SAE, bien qu'orientés vers la préservation durable, n'ont pas pour vocation de conserver l'ensemble des données.
Sorbonne Université a été retenue dans le cadre du dispositif DIAMAN, pour mener une étude de faisabilité sur le déploiement d'un SAE pour le dépôt, la gestion et la conservation des jeux de données de la recherche à « communicabilité restreinte ». Pour Océane Valencia, tout l'enjeu est précisément de déterminer les liens et interopérabilités possibles entre différentes plateformes pour la mise à disposition d'une part de données ouvertes et la préservation et sécurisation à long terme des données non-communicables d'autre part. Pour atteindre un niveau d'automatisation optimal dans ce projet complexe, les aspects techniques, comme l'harmonisation des métadonnées, sont certes essentiels, mais les dimensions organisationnelles le sont tout autant. Il est, en effet, crucial de définir précisément les acteurs et les rôles qui leur sont attribués.

Pour Philippe Prat, responsable du Département Archivage et Services aux Données au CINES des opérations de rationalisation sont à mettre en œuvre au niveau des services communs. Différentes compétences métier, archivistiques et techniques sont nécessaires pour s'interfacer efficacement avec un SAE et élaborer des standards de description métier.

S'agissant d'elabFTW[15], présentée par Henri Valeins, chargé de mission auprès du DGDR[16] pour l'offre de service des cahiers de laboratoire électronique du CNRS, une réflexion est amorcée sur l'archivage des données de laboratoire électronique. Le cahier de laboratoire consigne en détail le cheminement intellectuel des scientifiques. Traditionnellement papier, notamment en biologie, ce format présente des limites. Le CNRS a donc lancé en 2020 une initiative pour introduire elabFTW dans ses instituts.

Concernant l'archivage des données des cahiers de laboratoires, bien qu'il n'existe pas de freins techniques majeurs grâce à la possibilité de développer une API pour faciliter le dépôt des données au CINES, plusieurs questions subsistent. Il s'agit de déterminer la durée de conservation des données, leurs modalités d'accès, les conditions de leur réutilisation, et les standards de documentation à adopter pour assurer leur utilité future. Il est également nécessaire de clarifier le moment opportun pour l'archivage, les responsabilités associées et le besoin exact en matière d'archivage, autant d'éléments qui nécessitent une concertation et un retour d'expérience pour s'ajuster aux besoins réels.

Actuellement, très peu de données de recherche sont versées aux archives départementales. La démarche n'est pas systématique et chaque fond d'archive est un projet en soi dont on discute le bien-fondé du versement. Les archives ne sont par ailleurs pas toutes dotées de la même manière en matière de SAE, n'ont pas les mêmes capacités d'intégration des données volumineuses ni même les mêmes schémas de spécification d'entrée, ce qui pose des problèmes d'égalité de traitement au niveau du territoire.

Le programme VITAM a été conçu pour traiter de grandes masses de données, il permet d'organiser, de sécuriser et de documenter les flux d'informations, ce qui est essentiel, mais les solutions se construiront au fur et à mesure. Il s'agit désormais de s'appuyer sur des standards interopérables pour pouvoir faire communiquer les plateformes et faire voyager les données.

4. Les difficultés et facteurs de réussite dans la mise en place d'un écosystème pour le partage et la préservation des donnés

Des projets sont déployés au sein de nos infrastructures et organismes pour répondre aux besoins de gestion et préservation des données de la recherche. Les enjeux sont importants et nécessitent la mise en place de solutions techniques mais aussi opérationnelles.

Quatre d'entre eux nous ont été exposés et donnent un aperçu des difficultés rencontrées :

- Un projet pilote « Data Future » a été initié au CCIN2P3 pour étudier la faisabilité d'implémentation d'un service de préservation de données à long terme en suivant le modèle OAIS[17]. Le service est basé sur une action volontaire du producteur de données et l'objectif est la réutilisation des données. Il s'inscrit en complément de la mise en place de catalogues de données et d'un service d'accompagnement à la rédaction de PGD.
- Au sein de l'infrastructure Data Terra qui regroupe l'ensemble des données du système terre, c'est une plateforme numérique intégrée qui a été mise en place grâce au projet GAIA data

[15] https://www.elabftw.net/
[16] Directeur Général Délégué à la recherche du CNRS
[17] Open Archival Information System ou système ouvert d'archivage d'information. L'OAIS est un modèle conceptuel destiné à la gestion, à l'archivage et à la préservation à long terme de documents numériques. https://fr.wikipedia.org/wiki/Open_Archival_Information_System

pour permettre à l'utilisateur de mener des projets multi-source, multi-échelle, de plus en plus compliqués avec des données issues de nombreux pôles de données.

- Le Muséum d'Histoire naturelle, riche de quatre siècles d'histoire, a quant à lui choisi de mettre en place une gouvernance des données scientifiques en s'appuyant sur un Administrateur des Données, Algorithmes et Codes de la recherche (*ADAC*), un réseau de référent données dans les unités ou entités et un guichet ouvert de la donnée en cours de construction afin d'amener des bonnes pratiques de gestion au sein de l'établissement.
- L'Université d'Aix Marseille, pour finir a souhaité développer sa politique science ouverte sur des infrastructures existantes dont un data centre régional sur site, un mésocentre labellisé, l'IR Humanum et une offre de service couvrant l'ensemble du cycle de vie de la donnée proposé par son Centre De formation et de soutien aux Données de la REcherche (CEDRE).

Les présentations ont clairement souligné que l'enjeu majeur pour ces services réside dans la capacité à fédérer les ressources, soutenir les équipes, et favoriser l'ouverture et le partage des données produites. Toutefois, ce travail est complexe à mettre en place, voire colossal pour des infrastructures telles que Data Terra ou des institutions comme le Muséum d'histoire naturelle. Ces structures, qui regroupent à l'échelle nationale, européenne et internationale de nombreux partenariats, doivent composer avec une grande diversité d'infrastructures et une multitude d'entités aux finalités et modes de fonctionnement spécifiques.

Cécile Callou, responsable de la mise en place de la gouvernance des données et ADAC au Muséum National d'Histoire Naturelles témoigne dans son établissement de la difficulté de se parler en interne pour faire converger des pratiques de production et de normalisation hétérogènes bien que partageant des données communes.

Joël Sudre, coordinateur de GAIA Data, souligne la difficulté d'organiser l'interopérabilité des services dans un cadre composite qui intègre le moissonnage, la consultation et l'analyse des données provenant de catalogues spécifiques à chaque domaine d'observation. À cela s'ajoutent des données externes issues de projets européens, le tout soutenu par huit centres de calcul haute performance (HPC) répartis sur le territoire et 30 data lakes interconnectés. Il met également en évidence la complexité de la coordination humaine sur l'ensemble des projets dans ce contexte.

Pour relever ces défis, la possession d'une infrastructure adéquate est essentielle, mais les leçons tirées de l'expérience au CCIN2P3 montrent que cela ne suffit pas. La mutualisation joue un rôle clé dans le financement et la préservation des données scientifiques, et il est impératif de mettre en place un écosystème numérique complet, incluant centres de données, entrepôts et services de formation.

La structuration en réseau, la collaboration et la mutualisation sont des facteurs clé de réussite dans ce type de projet, comme le montre l'exemple du Muséum et de l'Université d'Aix-Marseille. Au Muséum, un travail collaboratif autour du concept de « spécimen étendu » ou « spécimen connecté » a permis de rassembler diverses informations (taxonomiques, observations, génomiques, imagerie 2D et 3D, etc.) sur un même spécimen. Ce projet a facilité les connexions entre différentes expertises et types de données, tout en permettant de dresser un état des lieux collectif pour identifier les lacunes et développer une politique de gouvernance. Cependant, la question de l'élimination des données se pose, comme pour tous les projets de ce type.

Cécile Callou insiste sur l'importance, avant de supprimer des données, d'identifier les manques. Cette démarche nécessite un travail collaboratif et une réflexion sur les améliorations à apporter au fonctionnement interne, ainsi que sur l'expansion des domaines de compétence et de recherche. La solution pour connecter les projets internes et externes et créer l'expérience collective nécessaire à la

sélection des données repose sur une structuration réfléchie des services et une mise en réseau de compétences complémentaires. Un data hub est actuellement en cours de création au Muséum pour connecter les systèmes d'information, homogénéiser et enrichir les données en interne. Cette étape est indispensable avant de pouvoir envisager la suppression des données sans risque.

De la même manière, l'Université d'Aix-Marseille cherche à structurer les expertises et compétences, souvent dispersées, pour soutenir efficacement ses équipes de recherche. La qualité des données et le développement d'une offre de service évolutive selon les besoins de la recherche sont essentiels. Toutefois, la formation et la sensibilisation des ingénieurs et techniciens occupent également une place centrale dans la gestion et la préservation des données de recherche. Selon Fabien Borget, chargé de mission science ouverte à l'université d'Aix Marseille et co-animateur du chapitre national français CoARA[18], il est crucial d'anticiper la montée en compétences des personnels, d'ajouter de nouvelles expertises et de définir de nouveaux métiers en lien avec les données pour mettre en place des dispositifs d'accompagnement efficaces.

Enfin, bien que la structuration et le dépôt des jeux de données dans les carrières restent un point sensible, il est désormais nécessaire de faire évoluer les pratiques pour en faire une étape incontournable.

5. La gestion des risques et préservation des données dans les entrepôts, les bonnes pratiques et les opportunités

Les risques de perte de données sont fréquemment évoqués lorsqu'il s'agit de la gestion des données de recherche. Assurer la préservation de ces données revient à faire face à une série de menaces. Selon Olivier Rouchon, responsable des données de recherche à la DDOR et président du Comité Directeur du Core Trust Seal, le risque résulte de la combinaison d'une vulnérabilité et d'une menace. On distingue plusieurs types de risques, notamment techniques (obsolescence, détérioration des supports), naturels (catastrophes, incendies), organisationnels (manque de ressources humaines), financiers, etc.

Bien qu'il soit impossible de se prémunir contre tous les dangers, il est possible de s'appuyer sur une méthodologie de gestion des risques pour gérer sereinement un entrepôt de données. Un plan d'action (plan de gestion des risques) peut être élaboré pour rendre les risques acceptables. Cette démarche se fait de manière collégiale, impliquant divers acteurs (archivistes, informaticiens, décideurs...), et vise à évaluer la probabilité des risques, leur impact, et à mettre en place des mesures adaptées (duplication hors site en cas de risques naturels, plan RH pluriannuel pour accompagner le personnel en cas de risques organisationnels, application de normes et standards interopérables pour les risques techniques, etc.). Cette approche garantit une qualité technique et organisationnelle, assurant la répétabilité des processus et le traitement homogène des données. Elle permet également de formaliser les processus métiers.

La certification Core Trust Seal (CTS) peut prolonger cette démarche. Le CTS est un standard international de certification qui évalue la fiabilité des dépôts de données numériques, en s'assurant qu'ils respectent de bonnes pratiques en matière de préservation et de gestion des données. Basée sur 16 critères de référence, cette certification est accordée aux dépôts qui prouvent leur capacité à préserver les données sur le long terme, en garantissant leur accessibilité, leur intégrité et leur réutilisabilité. Elle constitue un label de confiance pour les utilisateurs et permet aux administrateurs d'entrepôts de s'engager dans un processus d'auto-évaluation.

[18] Coalition on Advancing Research Assessment (*CoARA*) : https://coara.eu/

La table ronde dédiée aux entrepôt de données a permis de discuter plus particulièrement des pratiques en cours au sein de 3 entrepôts de données engagées dans une démarche de qualité ou de certification : Nakala, Data Indores et GBIF.

- Nakala, pour commencer est un entrepôt de données dédié à la préservation et la dissémination des données produites dans le cadre de projets de recherche français en Sciences Humaines et Sociales. Ce service est proposé par l'infrastructure de recherche Huma-Num.
- Data Indores est un entrepôt thématique en écologie environnement, déployé au centre de calcul de l'in2p3. Il est soutenu par le Museum National d'Histoire Naturelle et CNRS Ecologie et environnement. C'est aussi un des nœuds de GAIA Data, du PNDB (Pôle National de Données de Biodiversité).
- GBIF (Global Biodiversity Information Facility) enfin est un réseau international et une infrastructure de données qui a vocation à faciliter et promouvoir l'accès ouvert aux données sur la biodiversité.

Pour améliorer la qualité du dépôt et garantir la préservation à long terme des données, ces entrepôts ont mis en place plusieurs mesures visant à assurer la fiabilité, la pérennité et la réutilisabilité des données.

Ils ont d'abord adopté une politique de gestion des données claire, définissant des conditions de dépôt validées par un circuit de validation qui garantit la bonne destination des données et leur rattachement à l'organisme de recherche concerné. Des exigences spécifiques concernant les formats, les métadonnées et l'organisation des fichiers sont appliquées, avec un système de modération en place. Des conditions de confidentialité et d'accès sont également établies, comme dans le cas du GBIF, qui a mis en œuvre une procédure de floutage pour protéger l'accès aux données sensibles.

Afin de favoriser l'interopérabilité et l'ouverture des données, ces entrepôts alignent leurs fonctionnalités avec les principes FAIR (Findable, Accessible, Interoperable, Reusable). Ils utilisent des normes reconnues pour la documentation et les métadonnées, telles que le Darwin Core et le langage EML (Ecological Metadata Language) pour le GBIF, ou le Dublin Core qualifié pour Nakala. Ces normes encouragent une documentation complète et détaillée des jeux de données. Un identifiant unique est systématiquement attribué à chaque jeu de données, accompagné d'une licence permettant d'améliorer la traçabilité, la citabilité et la réutilisation des données.

Par ailleurs, des guides pratiques, des supports techniques, ainsi que des formations et un accompagnement sont proposés aux utilisateurs. Ces services sont assurés par des structures comme le pôle de données d'HumaNum et le réseau de référents données de Data Indores, afin d'aider les chercheurs et les déposants à comprendre les exigences de qualité et de préservation, et à se former à l'utilisation des outils et services de dépôt.

Lors de la table ronde, la question de l'engagement des entrepôts sur le long terme a suscité un débat. Si ces pratiques permettent d'assurer un niveau de qualité, d'accessibilité et de durabilité des données, la question de la préservation à long terme reste ouverte, notamment en ce qui concerne la définition du "long terme". Les entrepôts ne s'engagent généralement pas à conserver les données au-delà de 10 ans, en raison de la difficulté de garantir la pérennité de l'environnement qui permet à l'entrepôt de fonctionner. Le long terme est une période difficile à définir, mais elle concerne une période de temps suffisamment long pour que des changements technologiques majeurs surviennent. Les procédures mises en place au sein des entrepôts, notamment celles guidées par la certification Core

Trust Seal, sont conçues pour absorber ces changements technologiques (conversion de formats, évolution des supports), garantissant ainsi la pérennité des services et des données.

Une autre question abordée concernait la possibilité d'une passerelle entre les entrepôts et le CINES, en particulier pour Nakala qui collabore déjà avec cet organisme. Hélène Jouguet, responsable du pôle donnée et accompagnement des utilisateurs à Huma-Num, a précisé que le circuit d'archivage de corpus de données par HumaNum au CINES existait avant Nakala, et que ce dernier ne propose pas encore ce service, bien qu'il envisage de verser des données au CINES à l'avenir. Cependant, il est nécessaire de procéder à une sélection et à une concertation préalable avant de verser des données. Elle a ajouté que la préservation des données par le CINES, en tant que tiers archiveur, constitue une forme de coffre-fort, où l'accès aux données est limité au seul service versant, qui conserve généralement son propre outil de diffusion pour rendre les données accessibles.

Enfin, la question de l'implémentation de métriques dans les entrepôts a été soulevée : est-il possible d'évaluer l'utilité et l'intérêt de l'archivage des données ? Bien que la question soit encore en cours d'étude, des systèmes de suivi, des indicateurs de téléchargements et d'utilisation des données pourraient selon Amandine Hénon, être utiles pour évaluer les tendances de réutilisation au sein de la communauté, et encourager les pratiques de dépôt et de partage. GBIF dispose déjà d'un indicateur de suivi des citations des jeux de données dans les publications.

6. Les défis et solutions pour une production de données responsable dans un monde contraint

Dans un contexte de réchauffement climatique, la recherche et la gestion des données doivent évoluer pour répondre aux enjeux environnementaux. Il est impératif de réduire drastiquement les émissions de gaz à effet de serre générées par le numérique en adoptant des pratiques de recherche plus durables. Les présentations d'Alexis Arnaud, ingénieur de recherche à l'UGA et membre du GDS CNRS EcoInfo et de Guillaume Levavasseur, ingénieur de recherche à l'Institut Pierre Simon Laplace, ont mis en évidence une croissance exponentielle de la production de données dans la recherche scientifique, remettant en question nos méthodes de travail.

La recherche, largement numérique, repose sur la collecte, le traitement, le stockage et la diffusion des données. Cependant, il est difficile de quantifier précisément les impacts environnementaux du numérique, car celui-ci est omniprésent dans l'économie. Alexis Arnaud met en garde contre les effets d'un secteur en constante expansion, qui s'auto-alimente, crée de nouveaux besoins et attire toujours plus d'utilisateurs. Le numérique s'appuie également sur des objets matériels dont la production, depuis l'extraction des ressources jusqu'à leur fin de vie, a des conséquences environnementales importantes : consommation énergétique, changement climatique, pollution, toxicité, déplétion des métaux, entre autres.

Dans le domaine de la simulation climatique, Guillaume Levavasseur, qui contribue à l'élaboration du rapport du GIEC, constate que la multiplication des expériences et des trajectoires de modélisation allonge considérablement les temps de calcul. Cela soulève des questions tant écologiques que scientifiques, rendant parfois impossible l'analyse complète de tous les modèles produits. Se pose alors la question des moyens pour adapter les pratiques de recherche et de gestion des données aux enjeux environnementaux actuels et pour réduire le coût des données.

Certaines solutions sont déjà mises en œuvre. En modélisation climatique, il est par exemple possible de mieux sélectionner les données via des outils de filtrage (implémentées avec la communauté), afin de ne conserver que celles jugées nécessaires. De même, Jean-François Perrin, responsable informatique, a proposé des solutions au sein de l'ESRF (European Synchrotron Radiation Facility). Des mesures y sont appliquées pour limiter les données archivées, notamment en introduisant une

nouvelle politique de curation et de triage. Des outils ont été développés dans la communauté pour mieux comprendre comment les données sont utilisées et l'ESRF encourage les utilisateurs à utiliser des DOI pour suivre l'usage des données jusqu'à leur publication scientifique.

En partant du principe qu'une donnée non publiée est une donnée inutilisée, ce travail de liaison entre l'usage des données et leur production bénéficie d'une estimation du temps écoulé entre l'expérience et la publication. Cette information permet potentiellement de décider de manière éclairée s'il est pertinent ou non d'archiver une donnée.

Une concertation avec la communauté a permis de mettre en place des solutions pratiques pour éliminer, après analyse, les données sans valeur scientifique dès leur sortie du détecteur ou, lorsque cela est possible, de compresser avec perte afin de réduire le volume de données stockées. Cela a également conduit à faire évoluer les pratiques vers un service d'analyse au plus près de l'acquisition, en se concentrant sur l'archivage de données traitées à haute valeur ajoutée, réduisant ainsi le volume global des données et l'impact carbone associé.

La science ouverte vise à apporter des réponses aux défis de qualité, de diffusion et de réutilisation des données, ce qui permet de limiter les impacts environnementaux. Elle invite également à réfléchir aux données à préserver et au moyen d'y parvenir de façon durable. Bien qu'aucune solution universelle ne s'applique à tous les projets, Alexis Arnaud souligne l'importance d'intégrer les impacts environnementaux et sociétaux dans les outils de la science ouverte. Ralentir le rythme de production, réfléchir aux impacts écologiques à chaque étape du cycle de vie des données et utiliser un plan de gestion de données sont essentiels pour évaluer le coût environnemental de la recherche. Le plan de gestion doit toutefois être adapté aux besoins de sorte à questionner les besoins et finalités de la production des données de recherche.

Pour optimiser les coûts financiers et l'empreinte environnementale, le message principal porté par Denis Veynante, Directeur adjoint de la DDOR est de ne surtout pas développer sa propre solution de stockage, mais de s'associer aux centres existants.

Des solutions sont déjà en place, notamment l'entrepôt Recherche Data Gouv[19], une plateforme souveraine proposée par le ministère de l'Enseignement supérieur et de la Recherche pour le stockage des données de recherche, bien que celle-ci limite la capacité à 50 Go par dépôt et 5 To par organisme.

Des études de cas menées à la DDOR[20] ont permis d'identifier dans certaines communautés comme celle de la mécanique des fluides, des besoins de stockage massifs pour permettre la validation de modèles de simulation complexes et coûteux associés à d'importantes capacités de calculs et d'hébergement de matériels informatiques.

Ces études ont mené au lancement d'un projet national de centre de données, soutenu par la DGRI[21], visant à répondre aux besoins spécifiques non couverts par Recherche Data Gouv et les infrastructures actuelles, et conçu pour compléter le projet FITS, dédié aux infrastructures de recherche.

Ce projet, porté initialement par l'IDRIS[22] puis étendu à d'autres centres, a pour objectif d'analyser les besoins, de déployer une solution nationale, d'interconnecter différentes infrastructures de stockage avec des mésocentres et d'étudier les modèles économiques pour garantir la durabilité de

[19] Entrepôt Recherche Data Gouv : https://recherche.data.gouv.fr/fr/page/entrepot-recherche-data-gouv
[20] Direction des données ouvertes de la recherche du CNRS
[21] Direction générale de la recherche et de l'innovation du ministère de l'Enseignement supérieur et de la Recherche
[22] Institut du développement et des ressources en informatique scientifique du CNRS

l'infrastructure. Le déploiement du data centre est prévu pour mi-2025, tandis que le projet FITS devrait voir le jour d'ici 2029.

Cette stratégie de mutualisation et de rationalisation des infrastructures dans les data centres labellisés, encouragée par le Ministère et soutenue par le CNRS, représente un levier essentiel pour allier efficacité économique et responsabilité écologique, en répondant aux enjeux de durabilité et de maîtrise des coûts associés à la gestion des données.

7. Les pratiques et stratégies pour la sélection des données de recherche

Un archivage efficace garantit la préservation de données pertinentes. Dans ce processus, la question de l'évaluation, du tri et de la sélection des données est essentielle pour réduire les coûts de stockage, maximiser l'utilité des données et assurer que ce qui est conservé répond aux besoins de la recherche future et du patrimoine scientifique. Ce processus de sélection, qui implique divers acteurs, doit être anticipé dès les premières étapes de la recherche. Par ailleurs, pour une préservation efficace des données à moyen et long terme, il est crucial de disposer d'informations contextuelles suffisantes, afin que ces données puissent être gérées et réutilisées par d'autres. Le rôle des producteurs de données est central dans cette démarche, bien qu'ils n'en soient pas toujours pleinement conscients.

Pour explorer comment le tri et la sélection des données sont actuellement gérés dans différentes disciplines, une table ronde a réuni six communautés, chacune présentant ses pratiques en matière de gestion et d'archivage des données de recherche. Julien Peloton a représenté l'astrophysique, Françoise Genova l'astronomie, Denis Veynante la mécanique des fluides, Stéphane Renault l'archéologie, Isabelle Charpentier l'écologie et l'environnement, et Pierre Susbielle la robotique. Leurs interventions ont offert un aperçu des pratiques en vigueur pour la sélection des données à conserver.

Les présentations introductives ont montré une grande diversité dans la volumétrie et les formats des données, en fonction de leur provenance, qu'elles soient collectées sur le terrain ou produites dans des centres de calculs puissants. Les intervenants ont souligné cette hétérogénéité des pratiques, notamment en astrophysique, où les collaborations internationales influencent les méthodes de gestion des données, ou en archéoscience, où la diversité des communautés autour d'un sujet commun engendre des pratiques variées. Cette diversité se reflète également dans la création et l'adoption de standards : en astrophysique, le format FITS est largement utilisé, tandis qu'en archéologie, il existe plusieurs standards, et en astronomie, les formats définis par l'International Observatory Alliance sont nombreux et largement adoptés.

Sans réelle incitation ou en l'absence d'entrepôts thématique, certaines communautés, comme celles de la robotique et de la mécanique des fluides, ne sont pas encore habituées au partage des données. En robotique, l'enjeu principal est d'inciter les chercheurs à déposer leurs données et à garantir leur reproductibilité. En revanche, l'astronomie est un pionnier du partage de données et l'astrophysique, bien qu'elle ne soit pas systématique, pratique également l'ouverture des données. En archéologie, l'ouverture des données existe dans des réseaux structurés et des consortiums, mais il y a encore peu d'incitations directes à ouvrir davantage les données. En écologie et environnement, les pratiques sont variées et parfois inexistantes, avec pour principale difficulté la collecte de données sur le territoire.

Dans les observatoires, notamment en astronomie, les données sont conservées sur de longues périodes, car elles sont uniques et essentielles pour la recherche. La NASA a mis en place des archives thématiques, et des centres comme le CDS facilitent l'accès aux données de référence. Ces données sont largement réutilisées.

En robotique, bien que l'utilité des données ne soit pas toujours prévisible, la communauté estime que de nouveaux usages pourraient émerger. Il n'est pas utile de stocker les données à vie car tout est reproductible mais il y a un fort enjeu de pérennisation sur certains jeux d'équation ou modèles.

A l'inverse, en archéologie, les données de terrain sont difficilement reproductibles. La sélection des données numériques est un processus complexe à mener et la réflexion n'est pas aboutie. Le besoin de numériser d'importantes archives papier est également une préoccupation majeure.

En mécanique des fluides, les données de simulation sont conservées uniquement pendant la durée du projet (généralement un an) et disparaissent avec la fin des travaux du thésard. Ces données ont une durée de vie limitée à 5 ans, car elles sont reproductibles avec une meilleure précision à moindre coût. Les rares données anciennes, comme celles des études d'écoulements turbulents, sont encore utilisées, mais elles sont devenues obsolètes. En astrophysique, bien que peu de données soient triées, la tendance évolue vers une gestion plus sélective, avec pour principe de limiter la destruction des données. Pour la communauté écologie et environnement, la priorité est d'encourager la communauté à participer à la science ouverte et à partager les jeux de données, la question de la sélection n'étant pas encore au cœur des préoccupations.

Face à cette diversité de pratiques disciplinaires, il est difficile de définir une solution universelle pour la sélection de données. Cependant, plusieurs stratégies, soulevées lors des discussions, permettent d'aborder cette complexité :

- Adopter une méthode et une approche contextualisée

Plutôt que de chercher une méthode unique, il est essentiel de développer des critères de sélection spécifiques à chaque contexte disciplinaire, en analysant les besoins et pratiques propres à chaque communauté.

- Définir un cadre flexible de sélection et des critères adaptés

Un cadre général flexible peut-être définit offrant des lignes directrices permettant des ajustements disciplinaires. Divers critères peuvent être pris en compte pour la sélection des données tels que la qualité, la pertinence, la complétude, le potentiel de réutilisation et de reproductibilité, le niveau de curation, et les coûts associés.

- Utiliser des outils technologiques et des algorithmes

L'intelligence artificielle pourrait contribuer à la sélection des données de recherche en automatisant des tâches complexes et coûteuses. Une partie de la communauté robotique par exemple s'y intéresse fortement tandis qu'une autre s'en détourne totalement. Les indicateurs définis par les communautés et les institutions peuvent également jouer un rôle dans cette sélection, bien qu'il faille prendre en compte les biais possibles.

- Considérer le facteur « non-reproductibilité »

Le caractère non reproductible de certaines données, comme celles collectées en archéologie, pourrait constituer un critère naturel pour leur conservation. Toutefois, cette approche soulève la question de la finalité de la conservation : faut-il conserver une donnée pour qu'elle soit reproductible ou réutilisable ?

- Favoriser une approche disciplinaire et reconnaître l'ensemble des producteurs de données

Les activités liées aux données doivent être reconnues pour faire évoluer les pratiques. En écologie, environnement, le travail de terrain se fait avec les acteurs du territoire. Cette participation implique une reconnaissance du travail accompli, point fondamental pour obtenir les accords préalables à tout partage et toute ouverture des données. Cette approche est de nature à favoriser l'élaboration d'un processus de sélection inclusif.

Reconnaître les données à leur juste valeur, les auteurs scientifiques, et le personnel de soutien dans les publications scientifiques constitue également un préalable pour établir une culture de collaboration, de transparence et pour garantir des pratiques de recherche éthiques.

- Envisager une évaluation des données par les pairs

L'idée de mettre en place un peer reviewing spécifiquement pour l'évaluation des données de recherche permettrait de vérifier des critères clés tels que la qualité, la robustesse des données, et la présence de biais. Bien que cette approche puisse améliorer la confiance dans les résultats, sa mise en œuvre est complexe et coûteuse, nécessitant des experts qualifiés pour évaluer des jeux de données complexes. De plus, les critères d'évaluation restent à définir.

La valeur scientifique n'est pas le seul critère de sélection des données à archiver. Le processus dépend aussi fortement de la finalité visée, car les données ont une valeur patrimoniale significative, en tant qu'éléments clé de l'héritage scientifique, culturel et historique. Elles ne sont pas seulement destinées à être reproduites ou réutilisées, mais aussi à documenter les connaissances scientifiques, illustrer l'évolution des méthodes et des valeurs à une époque donnée, ou encore à offrir un aperçu de la société à un moment précis. Chaque finalité implique des critères de sélection différents.

Comment sélectionner les données patrimoniales ? Les demandes d'accès aux archives des chercheurs sont insoupçonnées en raison de leur valeur multiple et de l'importance des archives scientifiques pour la mémoire collective. Afin de mieux comprendre les raisons pour lesquelles ces données devraient être conservées, il serait pertinent d'analyser les demandes d'accès et de faire appel à des historiens des sciences ou des épistémologues. Toutefois, anticiper les besoins futurs en matière de données est un défi, car il est difficile de prévoir les évolutions technologiques à venir. Si le stockage ne sera peut-être plus un problème dans le futur grâce aux avancées technologiques, la gestion de l'organisation, de l'indexation et de l'accessibilité des données reste un défi majeur.

Conclusion

Le séminaire "RIP Data" a abordé des problématiques majeures pour la gestion des données de recherche et a permis de traiter des thèmes clé, notamment l'archivage, la préservation, les enjeux environnementaux et financiers, et la question du tri des données à conserver.

Les discussions ont souligné l'importance de l'archivage numérique et les défis liés à sa mise en œuvre. Les différents acteurs institutionnels (comme le Service Interministériel des Archives de France et le CINES) mettent en place des solutions, mais le volume croissant des données soulève la question de leur sélection pour éviter une surcharge des infrastructures de stockage. Les débats ont mis en exergue la nécessité d'une collaboration entre les chercheurs, les archivistes, et les institutions pour élaborer des normes et pratiques de sélection adaptées. Ils ont aussi insisté sur l'impact environnemental de la gestion des données, pointant des pratiques de gestion éco-responsables. Enfin, chaque discipline présente des besoins spécifiques en matière de conservation et de sélection des données.

La gestion des données de recherche impose de prendre du recul sur nos pratiques scientifiques et d'intégrer des choix réfléchis de sélection et de tri. Les échanges du séminaire ont révélé que cette gestion doit aussi prendre en compte les contraintes environnementales et financières. La rationalisation et la mutualisation des infrastructures, ainsi que l'implication de toutes les parties prenantes, sont des axes de progrès essentiels. À terme, la réussite de ces initiatives repose sur un cadre flexible, adapté à chaque discipline, et un engagement collectif.

Accès aux présentations des journées