L'archivage des données de la recherche à l'Inra. Eléments de réflexion, démarche et perspectives

Lina Sbeih, Fanny Dedet, Patrick Moreau, Esther Dzale

L'archivage des données de la recherche à l'Inra
Eléments de réflexion, démarche et perspectives

Lina Sbeih[1], Fanny Dedet[2], Patrick Moreau[3], Esther Dzalé[4]

Résumé. Dans un contexte de science ouverte (Open science), le partage et l'ouverture des données de la recherche (Open Data) emporte la nécessité de garantir leur accès, leur lisibilité et leur intelligibilité sur des durées plus longues que celles des projets qui en sont à l'origine. L'archivage pérenne des données devient par conséquent un sujet crucial. Cet archivage présente également des enjeux patrimoniaux et économiques : réduction des coûts liés à la multiplication des copies, à la perte de données stratégiques, au temps passé à retrouver et comprendre des données insuffisamment documentées.
Le but de cet article est double : (1) expliciter le concept de l'archivage pérenne et en décrire les enjeux, et (2) présenter l'état de l'art sur ce sujet à l'Inra ainsi que les actions envisagées pour proposer des services et des outils aux agents.

Mots-clés : Archivage numérique, archivage pérenne, données de la recherche, archives de la recherche, cycle de vie des données

Introduction

Dans un contexte de science ouverte (ou Open Science) promu par l'Inra, la transparence et la reproductibilité des résultats de la recherche représentent des enjeux majeurs. Par ailleurs, le développement des nouvelles technologies rend la recherche et la science de plus en plus collaboratives et axées sur les données, conduisant à un accroissement sans précédent de leur volumétrie. Le partage ou l'ouverture des données de la recherche (Open Data) emporte la nécessité de garantir leur accès, leur lisibilité et leur intelligibilité sur des durées plus longues que celles des projets qui en sont à l'origine. L'archivage pérenne des données qui consiste précisément à garantir l'accès, la lisibilité et l'intelligibilité sur le long terme devient par conséquent un sujet crucial pour la communauté scientifique dans son ensemble, les politiques, les financeurs de la recherche et les éditeurs scientifiques. L'archivage des données présente également des enjeux patrimoniaux et économiques : réduction des cas de reproduction (et donc de refinancement) inutile de données, des coûts liés à la multiplication des copies, à la perte de données stratégiques, au temps passé à retrouver et comprendre des données mal gérées ou mal documentées.
Le but de cet article est double : (1) expliciter le concept de l'archivage pérenne et en décrire les enjeux, et (2) présenter l'état de l'art sur ce sujet à l'Inra ainsi que les actions envisagées pour proposer des services et des outils aux agents.

[1] Mission Archives, DICSDAR Direction de l'Immobilier et de la coordination des services déconcentrés d'appui à la recherche, Inra, Paris, France, lina.sbeih@inra.fr
[2] UAR DSI-UA Direction du Système d'Information-Unité d'Appui, Inra, Paris, France, fanny.dedet@inra.fr
[3] UMR AGIR Département Environnement Agronomie, Inra, Toulouse/UAR Ingenum, Inra, Toulouse, France, patrick.moreau@inra.fr
[4] UAR DIST Délégation à l'IST, Inra, Versailles, France, esther.dzale-yeumo@inra.fr

Terminologie et contexte

Dans cette partie, le concept d'archivage pérenne sera précisé dans le but notamment de le distinguer des notions d'archive ouverte, de stockage et de sauvegarde.

Les données de la recherche sont aussi des archives

Les archivistes, au sein de la section des archivistes des universités, rectorats, organismes de recherche et mouvements étudiants (AURORE) de l'Association des archivistes français, ont souhaité proposer une définition du périmètre dans lequel ils interviennent.

« Les données de la recherche sont des informations, spécimens et matériaux produits, recueillis et documentés. Elles sont collectées ou exploitées à des fins de recherche et de preuves par les chercheurs et leurs équipes. À ce titre, elles constituent une partie des archives de la recherche. ».

Cette définition s'accompagne des précisions suivantes :

1. Les archives de la recherche englobent l'ensemble des documents et données produits ou reçus dans le cadre du processus de recherche, c'est-à-dire l'activité de recherche au sein des laboratoires ainsi que les fonctions relevant de l'appui à la recherche.
2. Les données de la recherche sont en grande partie électroniques mais peuvent exister également sur d'autres supports.
3. Les données de la recherche sont soit collectées soit exploitées dans le cadre du processus de recherche.[5]

Archives et archive ouverte

Les notions d'archives et d'archive ouverte sont fréquemment confondues.

Les archives sont « l'ensemble des documents, y compris les données, quels que soient, 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 »[6]. Elles sont conservées pour la bonne gestion des dossiers et des projets, leur valeur de preuve et la sauvegarde de la mémoire. A l'issue des durées de conservation, certaines archives sont à conserver et d'autres peuvent être détruites[7]. La communication des archives est régie par la loi. En effet, les archives publiques, également considérées comme des documents administratifs[8], sont communicables de plein droit, sous réserve de l'application de délais de communicabilité[9]. Les archives peuvent donc être ouvertes et sont consultables sauf si un délai spécifique est à prendre en compte. Un délai par exemple de 25 ans vise à protéger le secret en matière industrielle et commerciale et un autre de 50 ans protège la vie privée.

Les archives peuvent être publiques ou privées selon qu'elles sont produites par des institutions publiques ou des personnes privées. Les archives publiques[10] sont : imprescriptibles et inaliénables, soumises à des règles de gestion et au contrôle scientifique et technique de l'État sous l'égide du Service interministériel des archives de France[11]. Les archives d'institutions de recherche telles que l'Inra ou l'Inserm sont donc publiques.

---

[5] « Données de la recherche », article Wikipédia, https://fr.wikipedia.org/wiki/Données_de_la_recherche
[6] Code du patrimoine, livre II Archives, article L211-1.
[7] Code du patrimoine, livre II Archives, article L212-2.
[8] Code des relations entre le public et l'administration, livre III.
[9] Code du patrimoine, livre II Archives, article L 213-1 et 2.
[10] Code du patrimoine, livre II Archives, articles L212-1 à L212-5.
[11] Portail national des archives, https://francearchives.fr/

Selon une définition communément admise, le terme archive ouverte désigne un réservoir (ou entrepôt) où sont déposées des publications scientifiques (et parfois des données) dont l'accès se veut ouvert c'est-à-dire sans barrière. Cette ouverture s'appuie sur l'utilisation du protocole OAI-PMH[12] qui est un standard d'interopérabilité facilitant l'accessibilité de contenus provenant de plusieurs entrepôts maintenus par différents fournisseurs de données[13]. Outre le fait que l'archive ouverte a avant tout un caractère ouvert et concerne la production scientifique, elle renvoie souvent à un simple entrepôt Open Access (voir Directory of Open Access Repositories[14]) ou un entrepôt compatible avec le protocole OAI-PMH (voir OAI-PMH Registered Data Providers[15]).

L'archivage numérique ne se réduit pas au stockage des données

L'amalgame entre les notions de stockage, de sauvegarde, de pérennisation et d'archivage est courant. Il s'agit pourtant de termes distincts.

Le stockage est l'action d'enregistrer des informations sur un support pour les rendre accessibles.

La sauvegarde permet de dupliquer des informations pour prévenir toute disparition accidentelle.

La pérennisation (ou préservation) permet de faire face à la perte d'informations d'identification ainsi qu'à l'obsolescence des supports et des logiciels. Elle consiste en effet à identifier et à conserver des documents et des données pour les rendre accessibles sur le moyen (10 ans et plus) et le long terme (50 ans et plus).

L'archivage, qu'il concerne des supports électroniques ou papier, est une démarche qui vise à définir et mettre en œuvre des méthodes et des outils pour gérer et conserver des documents et des données. Le but est de rendre ceux-ci disponibles et accessibles pour répondre aux besoins métiers et scientifiques ainsi qu'aux obligations légales et sauvegarder les documents et données ayant un intérêt patrimonial. L'archivage détermine également les responsabilités de différents acteurs.

Afin d'être en mesure de conserver les documents et les données, ceux-ci doivent être décrits grâce à des métadonnées, enregistrés dans des formats pérennes et sélectionnés pour leur intérêt juridique, scientifique ou historique.

Bien souvent, les données de la recherche déposées dans des bases de données ou sur des serveurs sont stockées et sauvegardées mais elles ne sont pas archivées.

Contexte et enjeux de l'archivage

Les enjeux de l'archivage sont à la fois scientifiques, économiques, techniques, patrimoniaux et juridiques.
D'un point de vue scientifique, plusieurs éléments de contexte rendent la question de l'archivage prégnante. Par exemple, l'explosion du volume de données produites par la recherche induit la nécessité de sélectionner les données à archiver sur le long terme. Cette sélection doit tenir compte notamment de la valeur stratégique, économique et patrimoniale des données. Par ailleurs, dans un contexte de science ouverte et de défiance envers la science, les données de la recherche représentent un enjeu majeur pour la reproductibilité et l'intégrité de la science. A ce titre, leur sécurisation et leur conservation sur le long terme s'avèrent importantes.
De nombreux textes de loi régissent la gestion des données dont notamment le code du patrimoine, le code des relations entre le public et l'administration, la loi « Informatique et libertés », le règlement relatif à la protection

[12] http://www.openarchives.org/pmh/
[13] https://guides-formadoct.u-bretagneloire.fr/c.php?g=491583&p=3362295
[14] http://v2.sherpa.ac.uk/opendoar/
[15] https://www.openarchives.org/Register/BrowseSites?viewRecord

des données à caractère personnel, la loi relative à la gratuité et aux modalités de la réutilisation des informations du secteur public, dite loi Valter, la loi pour une République numérique appelée loi Lemaire, etc.

Au niveau institutionnel, plusieurs acteurs sont concernés par le sujet de l'archivage et peuvent intervenir à différentes étapes du cycle de vie des données. En amont on trouve bien sûr les producteurs de données ainsi que ceux qui les traitent et les analysent. Ce sont eux qui sont le mieux à même de documenter les données (métadonnées, autres documents de contexte) qui serviront à les rendre intelligibles et réutilisables dans le futur. Les archivistes apportent leur expertise métier. Ils ont, comme les professionnels de l'IST un rôle d'accompagnement des utilisateurs dans la mise en place des bonnes pratiques. Ils contribuent également au développement et à la mise en œuvre des offres de services en partenariat avec les informaticiens de l'Institut.
D'autres acteurs pourront également intervenir de par leur expertise sur certains sujets comme la DSI (Direction des systèmes d'information) le Délégué informatique et libertés (DIL) de l'Institut, le Responsable de la sécurité des systèmes d'informations (RSSI), le Fonctionnaire sécurité défense (FSD) ou la Direction des Affaires Juridiques (DAJ) pour toutes les questions liées à la sécurité et à l'utilisation d'un point de vue juridique des données scientifiques produites au sein de l'Institut.

Depuis 2015, la Mission Archives a mené des actions de sensibilisation et un site web[16]publié en Intranet a été créé afin de transmettre au plus grand nombre des informations sur les archives. Des outils ont également été produits dont les référentiels de gestion et de conservation des archives qui permettent à chacun de savoir que conserver et pendant combien de temps.

Dans le cadre du chantier stockage-archivage du projet Data Inra, les différents acteurs concernés dont la Mission Archives et la future direction pour la science ouverte (DipSO) collaboreront afin de proposer une offre de service pour l'Institut.

Vers une offre de service autour de l'archivage des données la recherche à l'Inra

Face aux enjeux d'ordre patrimonial, économique et scientifique évoqués plus haut, l'Inra souhaite mettre à disposition de ses agents des solutions pour répondre aux besoins d'archivage des données de la recherche. La quantité de données concernées, sans cesse croissante, amène à considérer qu'il n'est plus envisageable de conserver un accès immédiat à toutes les données. Pour certaines l'accès devient de moins en moins fréquent mais reste toujours essentiel. La façon de stocker les données peut être étudiée par rapport à la nature de l'activité concernée selon que le traitement justifie ou non de certaines caractéristiques de ce stockage. Par contre l'archivage doit être pris en compte plus globalement et envisagé tout comme le stockage dès le début du processus. La finalité même du processus d'archivage, qui est de conserver les données de la recherche sur des périodes longues, est à considérer comme une étape complémentaire, au sens fonctionnel et non forcément temporel, du stockage et selon un dispositif central à l'Institut. La mise en œuvre de ce dispositif d'archivage devra s'accompagner au plus près des quatre principes du FAIR[17].

Etat des lieux des besoins et des usages

Un groupe de travail[18] constitué début 2019 par le comité de pilotage du projet Data Inra a pris en charge le cadrage d'un chantier sur la question du stockage et de l'archivage pérenne. Sur ces deux problématiques étroitement liées, sans toutefois avoir les mêmes composantes, il a été décidé de démarrer ce chantier par un

[16] Site Intranet des Archives de l'INRA : https://intranet.inra.fr/archives-inra

[17] Principes FAIR : Findable, Accesible, Interopable, Re-usable

[18] Ce groupe de travail se compose des personnes suivantes : Eric Cahuzac - Sébastien Cat - Fanny Dedet - Esther Dzale-Yeumo - Ollivier Levy - Mikael Loaec - Patrick Moreau - Lina Sbeih

état des lieux prenant en compte le plus largement possible les usages et les besoins au sein de la communauté scientifique. Pour cela, le groupe de travail s'appuiera sur l'étude des volets data des dossiers de labellisation des Infrastructures Scientifiques Collectives (ISC) et sur un questionnaire qui sera adressé aux équipes dans les laboratoires par le canal des Commissions Locales des Systèmes d'Information (CLSI).
Les analyses des réponses donneront des indications sur le positionnement actuel d'un grand nombre de personnes et/ou de projets concernées par la manipulation (production, traitement, stockage) des données scientifiques à l'Inra. Elles pourront donner des informations sur les attentes actuelles et permettront d'établir des liens entre les initiatives qui auraient pu ici ou là se mettre en place. Le groupe de travail émettra des recommandations à la fois pour les scientifiques (où stocker vos données et quand) et pour l'établissement (quels outils et services communs mettre en place).

Data Inra

Data Inra (https://data.inra.fr) est un entrepôt de données ouvert dont la principale mission est de rendre les données plus faciles à trouver, accessibles, interopérables et réutilisables. A l'instar de la plupart des entrepôts de ce type, Data Inra a pour mission de (1) garantir l'accès des données sur une durée moyenne de dix ans à des fins de réutilisation, de preuve ou de reproductibilité de la science, et (2) préserver sur le long terme les données qui doivent l'être selon les critères fixés par la gouvernance des données de l'établissement (incluant un volet lié à l'archivage).
Les données de Data Inra qui ont vocation à être archivées sur le long terme seront confiées à un opérateur externe. Plusieurs pistes sont actuellement explorées pour l'archivage pérenne des données de Data Inra. Dans le cadre d'un projet européen (EOSC Pillar 5b) notamment, nous développerons un connecteur entre Data Inra et B2Safe[1] en collaboration avec le CINES.
Deux types de critères seront pris en compte pour sélectionner les données qui seront archivées sur le long terme : (1) **l'opportunité d'archiver** et (2) la **capacité d'archiver**. Concernant l'opportunité d'archiver, un chantier en cours sur la gouvernance des données qui inclut un volet archivage permettra d'identifier les critères de choix des jeux de données qui feront l'objet d'un archivage au-delà de dix ans. Le deuxième critère repose sur la présence ou non de métadonnées suffisantes pour préserver l'intelligibilité des données, et sur la nature durable des formats dans lesquels elles sont (ou la possibilité de les convertir dans des formats durables). Plusieurs niveaux de service sont envisagés in fine par Data Inra, approche inspirée de politiques d'archivage d'entrepôts similaires[1].
Data Inra met en œuvre les bonnes pratiques de gestion des données favorables à l'archivage pérenne. Parmi ces pratiques, nous pouvons citer l'utilisation de standards de métadonnées reconnus (Dublin Core, DDI, DataCite, etc.), l'extraction automatique des métadonnées des fichiers tabulaires pour permettre la découverte et la réutilisation des données, l'attribution automatique d'un identifiant pérenne DOI. Par ailleurs, depuis sa version 4.11, l'outil Dataverse sur lequel s'appuie Data Inra permet de générer des paquets aux formats OAI-ORE et BagIT qui sont des standards pour l'archivage pérenne. Cette version sera implémentée dans Data Inra au cours des prochains mois.

Mise en place des bonnes pratiques de gestion des données

Afin que les données puissent être archivées, un travail de préparation est nécessaire. Dans le but de permettre à chaque projet ou unité de recherche de prendre en compte les modalités d'archivage dès la production des données et durant tout leur cycle de vie, la mise en place de bonnes pratiques est essentielle.

Il est actuellement possible de renseigner des plans de gestion des données pour traiter l'ensemble des questions liées aux données dans le cadre d'un projet dont le devenir de celles-ci. L'Inra[19], a conçu des guides pour accompagner les scientifiques dans cette tâche. Ces guident sont disponibles dans l'outil DMP Opidor[20] où l'on retrouve également des guides de la Commission Européenne, de l'ANR et d'autres établissements publics de recherche tels que Paris Diderot, Descartes et Sorbonne Paris Cité[21].

Quelques aspects méritent plus particulièrement notre attention pour gérer au mieux les données et envisager un archivage de qualité. Lors du lancement d'un projet de recherche, une analyse des processus, des activités et des actions qui mènent à la production des données peut être effectuée. Cette étude permet de lister les méthodes de production des données, le type et la nature de celles-ci, les formats de production ou de traitement ainsi que les outils utilisés. Le lien entre les données et les documents produits comme les documents projet ou les cahiers de laboratoire est à maintenir. Ces éléments sont cruciaux pour contextualiser et décrire les données. En effet, sans ces informations, les données ne pourront être intelligibles sur le long terme.

La conservation des données est une étape parmi d'autres du cycle de vie des données. Les données sont produites ou collectées, elles sont traitées et analysées, elles sont stockées, elles peuvent être partagées ou réutilisées, elles sont conservées ou détruites. Pour cette raison, une réflexion au sujet de la pérennisation des données est à mener dès le lancement d'un projet de recherche.

L'identification des données grâce à des métadonnées est primordiale. Il existe des standards de métadonnées généraux comme Dublin Core[22], certains sont plus spécifiques à certains types de données, d'autres concernent l'archivage. Le format des données est également un préalable pour la pérennisation des données. Celui-ci doit être ouvert et si possible normalisé ou standardisé ou spécifié. La durée de conservation des données ainsi que leur sort final sont à définir. En effet, il est nécessaire de conserver les données pendant une durée pour la bonne gestion du projet et la valeur de preuve que constituent les données. A l'issue de cette durée, les données peuvent être pérennisées si elles ont une valeur de preuve, un intérêt scientifique ou historique et patrimonial. Cependant, si elles ne présentent pas d'intérêt, elles peuvent être détruites. Toutes les données ne sont pas à conserver sur le long terme. Une réflexion concernant le tri et la sélection des données a donc son importance. Afin de déterminer si les données sont à pérenniser, il est par exemple possible de se demander si elles sont reproductibles, si elles pourront faire l'objet de traitements complémentaires, être exploitées différemment ultérieurement ou être réutilisées. Une étude sur l'évaluation des risques et les conséquences de la perte d'exploitation des données peut être engagée. Une analyse des coûts de production comparés aux coûts de sauvegarde et d'archivage peut également être menée notamment pour les volumétries importantes. De plus, les données sont à organiser physiquement en les classant selon un plan et en instaurant des règles de nommage. Un lieu de stockage est également à définir. Celui-ci peut-être un espace intermédiaire pour un temps avant une pérennisation grâce au recours à une plateforme d'archivage numérique.

Offres existantes en matière d'archivage numérique

L'état des lieux des besoins et usages permettra d'avoir une vision plus précise des pratiques et outils utilisés au sein de l'Inra en matière de gestion et d'archivage des données de la recherche. Aujourd'hui il n'existe pas d'offre institutionnelle d'archivage numérique des données et les initiatives locales, au sein des unités ou des infrastructures de recherche sont peu nombreuses.

---

[19] https://www.inra.fr/datapartage/Gerer/Rediger-un-plan-de-gestion
[20] https://dmp.opidor.fr/
[21] *Réaliser un plan de gestion de données « FAIR » : guide de rédaction*, A. Cartier, R. Delemontez, M. Moysan, N. Reymonet, 2018, https://archivesic.ccsd.cnrs.fr/sic_01690547/document
[22] http://dublincore.org/

Le processus complet de l'archivage numérique des données sera probablement trop lourd pour être traité dans son intégralité en interne. L'Institut aura sans doute recours à des offres externes.

Des plateformes d'archivage numérique sont proposées par des sociétés d'archivage et par des institutions publiques. Le programme VITAM[23] est un programme interministériel d'archivage numérique piloté par le Comité Interministériel aux Archives de France (CIAF) et la Direction Interministérielle du Numérique et du Système d'Information et de Communication de l'État (DINSIC). Il est porté conjointement par les Ministères des Armées, de la Culture et de l'Europe et des Affaires Etrangères. Ce projet a entre autre pour objectif la réalisation d'une solution logicielle libre d'archivage numérique permettant la prise en charge, la conservation et la consultation sécurisée de très gros volumes d'archives numériques. De son côté le CINES[24] est depuis 2004 l'opérateur pour l'archivage des données et documents numériques produits par la communauté Enseignement Supérieur et Recherche française (ESR). Il propose des solutions d'archivage numérique sur le moyen et long terme, mutualisées et personnalisables. A l'échelle européenne, la solution B2SAFE proposée par EUDAT[25] permet également d'archiver de gros volumes de données sur le long terme.

Le groupe de travail analysera et étudiera les solutions existantes. Il préconisera l'offre la plus adaptée aux besoins de l'Inra, en s'appuyant éventuellement sur plusieurs services. Cela permettra de répondre plus finement aux besoins de l'Institut et de fixer un cadre national en matière d'archivage des données de la recherche mais aussi des données au sens large produites par l'Inra (données administratives, données documentaires, etc.).

Conclusion

Le but de cet article est de déterminer un cadre commun sur lequel construire une réflexion sur l'archivage des données de la recherche à l'Inra. Il est en effet essentiel que les termes et les concepts employés soient définis de manière claire et précise afin qu'ils aient la même signification pour toutes les parties prenantes du projet. Ces objectifs, ont d'autant plus d'importance du fait que l'Inra produit des données dans des domaines variés et qu'un nouvel institut naîtra prochainement du regroupement avec l'Irstea. Une prise de conscience de la nécessité d'avoir une approche globale de la gestion des données sur tout leur cycle de vie et de se préoccuper du devenir de celles-ci dès le début du processus de production, est primordiale pour la réussite de ce chantier. Aujourd'hui, au sein de l'Inra, les connaissances à ce sujet sont hétérogènes. Il est donc nécessaire d'accompagner chacun dans sa réflexion et d'apporter les éléments utiles pour mettre en place un dispositif de conservation des données de l'Institut. Des instituts de recherche, comme l'Inserm, se sont d'ores et déjà lancés dans de tels travaux et l'Inra s'apprête à s'engager dans cette voie en s'appuyant sur les retours d'expériences de ses partenaires.

[23] http://www.programmevitam.fr/
[24] https://www.cines.fr/archivage/
[25] https://www.eudat.eu/b2safe

Vers une prise en charge de l'archivage des données de la recherche, l'exemple de l'Inserm

Hélène Chambefort[1]

Le service des archives de l'Institut national de la santé et de la recherche médicale (Inserm) a été créé au tout début des années 1990, avec d'emblée une mission d'archivage des documents administratifs et scientifiques. Les premières collectes d'archives de laboratoires sont effectives dès 1993 et ne cessent de croître depuis, qu'ils s'agissent d'archives papier comme d'archives électroniques. Une réflexion autour de la mise en place de l'archivage électronique est menée à la fin des années 2010 et aboutit à la mise en place d'un logiciel hybride de gestion des archives, couplée au déploiement d'un système d'archivage électronique, en cours actuellement.

La réflexion sur l'archivage électronique s'est construite autour d'un partenariat entre le service des Archives, le département du système d'information (DSI) Inserm, les services producteurs et le Centre informatique national de l'enseignement supérieur (CINES).

Le premier objectif, après la signature d'une convention d'archivage sur le long terme avec le CINES en 2013, a été d'effectuer un versement parlant en termes de données de recherche : le choix s'est porté sur les images fixes et animées issues de la photothèque de l'Inserm « Inserm images », dont la production est pour près de 50 pourcents issue d'imagerie scientifique. La particularité est qu'il s'agit d'une base de données vivante, que l'on ne souhaitait pas archiver entièrement tous les ans, afin d'avoir une vision la plus exhaustive possible des images représentées dans cette base et par la même de ce que va être l'imagerie scientifique sur un temps défini. Par exemple les images représentant le microbiote intestinal étaient presque inexistantes en 2014 lors du premier versement au CINES et sont depuis bien présentes dans la base de données.

Le travail initial a été de définir les métadonnées pour chaque image, lesquelles ont été choisies en Dublincore (métadonnées propres à l'édition) en ce qu'elles collaient presque parfaitement aux propriétés de chacune des images. Chaque paquet (données, métadonnées …) est ensuite encodé et transféré via ftp vers le CINES, grâce à une moulinette permettant l'automatisation du transfert. Cette moulinette a été développée par Algoba, l'éditeur du logiciel DAM (digital asset management) utilisé pour la base « Inserm images ».

La première année toute la base a été versée, les images nouvelles sont depuis taguées et seules ces dernières, où les données modifiées, sont envoyées tous les ans pour archivage. Le versement de la photothèque a été très important puisqu'il a permis de montrer que l'archivage électronique des données de la recherche était possible et qu'il pouvait être assez simple bien que chronophage la première année. Mais il nécessitait un véritable échange de pratiques entre plusieurs structures.

La deuxième expérience relevait de données à très forte valeur légale, les cahiers d'observation issus de la recherche clinique à l'ANRS, agence publique française de recherches sur le Sida et les hépatites virales. Ces données devaient être conservées de manière électronique puisque ces cahiers contenaient des données nativement numériques, qui faisaient donc foi juridiquement en tant que document original.

Ce travail, qui a impliqué le choix des métadonnées, le transfert des données initiales, qui bien qu'en PDF a posé malgré tout pas mal de souci de reconnaissance et d'acceptation de format, l'encodage des paquets de données et le calcul des empreintes, a été très long et a occupé presque la totalité des six mois du stage de master, mais là encore il a démontré que l'archivage des données de recherche était possible. Cela a permis

également au service des archives d'avoir déjà un choix construit autour des métadonnées scientifiques. En revanche, à la différence d'« Inserm images » et de son éditeur Algoba, le travail d'automatisation n'a pas été possible via l'éditeur du logiciel de gestion des essais cliniques parce que trop long et surtout beaucoup trop coûteux à faire développer.

La réponse au déploiement d'un archivage des données de la recherche, vient en interne de l'acquisition d'un logiciel d'archivage hybride en 2018, lequel va permettre de développer et surtout d'automatiser à terme cet archivage en poursuivant le travail autour des essais cliniques et en développant l'archivage des cahiers de laboratoires électroniques. Ce projet déployé à grande échelle pour la totalité ou presque des laboratoires de l'Inserm est porté par le département des systèmes d'information. C'est avec lui que le service des archives réfléchit à la construction de l'archivage électronique dès 2015.

L'archivage électronique des données de la recherche ne peut se faire qu'en avançant par constructions successives, avec des partenariats solides en interne et en externe et si l'on veut le développer à plus grande échelle il faut investir dans des outils performants afin de gagner du temps dans toutes les étapes du travail, de la prise en charge au transfert final. Pour cela la mise à disposition d'outils en direction de la communauté scientifique, tels que les règles de nommage des fichiers électroniques, les formats pérennes de données... est également un plus pour mener à bien un archivage qui réponde avant tout à une conservation sinon historique, du moins légale et probante sur de très longues durées.

---

[1] Service des archives, DISC, Inserm, Paris, France, helene.chambefort@inserm.fr