Les problématiques liées à l'archivage des données de la recherche

openscience.pasteur.fr/2024/01/18/les-problematiques-liees-a-larchivage-des-donnees-de-la-recherche/

CeRIS - Institut Pasteur 18 janvier 2024

Selon le CINES, l'archivage numérique pérenne des documents électroniques consiste à **conserver le document et l'information qu'il contient**, dans son aspect physique comme dans son aspect intellectuel, sur le très long terme, de manière à ce qu'il soit en permanence accessible et compréhensible. Cela consiste ainsi à lutter contre plusieurs menaces : l'obsolescence matérielle, l'obsolescence logicielle, l'obsolescence du format de fichier et la perte de la signification du contenu.

A cela s'ajoute la problématique de l'**augmentation exponentielle du volume de données scientifiques**. Dès lors, la question de l'archivage numérique pérenne de ces données devient préoccupante pour les organismes de recherche. Que faire de ces données qui prennent énormément de place sur les espaces de stockage ? Doit-on les conserver, les supprimer ? Cette question n'a pas de réponse simple et elle en suscite de nouvelles. Nous vous proposons donc ici un état des lieux non exhaustif des **problématiques liées à l'archivage des données de la recherche**.

Il est clair qu'il n'est pas possible de conserver l'ensemble des données de recherche sur le long terme. Mais **comment sélectionner les données à préserver** ? Qui a l'expertise suffisante pour savoir si des données ont un intérêt scientifique/historique sur le long terme ? Une option pourrait consister à conserver uniquement les données liées aux publications ainsi que celles soumises à une obligation de conservation. Mais dans cette hypothèse, ne perd-on pas une grande partie du patrimoine scientifique ?

Arrive ensuite la question du pourquoi : **pourquoi souhaite-t-on préserver les données** ? Pour des questions juridiques (obligation de conservation, en cas de contentieux…), pour des questions d'intégrité scientifique et de reproductibilité, pour les réutiliser, pour un intérêt patrimonial et historique (témoignage de l'activité scientifique d'un organisme à un moment donné) ? La réponse à cette question aura un impact sur la façon dont les données seront archivées (format, niveau de sécurité, niveau d'accessibilité…). Et cette question du pourquoi en amène une autre : **combien de temps conserver les données de recherche** ?

Pour préserver la connaissance du contenu des fichiers, l'un des facteurs clés est la description des données archivées ainsi que la documentation de leur contexte de création. Mais **qui se charge de cette description et documentation** ? Ces activités sont chronophages et devraient être réalisées par les producteurs des données, c'est-à-dire les scientifiques. Hors, cette activité n'est actuellement pas valorisée et les chercheurs n'ont pas

beaucoup de temps à y consacrer. Par ailleurs, il faudrait **conserver le lien entre les données de recherche et les éléments contextuels** (cahier de laboratoire électronique, plan de gestion des données, documentation sur les projets, les équipes…) ? Mais de quelle manière ?

La question du moment de la collecte se pose également : **quand doit-on collecter les données à préserver** ? Les services d'archives sont généralement contactés au moment de la fermeture d'un laboratoire ou du départ en retraite d'un chercheur mais c'est souvent trop tard : un archiviste ne peut pas, à lui seul, réorganiser un « vrac numérique » de plusieurs années, ni le documenter sans l'expertise des producteurs de données. Mais alors, quel serait le moment le plus propice pour archiver les données ? De façon régulière, à la fin d'un projet de recherche… ?

Une question technique se pose enfin : **quel outil utiliser pour archiver ces données** ? Peut-on faire confiance aux entrepôts de données pour en assurer la préservation ? La difficulté repose sur le fait que ces entrepôts ont pour objectif premier de faciliter la diffusion et la découverte des données scientifiques, la plupart ne sont pas conçus pour en assurer la pérennité. Il arrive également que certains entrepôts disparaissent sans nécessairement adopter de stratégie pour éviter la perte de données. Par ailleurs, la multiplicité des entrepôts de données entraine un éparpillement des données et souvent la perte des liens entre elles et avec les éléments contextuels. Dès lors, ne devrait-on pas utiliser un même outil pour archiver les données de recherche et les documents contextuels ? Ne devrait-on pas archiver l'ensemble de la production d'un laboratoire au même endroit ?

Toutes ces problématiques sont complexes et n'ont pas de réponses tranchées pour le moment. Plusieurs bonnes pratiques restent cependant essentielles pour faciliter la préservation des données :

- Adopter des **bonnes pratiques d'organisation, de nommage et de description des fichiers** de façon à faciliter le repérage et la compréhension des données (on vous explique comment faire ici) ;
- Privilégier les **formats de fichiers ouverts, standardisés ou largement utilisés** dont le risque d'obsolescence est plus faible.
- Mettre en place un plan de gestion des données (PGD), par exemple un PGD d'entité pour réfléchir de façon collaborative à la gestion et la conservation des données, et se mettre d'accord sur les bonnes pratiques à adopter.