Glossaire

Les définitions et tentatives d'explications partagées ici sont issues pour la plupart de Wikipedia. Ce sont des interprétations et parfois des simplifications qui ne font pas autorité et sont données à titre indicatif pour éclaircir des notions qui pourraient être abordées lors de la journée.

Données de la recherche

Les données de la recherche sont l'ensemble des informations produites et utilisées par la recherche scientifique. Elles peuvent recouvrir différents types de données selon le point de vue adopté, notamment par les différents métiers qui composent la recherche scientifique ainsi que l'information scientifique et technique.

L'Association des archivistes français donne la définition suivante : « Les données de la recherche concernent, en plus des métiers de la recherche, les métiers qui viennent en appui à celle-ci, tels que la documentation, les archives et l'informatique. Chacun de ces métiers a un rôle à jouer dans le cycle de vie des données. L'archiviste apporte son expertise pour leur gestion, leur conservation voire leur communication. 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 ». Les archives de la recherche englobent donc l'ensemble des documents et données produits ou reçus dans le cadre du processus de recherche. C'est-à-dire à la fois l'activité de recherche au sein des laboratoires et par les chercheurs et l'administration de la recherche au sein des organismes ainsi que par les fonctions venant en appui à la recherche. Elles sont en grande partie électroniques mais peuvent exister également sur d'autres supports. Elles sont soit collectées soit exploitées dans le cadre du processus de recherche.

Selon l'OCDE : « Les données de la recherche sont définies comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider les résultats de la recherche. »

La reproductibilité

La reproductibilité d'une expérience scientifique est une des conditions qui permettent d'inclure les observations réalisées durant cette expérience dans le processus d'amélioration perpétuelle des connaissances scientifiques. Cette condition part du principe qu'on ne peut tirer de conclusions que d'un événement bien décrit, qui est apparu plusieurs fois, provoqué par des personnes différentes. Cette condition permet de s'affranchir d'effets aléatoires venant fausser les résultats ainsi que des erreurs de jugement ou des manipulations de la part des scientifiques.

Le critère de reproductibilité est une des conditions sur lesquelles le philosophe Karl Popper distingue le caractère scientifique d'une étude.

Pour toutes les sciences expérimentales, les probabilités fournissent un modèle mathématique expliquant la variabilité des résultats.

PGD

Un plan de gestion des données, Data management plan ou PGD est un document évolutif qui aide le(s) chercheur(s) ou le chargé de projet de la recherche à définir un plan pour gérer les données utilisées et générées dans le cadre de son activité ou de son projet de recherche. Initié au début du projet, ce plan est mis à jour de manière périodique pour s'assurer de son adéquation avec le déroulement de l'activité ou du projet.

Plus spécifiquement, le plan de gestion aborde les éléments suivants :

- la description des données collectées et/ou créées,
- les standards, formats et méthodologies appliqués sur le paquet de données,
- les questions d'ordre éthiques, de propriété intellectuelle et de restrictions (les éventuelles périodes d'embargo),
- les prévisions pour le partage et l'ouverture des données,
- et la stratégie de la préservation à long-terme (archivage).

Cette description se compose des éléments régissant le cycle de vie des données de la recherche à savoir : la création, le traitement, l'analyse, la conservation, l'accès et la réutilisation des données.

Le plan de gestion des données est une réflexion en amont qui permet de donner une ligne directrice à la donnée lors de sa création mais également à garantir sa bonne exploitation dans le futur par des politiques d'archivages et des conditions de réutilisation prédéfinies.

Intéropérabilité

L'interopérabilité est la capacité que possède un produit ou un système, dont les interfaces sont intégralement connues, à fonctionner avec d'autres produits ou systèmes existants ou futurs et ce sans restriction d'accès ou de mise en œuvre.

Il convient de distinguer « interopérabilité » et « compatibilité ». Pour être simple, on peut dire que la compatibilité est une notion verticale qui fait qu'un outil peut fonctionner dans un environnement donné en respectant toutes les caractéristiques et l'interopérabilité est une notion transversale qui permet à divers outils de pouvoir communiquer - quand on sait pourquoi, et comment, ils peuvent fonctionner ensemble.

FAIR data

Dans le contexte de l'accessibilité de l'Internet, du Big data des données de la recherche et des sciences ouvertes (Open science) et plus largement du partage et l'ouverture des données, la notion de FAIR data (ou Fair data) recouvre les manières de construire, stocker, présenter ou publier des données de manière à permettre que la donnée soit « trouvable, accessible, interopérable et réutilisable ».

Le mot Fair fait aussi référence au Fair use, fair trade, fair play, etc., il évoque un comportement proactif et altruiste du producteur de données, qui cherche à les rendre plus facilement trouvables et utilisables par tous, tout en facilitant en aval le sourçage (éventuellement automatique) par l'utilisateur des données.

Metadonnée (metadata)

Une métadonnée (mot composé du préfixe grec meta, indiquant l'auto-référence ; le mot signifie donc proprement « donnée de/à propos de donnée ») est une donnée servant à définir ou décrire une autre donnée quel que soit son support (papier ou électronique).

Un exemple type est d'associer à une donnée la date à laquelle elle a été produite ou enregistrée, ou à une photo les coordonnées GPS du lieu où elle a été prise.

Normes de métadonnées

Les normes de métadonnées sont des normes qui décrivent les données sur les données, employées pour la structuration des ressources informatiques en général (pas seulement les documents électroniques) et l'interopérabilité informatique.

Étant donné les multiples utilisations des métadonnées, à la fois dans les ressources informatiques et les systèmes, il est nécessaire d'employer des normes.

La plupart de ces normes ne sont disponibles qu'en anglais. Il n'existe que deux normes disponibles en français : celle sur le référentiel Dublin Core, et celle sur le patrimoine culturel immatériel (ISO 21127).

Les normes de métadonnées portant sur des domaines particuliers découlent de la définition de communautés d'intérêt (cf DoDAF).

Ontologie

En informatique et en science de l'information, une ontologie est l'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances. L'ontologie constitue en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que des relations entre ces concepts. Elle est employée pour raisonner à propos des objets du domaine concerné. Plus simplement, on peut aussi dire que l'« ontologie est aux données ce que la grammaire est au langage ».

Open Science

La science ouverte (open science ou open research pour les anglophones) est un mouvement visant à rendre la recherche scientifique et les données qu'elle produit accessibles à tous et dans tous les niveaux de la société.

Pour cela la science ouverte s'appuie fortement sur le recours à l'Internet ouvert, à l'open data, aux outils de travail collaboratif (dont Wikipédia et Wikiversité ou Wikispecies font partie), à la formation en ligne et au web social de manière à rendre la recherche scientifique et ses données accessibles à tous (amateurs et professionnels)

Archivage électronique

L'archivage de contenus électroniques est l'ensemble des actions, outils et méthodes mis en œuvre pour réunir, identifier, sélectionner, classer, détruire et conserver des contenus électroniques, sur un support sécurisé, dans le but de les exploiter et de les rendre accessibles dans le temps, que ce soit à titre de preuve (en cas d'obligations légales notamment ou de litiges)

ou à titre informatif. Le contenu archivé est considéré comme figé et ne peut donc être modifié. Ceci est notamment possible en garantissant l'authenticité via l'empreinte électronique, la signature électronique, la traçabilité des accès et bien d'autres moyens. La durée de l'archivage est fonction de la valeur du contenu et porte le plus souvent sur du moyen ou long terme. La conservation est l'ensemble des moyens mis en œuvre pour stocker, sécuriser, pérenniser, restituer, tracer, transférer voire détruire, les contenus électroniques archivés.

Entrepôt de données

Un Data Warehouse (Entrepôt de données) est une base de données relationnelle pensée et conçue pour les requêtes et les analyses de données. En Science, à la fois source et lieu de stockage d'information, les entrepôts de données jouent un rôle clé dans le mouvement de l'open science. Le choix de l'entrepôt est essentiel pour mettre à disposition des données FAIR (Facilement trouvables, Accessibles, Interopérables et Réutilisables) et pour répondre à l'obligation faite aux établissements publics d'ouvrir à tous leurs données.

Dataverse

C'est une application web permettant de préserver, partager, citer, rechercher et analyser des données de recherche.

Le répertoire principal, Harvard Database, héberge plusieurs plateformes dataverses. Chacune d'entre elles contient des jeux ou paquets de données décrits à l'aide de métadonnées et attachés aux fichiers (en incluant la documentation et le code).

Identifiant pérenne

Un identifiant pérenne est une chaîne de caractères alphanumériques formant référence stable à un document, une page internet, ou tout autre objet.

La notion d'identifiant pérenne répond à un problème d'archivistique. Un document doit être désigné sans ambiguïté de telle façon qu'un utilisateur puisse le retrouver, quelles que soient les réorganisations de l'espace ou des espaces où il se trouve conservé. L'International Standard Book Number (« numéro international normalisé de livre » ISBN) donne ainsi, dans une bibliothèque, accès à une édition précise d'un ouvrage imprimé.

Une ressource en ligne est accessible par une chaîne de caractères appelée Uniform Resource Locator (url) qui indique son emplacement dans un serveur informatique. L'organisation des archives amène à modifier l'url, par exemple en changeant la structure des dossiers, alors qu'il s'agit toujours du même document. On définit, pour désigner cette ressource indépendamment de son emplacement, un identifiant pérenne, qu'un intermédiaire pourra convertir en url pour donner accès à la ressource. L'identifiant comprend un préfixe, indiquant le système de résolution des identifiants, et une chaîne de caractères indiquant l'objet.

Certains systèmes d'identifiant pérenne comme Archival Resource Key promeuvent l'opacité des identifiants, d'autres acceptent des identifiants explicites.

Digital object identifier

Digital object identifier (DOI, littéralement « identifiant numérique d'objet ») est un mécanisme d'identification de ressources, qui peuvent être des ressources numériques, comme un film, un rapport, des articles scientifiques, mais également des personnes ou tout autre type d'objets. Le

but des DOI est de faciliter la gestion numérique sur le long terme de toute chose en associant des métadonnées à l'identifiant de la chose à gérer. Les métadonnées peuvent évoluer au cours du temps, mais l'identifiant reste invariant. C'est une alternative aux URI. Depuis 2012, le système d'identifiant numérique d'objet a été normalisé sous la forme de la norme ISO 26324.

Les DOI sont notamment utilisés dans les bases de données bibliographiques. Depuis février 2010, l'Institut de l'information scientifique et technique (INIST, du CNRS) est doté d'un statut « agence DOI », faisant partie du consortium DataCite.

Le DOI d'un document permet notamment une identification pérenne de celui-ci. Par exemple, il permet de retrouver l'emplacement d'un document en ligne si son URL a changé.

Les DOI permettent ainsi de faciliter l'utilisation des bases de données bibliographiques, des logiciels de gestion bibliographique, et de produire des citations plus fiables et plus pérennes.

Cycle de vie des données

Le cycle de vie des données de la recherche est l'ensemble des étapes de gestion, conservation, diffusion et réutilisation des données scientifiques liées aux activités de recherche. Ces étapes impliquent des activités particulières : élaborer le plan de gestion de données, décrire les métadonnées décrivant les données, choisir les entrepôts pour déposer les données, administrer les infrastructures de conservation des données, découvrir et explorer les données, réutiliser les données, définir le cadre législatif, juridique ou contractuel pour diffuser les données.

Stockage capacitif

C'est un stockage de données privilégiant une grande capacité à faible coût

Stockage performant

C'est un stockage de données privilégiant la performance (en vitesse et nombre d'accès simultanés)

Stockage (à plusieurs termes (court, long, pérenne))

Les solutions techniques et organisationnelles seront différentes selon la période de temps pendant laquelle on souhaite que la donnée soit disponible. Par exemple, pour un stockage pérenne sur des dures qui se chiffrent en décennies, on devra rechercher des solutions qui soient à la fois très fiables et très peu couteuses. Pour un stockage à court terme, l'impact du coût sera moins important

Préservation des données électroniques

Actions mise en place pour garantir, dans le temps, l'accès aux données pour leur réutilisation

Web sémantique

Le Web sémantique, ou toile sémantique, est une extension du Web standardisée par le World Wide Web Consortium (W3C). Ces standards encouragent l'utilisation de formats de données et de protocoles d'échange normés sur le Web, en s'appuyant sur le modèle Resource Description Framework (RDF).

Le web sémantique est par certains qualifié de web 3.0 .

Selon le W3C, « le Web sémantique fournit un modèle qui permet aux données d'être partagées et réutilisées entre plusieurs applications, entreprises et groupes d'utilisateurs ». L'expression a été inventée par Tim Berners-Lee (inventeur du Web et directeur du W3C), qui supervise le développement des technologies communes du Web sémantique. Il le définit comme « une toile de données qui peuvent être traitées directement et indirectement par des machines pour aider leurs utilisateurs à créer de nouvelles connaissances ». Pour y parvenir, le Web sémantique met en œuvre le Web des données qui consiste à lier et structurer l'information sur Internet pour accéder simplement à la connaissance qu'elle contient déjà.

RGPD

Le règlement général sur la protection des données est un règlement de l'Union européenne qui constitue le texte de référence en matière de protection des données à caractère personnel. Il renforce et unifie la protection des données pour les individus au sein de l'Union européenne. Les principaux objectifs du RGPD sont d'accroître à la fois la protection des personnes concernées par un traitement de leurs données à caractère personnel et la responsabilisation des acteurs de ce traitement. Ces principes pourront être appliqués grâce à l'augmentation du pouvoir des autorités de régulation