Les types de données

1. Définition

Les données de la recherche (*Research Data*) constituent la « matière première » à partir de laquelle les recherches scientifiques produisent et justifient leurs résultats. Tout résultat de recherche, pour être considéré comme scientifiquement fondé, doit s'appuyer sur l'analyse de données primaires ou secondaires, et ce quelle que soit la discipline scientifique considérée.

L'Organisation de Coopération et de Développement Économiques (OCDE) définit les données de recherche 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 » (OCDE, 2011).

Les données de recherche sont donc ce qui rend possible la production de connaissances scientifiques. Elles sont au fondement de l'administration de la preuve.

2. Types de données

L'INIST (l'Institut de l'Information Scientifique et Technique) du CNRS définit 5 types de données :

- Les données d'observation : données capturées en temps réel, habituellement uniques et donc impossibles à reproduire.
- Les données expérimentales : données obtenues à partir d'équipements de laboratoire, qui sont souvent reproductibles mais parfois coûteuses.
- Les données computationnelles ou de simulation : données générées par des modèles informatiques ou de simulation, souvent reproductibles si le modèle est correctement documenté.
- Les données dérivées ou compilées : données issues du traitement ou de la combinaison de données « brutes », elles sont souvent reproductibles mais coûteuses.
- Les données de référence : collection ou accumulation de petits jeux de données qui ont été revus par les pairs, annotés et mis à disposition.

Il importe de garder à l'esprit que les données ne sont jamais « données » mais plutôt « obtenues » (Latour, 1996, p. 188) grâce à des procédés impliquant des humaines et/ou des machines. En effet, si la température est une donnée pour le météorologue, les données que traite ce dernier sont des enregistrements factuels obtenus à partir d'une chaîne de transmission impliquant des capteurs, des transmetteurs et des récepteurs. De la même manière, si un texte ancien est une donnée pour l'historien, ce dernier a dû en découvrir l'existence par des recherches, obtenir des autorisations pour y avoir accès, le scanner, le

traduire, le reconstituer... Dans la même logique, le physicien comme le psychologue obtiennent des données à partir de leurs expérimentations, le sociologue les obtient en faisant passer un questionnaire ou en réalisant des entretiens, le géographe constitue son corpus en réalisant des photographies ou des plans qu'il va ensuite traiter pour obtenir des données cartographiques, l'archéologue réalise des fouilles puis date et classifie ses prélèvements, l'épidémiologue obtient ses données auprès des laboratoires d'analyse (résultat de tests par ex.) qui eux-mêmes les obtiennent auprès des structures médicales (échantillons prélevés) qui eux-mêmes les obtiennent des patients, etc.

Une donnée est donc toujours une information qui a été produite par une démarche méthodologique impliquant des agents humains et non-humains. Toute discipline scientifique gagne donc à réfléchir sur ses propres modes de production des données afin de ne pas confondre ces dernières avec le réel qu'elles cherchent à saisir.

3. Caractéristiques des données

Primaires ou secondaires : Lorsqu'un protocole de recherche produit ses propres données, on parle de données primaires. Mais toute recherche scientifique ne produit pas systématiquement son propre set de données avant de réaliser ses analyses. Les données de la recherche peuvent en effet être produites et fournies aux équipes de recherche par d'autres équipes ayant mis leurs données en partage sur des répertoires de données (*Data repositories*) ou par des organismes tiers chargés de constituer des bases de données (observatoires nationaux par ex.). On parle alors de données secondaires (ou de seconde main) lorsque les équipes de recherche exploitent et analysent des données qu'elles n'ont pas produites.

Formatées et regroupées : les données de la recherche doivent faire l'objet d'un traitement afin d'être lisibles, compréhensibles, contextualisées, associables entre elles. Une fois formatées et regroupées dans un même espace, elles forment entre elles un corpus ou set de données. C'est seulement une fois rassemblées qu'elles peuvent être analysées puisque l'administration de la preuve scientifique repose sur la recherche et l'analyse de répétitivités.

Sensibles : Certaines données personnelles peuvent être qualifiées de sensibles et nécessiter des précautions particulières afin que leur utilisation ne nuise pas aux individus (cf. « Données personnelles » dans le glossaire). Il s'agit notamment des données sur la santé.

Intègres : Obtenues dans le cadre de la mise en œuvre d'un protocole de recherche, les données sont en elles-mêmes un produit scientifique qui a à la fois une certaine valeur (scientifique, historique mais aussi commerciale) et une certaine confidentialité (informations sensibles, propriété intellectuelle). Les données doivent donc être conservées de manière

sécurisée et leur partage doit être l'objet d'une régulation afin que leur usage, strictement scientifique, ne soit pas détourné à des fins politiques ou commerciales.

FAIR : Dans le cadre des politiques en faveur de l'Open Science, les données doivent être FAIR, c'est-à-dire faciles à trouver (*Findable*), accessibles (*Accessible*), interopérables (*Interoperable*) et réutilisables (*Reusable*). Pour désigner l'ensemble des opérations de formatage, enregistrement et partage des données en conformité avec les politiques en faveur de l'Open Data, on parle donc aujourd'hui de « données FAIR » ou de « fairisation » des données.

Quantitatives ou qualitatives : Selon les disciplines, les données peuvent être quantitatives (données codées en quantité importante) ou qualitatives (données observationnelles, discours et textes devant faire l'objet d'une interprétation). Si ces deux types de données peuvent être utilisées de manière complémentaires ou interconnectées (*Grounded theory*), elles relèvent de deux méthodologies distinctes :

- Les données quantitatives relèvent d'une méthode dite « hypothético-déductive » : l'hypothèse de recherche est antérieure à la production des données et l'analyse de ces dernières a pour objectif de confirmer ou infirmer l'hypothèse de travail. Ce type de données relève donc prioritairement des sciences expérimentales.
- Les données qualitatives relèvent d'une méthode dite « empirico-inductive » : l'hypothèse s'élabore et s'affine durant la phase de production et d'interprétation des données. Ce type de données relève donc prioritairement des sciences humaines et sociales.

Références

Organisation de Coopération et de Développement Économiques (2021). – *Recommandation du Conseil concernant l'accès aux données de la recherche financée sur fonds publics.* https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0347

Fournier, T. (2014). Les données de la recherche : définition et enjeux. *Arabesques*, 73, 4-6. https://10.0.138.234/arabesques.985

Latour, B. (1996). *Petites leçons de sociologie des sciences*. La Découverte.

URD Data (2022). *Les données de la recherche*. https://data.ird.fr/gerer/quelles-donnees/#Les_types_de_donnees

GO-FAIR (2022). *FAIR principles*. https://www.go-fair.org/fair-principles/

Les types de données © 2021 by Groupe de travail Guidelines de la Communauté Open Science HES-SO is licensed under Attribution 4.0 International