La qualité des données

Synthèse webinaire du 5 juillet 2021
Ce qu'il faut retenir

Cette synthèse reprend les principaux éléments échangés lors du webinaire « Qualité des données » organisé le 5 juillet 2021 par le groupe de travail Atelier Données, un groupe inter-réseau mis en place en 2016 par la MITI (Mission pour les Initiatives Transverses et Interdisciplinaires) du CNRS.

Ce webinaire s'insère dans un cycle thématique qui illustre une série de thèmes traités dans le guide de bonnes pratiques sur la gestion des données de la recherche réalisée par ce même Atelier Données et paru en janvier 2020. Les thématiques abordées dans ce cycle de webinaires tout comme dans le guide s'attachent à montrer une application aux données de la recherche vue sous l'angle de pratiques observées dans différents métiers de la recherche.

Pour illustrer la thématique du jour dédiée à la qualité des données, deux principales questions ont été retenues

- Qu'est-ce qu'une donnée de qualité ?
- Quelle organisation faut-il mettre en place pour arriver à obtenir des données de qualité ?

Les concepts généraux de la thématique ont été illustrés par des retours d'expérience.

A l'occasion de cette demi-journée introduite et clôturée par Marie Claude Quidoz[1], nous avons accueilli tout d'abord Alain Rivet[2] et Henri Valeins[3] du réseau « Qualité en recherche » (QeR) qui ont eu la lourde tâche de définir le lien existant entre qualité et données. Christine Coatanoan[4] a ensuite présenté le processus de contrôle et de qualification des données en usage dans un système d'observation océanographique. Après la pause, trois interventions ont suscité notre intérêt. La première, présentée par Geoffrey Aldebert[5], a témoigné de la démarche d'open data mise en œuvre au sein d'Etalab depuis la fédération d'une communauté jusqu'à la production de données en passant par la conception d'un schéma de données permettant d'assurer la qualité. La seconde intervention présentée conjointement par Véronique Humbert[6] et Blandine Nouvel[7] a porté sur les enjeux de la qualité des référentiels et des métadonnées pour la communauté scientifique. A travers l'exemple des outils de Frantiq, les intervenantes ont montré l'intérêt des vocabulaires contrôlés et des alignements pour une meilleure « fairisation » et réplicabilité d'un modèle de données. La dernière intervention a été consacrée à un retour d'expérience en écologie végétale sur les étapes d'homogénéisation des données et a été présentée par Eric Garnier[8].

[1] Marie-Claude Quidoz, Ingénieure de recherche au CEFE, Co-animatrice du Groupe de travail Atelier Données
[2] Alain Rivet, Ingénieur de Recherche au Cermav, Réseau QeR
[3] Henri Valeins, Ingénieur de Recherche au RMSB, Réseau QeR
[4] Christine Coatanoan, Ingénieur Gestion des données au Sismer, Ifremer, Brest
[5] Geoffrey Aldebert, Data Engineer, Etalab, Département de la Direction Interministérielle Du Numérique (DEDUM)
[6] Véronique Humbert, Ingénieure de recherche, Archéologie des Sociétés Méditerranéennes, GDS Frantiq
[7] Blandine Nouvel, Ingénieure de recherche, Centre Camille Jullian, GDS Frantiq
[8] Eric Garnier, Chercheur au CEFE

Christine Hadrossek – DDDOR – Août 2021

Cette synthèse se structure autour de 3 points essentiels. Ils reflètent le cœur du sujet, la richesse des débats qui ont animé ce webinaire et attestent du rôle déterminant qu'occupe la notion de qualité dans la gestion des données de la recherche.

Données et qualités : des notions aux contours et dimensions multiples parfois difficiles à cerner

La norme ISO 9000 définit la qualité comme étant l'aptitude d'un ensemble de caractéristiques intrinsèques (traits permanents) à satisfaire des exigences (besoins formulés ou imposés).

Cette définition est claire et compréhensible, elle n'en reste pas moins large et complexe dès lors qu'il s'agit d'appliquer ou de définir des critères de qualité. De même que la notion de donnée est beaucoup plus diverse qu'il n'y paraît, il existe différentes visions possibles de la qualité.

Alain Rivet et Henri Valeins ont en particulier pointé la subjectivité induite par cette définition et les différentes réponses possibles que pourrait formuler un chercheur à la simple question « qu'est-ce qu'une donnée de qualité », en fonction de ses objectifs, son domaine scientifique ou de ses exigences disciplinaires.
Ils nomment quatre qualités intrinsèques associées à la donnée : la répétabilité, la fiabilité, la reproductibilité, la robustesse tandis que Geoffrey Aldebert mentionne 5 dimensions de la qualité axées cette fois sur le contenu de la donnée : la complétude, l'unicité, la consistance, la validité et la précision.

On constate une première difficulté qui résulte de la **variété des caractéristiques, facteurs, dimensions, attributs** que l'on peut associer à la notion de qualité et **l'absence d'une définition consensuelle** pour qualifier ce qu'est une donnée de qualité.

Alain Rivet et Henri Valeins ont rappelé cependant très justement **qu'une donnée n'a pas de valeur en dehors de son contexte et que la qualité d'une donnée ne peut être décorrélée de la qualité de ses métadonnées.**

La connaissance (information intellectualisée) s'acquière par l'interprétation de la donnée qui doit être signifiante (porteuse d'information) et c'est précisément grâce à l'apport des métadonnées que l'on chemine de la donnée vers l'information puis la connaissance.

On perçoit ici très distinctement l'utilité, **la fonction charnière des métadonnées dans le processus de construction de la connaissance** et l'intérêt de mettre en place une gouvernance des données ainsi qu'une démarche qualité, c'est-à-dire une démarche organisationnelle qui comprend une approche processus (production, gestion, contrôle, conservation, accessibilité et diffusion des données), une amélioration continue (qui s'appuie principalement sur le cycle PDCA : Plan – Do – Check – Act) et une traçabilité des actions (avec des indicateurs de traçabilité corrélés aux caractéristiques intrinsèques des données).

La norme ISO 9001 définit par ailleurs la connaissance comme un « Ensemble disponible d'informations constituant une conviction justifiée et ayant une forte certitude d'être vraies »

A noter ci-dessous quelques éléments relevés par Alain Rivet et Henri Valeins qui permettent d'évaluer le niveau de qualité d'un jeu de données (et disposer de données FAIR).

- Des éléments sur les données elles-mêmes et leur structure : format de fichier, structure du fichier…

- Des éléments attestant du potentiel de réutilisation et de croisement des données : Le respect de standards, référentiels et schémas déjà établis ; La présence de données et colonnes pivots pour lier les données à un référentiel (Le code SIRET ou SIREN par exemple).
- Des éléments qui accompagnent les données (documentation claire et rigoureuse, gestion des versions et des mises à jour éventuelles des données…)

La qualité d'une donnée n'est pas simple à appréhender. Se pose alors la question de savoir ce qu'est une donnée valide ? Comment valide-t-on une donnée et quels sont les risques d'introduire des biais (*Un point de vue situé peut-il définir un domaine de validité de la donnée ? Associer un standard ne comporte-t-il pas un risque de se conformer à un certain modèle de connaissance ?*).

Répondre à ces interrogations peut s'avérer complexe mais il est urgent pour pallier aux risques de « non qualité » des données de pouvoir disposer d'informations fiables, de données qui ont une forte certitude d'être vraies, dotées d'un maximum de robustesse et décrites le mieux possible. Le véritable risque étant d'ignorer ou sous-estimer l'importance des métadonnées dans le processus de gestion des données de la recherche.

Le contrôle qualité et la curation des données, un mode opératoire déterminant pour qualifier, mesurer, assurer la pertinence des données

Alain Rivet et Henri Valeins considèrent « qualité des données » et « qualité de la recherche » comme des notions très proches. La confiance en la qualité d'une recherche peut s'appuyer sur le fait de **vérifier que les différentes étapes d'une étude pourront être répétées en obtenant un résultat identique** (cela passe par exemple par la maîtrise d'un équipement et le fait qu'à chaque instant l'équipement réalise des opérations conformes à ce que l'on attend). **La qualité d'une donnée** (une donnée de confiance) **se mesure à travers ses caractéristiques intrinsèques** (Répétabilité, Fiabilité, Reproductibilité, Robustesse) qui sont validées, contrôlées et garantissent ainsi un premier niveau de qualité.

Christine Coatanoan illustre ce point de manière concrète en présentant un processus de contrôle et de qualification de données dans un système d'observation océanographique. Le contrôle se réalise au moyen de programmes informatiques, d'outils, de tests automatiques qui permettent de travailler sur de gros volumes de données : tests de positions (latitude/longitude), de vitesse entre deux profils, test de dérive de capteurs, tests de rang présentant des valeurs maximum et minimum etc. Un code qualité est attribué aux données [0,1,2,3…] pour permettre d'évaluer la cohérence des mesures et l'intervention d'un expert peut être requise dans un second temps pour vérifier les jeux de données au moyen d'un outil de contrôle visuel et des connaissances scientifiques de l'expert. Christine Coatanoan définit ici le contrôle qualité comme étant avant tout une opération qui mesure la qualité de la donnée et qui consiste à distinguer une mesure aberrante d'une mesure qui reflète un phénomène réel.

Dans un autre domaine d'étude, en Archéologie, le réseau Frantiq s'engage dans des actions de curation visant à maintenir la cohérence et la pertinence de ses outils (un catalogue de 600 000 notices alimenté par 40 bibliothèques et un thésaurus constitué de 60 000 concepts) en procédant à **des alignements de référentiels** et **a recours à une expertise scientifique** pour contrôler, valider le vocabulaire de son thésaurus Pactols.
La procédure est assez proche au laboratoire CEFE, en Ecologie végétale ou un contrôle qualité s'opère pour **vérifier l'exactitude syntaxique et sémantique des données**, garantir leur

homogénéité, leur **conformité au modèle**, s'assurer ainsi de la qualité de la base et de la possible réutilisation des données.

Au sein d'Etalab, le contrôle s'effectue au moyen de **l'outil Publier.Etalab.Studio** qui référence **des schémas** avec l'objectif de publier sur la plateforme data.gouv et faciliter la consolidation automatique de données issues d'un même schéma. La vérification se fait d'après le référentiel Etalab. Cet outil permet à un utilisateur de sélectionner un modèle puis saisir, charger, valider ou corriger des données si elles sont erronées (l'outil propose des boucles de correction à l'utilisateur).

Ces retours d'expériences variés montrent que **plusieurs approches sont possibles et complémentaires pour contrôler les différents aspects de la qualité**. Elles témoignent de l'importance, de l'utilité et des applications possibles d'une démarche qualité dans différentes communautés de recherche ne serait-ce que pour détecter les données non fiables et éviter le risque de « non qualité » évoqué par Alain Rivet et Henri Valeins (données inexactes, non conformes, non contrôlées, non sécurisées ou non fiables).
Elles montrent toutefois à l'instar des échanges qui ont alimenté le débat que ce type d'opérations ne coule pas de source et soulève même de nombreuses questions.

Ces démarches peuvent s'avérer **compliquées à implémenter** dans la mesure où elles impliquent de réaliser un certain nombre **d'opérations techniques de normalisation** (automatisées ou nécessitant une expertise), de faire des **choix d'outils** (utiliser un outil existant ? en créer un nouveau ?) et **de référentiels** (quelles est la bonne norme et pour quelle finalité ?, quel schéma adopter ?) de **définir des codes** qualité (lesquels ? associés ou non aux métadonnées ?), **un circuit de validation** (avec un processus de validation itératif ou modulable tout au long du projet ?), de **mobiliser une communauté d'expert** (combien ? à quelle fréquence ?) etc.

Elles suscitent également des **questions techniques** sur la **faisabilité des opérations** (Peut-on avoir recours au machine learning pour faciliter la correction des données ? **leur traçabilité** (Est-ce que les logiciels assurent la traçabilité des actions par les opérateurs ?) les **fonctionnalités** associées aux outils (Existe-t-il des systèmes d'auto-complétion en cas d'erreur de saisie, un suivi de correction pour les données incomplètes ?)

Et elles interrogent inévitablement sur les compétences et expertises nécessaires pour réaliser le travail : quelles sont les pratiques en la matière ? Qui peut, doit être sollicité ? pour quel type de traitement ? avec quel degré d'expertise ? Combien de temps consacre-t-on à la curation des données ? A-t-on les moyens humains et les ressources nécessaires (combien de personnes sont impliquées dans le contrôle qualité, la correction des données, la construction de référentiels) ?

La question centrale des standards et vocabulaires contrôlés dans le processus de production de données de qualité

La mise en place d'un contrôle qualité va de pair avec l'établissement de référentiels, standards, indispensables à chaque étape du cycle de vie pour assurer la qualité, l'interopérabilité, le partage et la réutilisation des données de la recherche.

Eric Garnier dans son intervention a pointé les différents **obstacles à l'interopérabilité des données** dans le champ de l'écologie fonctionnelle végétale (obstacles techniques comme par exemple la dispersion des données, leur hétérogénéité syntaxique et sémantique, le manque de traçabilité, les obstacles socio-culturels également ou le manque de temps dévolu à la structuration et au stockage des données).

Christine Coatanoan a évoqué la **grande variété des données recueillies** par divers systèmes d'information (navires, observatoires, satellites, réseaux d'observateurs) qu'il faut pouvoir organiser, classer, analyser et pour lesquelles il est important de pouvoir identifier l'origine et les traitements réalisés afin de les rendre lisibles, produire des informations fiables, standardisées et de confiance.

Geoffrey Aldebert a mentionné 200 000 ressources hébergées sur la plateforme DataGouv.fr et bon nombre de jeux de **données parcellaires avec des formats hétérogènes** difficilement réutilisables.

L'usage des métadonnées et des standards sont la base de l'interopérabilité, elle-même clé du système pour pouvoir lire les données. Cette interopérabilité s'opère via **l'utilisation de vocabulaires contrôlés** (plus de 70 000 termes et 70 listes de vocabulaire dans le cas de SeaDataNet sur des navires, des unités de mesure, des codes ou listes de températures), de formats ouverts (net CDF, ODVA, ASCII en océanographie) et **l'adoption de normes et de standards** largement partagés (norme de métadonnée ISO1915 pour SeaDataNet, format UNIMARC pour homogénéiser les métadonnées de Frantiq etc.)

La démarche initialisée autour de Schéma data.gouv et présentée par Geoffrey Aldebert montre bien quant à elle la position centrale du schéma de métadonnée dans le processus mis en place par Etalab pour parvenir à publier des données de qualité. On saisit clairement l'intérêt de pouvoir disposer d'une plateforme qui recense l'ensemble des schémas partagés par différents producteurs de données et permet d'accéder à la documentation du modèle avec ses évolutions au cours du temps.

L'utilisation de **vocabulaires communs**, de **termes normalisés** est une condition préalable importante pour **garantir la cohérence et la compréhension des données**.

Dans le cas du projet TRY présenté par Eric Garnier (constitution d'une base de données mondiale comprenant 8000 sites de mesures, 130 contributeurs 239 datasets et 973 différents traits etc.), la constitution du thésaurus TOP (Thesaurus of Plant Carateristics) a été nécessaire pour réaliser le travail d'homogénéisation sémantique et syntaxique permettant de comprendre et d'analyser l'espace phénotypique des plantes.

L'intervention de Véronique Humbert et Blandine Nouvel présentant le réseau Frantiq, producteur de métadonnées bibliographiques et thématiques pour l'archéologie met en également en lumière l'apport des standards et outils normalisés pour organiser et faciliter l'accès à de multiples ressources. En témoigne aujourd'hui la place occupée par le thésaurus Pactols, positionné au centre d'un réseau de ressources, figurant comme outil de référence, outil pivot d'interopérabilité entre différentes bases de données.

Les expériences relatées autour de Frantiq, Try, Schéma.gouv.fr ou au SeaDataNet montrent clairement les **besoins d'homogénéisation des termes scientifiques et techniques** et positionnent les vocabulaires communs, les normes, standards et métadonnées au cœur même des activités de recherche et du processus de qualité. Elles posent en même temps l'épineuse question de la difficulté d'élaborer des référentiels et de l'organisation à déployer pour y parvenir.

La variété des échanges avec les intervenants a montré l'intérêt suscité par ces retours d'expérience et a révélé des interrogations liées

- À l'organisation nécessaire pour mener une démarche vers la qualité : *Comment construit-on un modèle de schéma ? D'où vient l'initiative ? Comment et sur quelles bases se constituent les groupes de travail, les communautés d'acteurs ? Comment se gère le risque de développer des standards de métadonnées parallèles, Comment s'opère la coordination et à quelle échelle ?*

- À la difficulté de créer et maintenir des standards ou référentiels : *Quelle est la marche à suivre pour créer un référentiel ? Quelle est la bonne procédure d'adoption des vocabulaires communs ? Comment s'organise le consensus dans l'espace et dans le temps ? Avec quels outils ? Comment se gère l'ajout de nouveaux termes ? Comment se prennent les décisions ? comment faire adopter un référentiel ?*
- Au temps et à l'investissement nécessaire à une telle démarche : *Combien de temps a pris le développement du thésaurus TOP ? Est-il réaliste de penser que l'on peut développer une démarche similaire pour chaque les thématiques scientifiques ?*

Elle a aussi marqué en particulier le souci de

- Voir les actions entreprises consolidées et pérennisées en les inscrivant dans un cadre européen ou international (*Est-ce que l'initiative sur les schémas a une équivalence européenne ou internationale ? Est-ce que le thésaurus Pactols est en lien avec des projets internationaux ?*)
- Mener ces actions dans le sens d'une démarche durable, concertée, en adéquation avec les actions de politique nationale ou la législation (RGPD) (*Comment se positionne le projet de plateforme nationale fédérée des données de la recherche par rapport à la plateforme data.gouv ? Comment s'organise l'incitation à publier ? Comment se gère le conflit de respect de l'anonymat sur la plateforme des transactions immobilières ?)*

Le webinaire a pour finir permis de relever un certain nombre de considérations importantes à prendre en compte dans la démarche organisationnelle et pratique visant la production de données de qualité. On retiendra plus particulièrement :

- La nécessité de se conformer à des règles, d'utiliser des outils normalisés et ouverts, de documenter ses choix
- De mobiliser un réseau d'experts et de faire vivre une communauté d'utilisateurs notamment en proposant des guides méthodologiques pour faire adopter une démarche
- De mettre en place des actions de formation et d'associer différents types de compétences (documentaires, informatiques, scientifiques)
- De prendre conscience de participer à un processus d'enrichissement réciproque

Voir les présentations et vidéos de la journée