Gestion des données ouvertes en Agriculture et Nutrition

Ce cours en ligne est le fruit d'une collaboration entre les partenaires de GODAN Action, y compris Wageningen Environmental Research (WUR), AgroKnow, AidData, l'Organisation des Nations Unies pour l'Alimentation et l'Agriculture (FAO), le Forum Mondial sur la Recherche Agricole (GFAR), l'Institut des Etudes du Développement (IDS), le Land Portal, l'Open Data Institute (ODI) et le Centre Technique de Coopération Agricole et Rurale (CTA).

GODAN Action est un projet de trois ans du Département pour le Développement International du Royaume-Uni pour permettre aux utilisateurs, producteurs et intermédiaires de données de s'engager efficacement avec les données ouvertes et maximiser leur potentiel d'impact dans les secteurs de l'agriculture et de l'alimentation. Nous travaillons en particulier à renforcer les capacités, à promouvoir des normes communes et les meilleures pratiques et à améliorer la manière dont nous mesurons l'impact. [www.godan.info]

MODULE 4 : PARTAGE DES DONNÉES OUVERTES

LEÇON 4.1: Cadre d'orientation pour le partage des données

Photo by Fintrac Inc. licensed under CC BY NC 2.0

Objectifs et résultats d'apprentissage

Cette leçon a pour objectif de :

- Présenter les cadres directeurs les plus connus pour le partage des données (FAIR, les 5 étoiles du TBL, LOD, W3C Best Practices)
- Établir la correspondance entre les principes de ces cadres et les lignes directrices pratiques sur la manière de les mettre en œuvre.

À la fin de cette leçon, vous devez être en mesure :

- de comprendre les plus importants principes de ces cadres d'orientation pour le partage des données ouvertes
- de comprendre leur principales implications pratiques
- d'évaluer outils et différentes façons de publier les données, en fonction de ces cadres.
- d'identifier les problèmes de durabilité dans une politique de données ouvertes, et être conscient de la façon de les surmonter.

Sommaire

Liste des illustrations

1. Cadres directeurs pour les données : De l'ouvert au FAIR

Le partage des données, surtout parmi les chercheurs, s'est toujours fait de différentes façons au fil des siècles. Dans l'agriculture comme dans d'autres domaines, l'échange de données est indispensable pour la recherche et l'innovation, mais aussi pour le fonctionnement des entreprises (par exemple, les prix du marché, les infrastructures, les informations météorologiques) ainsi que pour les politiques et réglementations (par exemple, le suivi des produits alimentaires et l'utilisation des pesticides). Les parties concernées sont donc variés et leurs intérêts sont parfois alignés ou non, mais en fin de compte, ils ont tous besoin de données provenant d'autres acteurs ou d'autres parties de la chaîne de valeur.

Un facteur clé qui a changé le mode de partage des données au cours des dernières décennies est bien sûr l'avènement de l'Internet et, plus récemment, des technologies de cloud et de bigdata qui ont multiplié le potentiel de puissance de traitement des données. Au début de la toile, son inventeur Tim Berners-Lee l'a vue comme une " toile de données " et au cours de la dernière décennie, le monde en est venu à reconnaître que les principaux consommateurs de données et les véritables intermédiaires dans le processus de partage des données sont les machines et que, pour être partagées, les données doivent donc être lisibles par machine.
D'autre part, la facilité de partage des données sur le web entraîne des préoccupations juridiques quant à l'utilisation qui peut être faite de ces données : le fait qu'elles soient sur le web signifie seulement qu'elles peuvent être lues, mais rien n'est dit sur la permission de réutiliser, modifier ou redistribuer. Si les autorisations ne sont pas explicites, les données ne peuvent pas être légalement réutilisées.

Le partage des données nécessite donc des accords sur la manière dont les données doivent être " écrites " pour les machines (parce que les machines doivent être programmées pour les lire et doivent connaître les règles mais aussi sur la manière dont les données peuvent être utilisées– en bref, les droits d'utilisation et la manière de les formaliser également (pour plus d'informations sur la question la licence des données, voir le module 5). S'accorder sur ce point peut n'être qu'une question d'accords ad hoc ou de spécifications techniques, ou, dans une perspective plus ambitieuse et à long terme, elle peut conduire à la création de cadres directeurs généraux faisant autorité. Dans cette leçon, nous présenterons deux principaux cadres qui ont été établis et approuvés pour le partage des données.

1.1. Le plan de déploiement 5 étoiles de Tim Berners-Lee pour les données ouvertes

Le premier cadre directeur pour le partage des données a été conçu par l'inventeur du Web lui-même, Sir Tim Berners-Lee. Le cadre technique qu'il a conçu pour le réseau de données est la "Linked Open Data" (LOD) ou simplement une bonne pratique de gestion des données liées[1], qui a été entièrement formalisé en 2006 ; ce cadre consiste en des directives très techniques pour relier les données et elle recommande essentiellement:

- Utiliser les URI comme noms
- Utiliser les URI HTTP pour que les gens puissent les consulter lorsque quelqu'un cherche une URI, fournir des informations utiles; et
- Inclure des liens vers d'autres URI, afin de favoriser la découverte.

Toutefois, étant donné la forte barrière à l'entrée de ces solutions technologiques, il a publié vers 2010 le schéma plus complet de déploiement de données ouvertes 5 étoiles, "afin d'encourager les gens - en particulier les propriétaires de données gouvernementales - sur la voie d'une bonne liaison des données ".[2] Le schéma 5 étoiles illustre le " continuum " de la publication de données qui mène aux dernières étapes de la publication de données ouvertes entièrement liées (le cadre LOD n'en est qu'un sous-ensemble, les deux dernières étoiles).

★	Rendez votre matériel disponible sur le web (quel que soit le format) sous licence ouverte.
★★	Rendez-le disponible sous forme de données structurées (par exemple, Excel au lieu d'une image numérisée d'un onglet).
★★★	Format non propriétaire (par exemple CSV au lieu d'Excel)
★★★★	Utilisez les URI pour identifier les objets, afin que les gens puissent retrouver votre matériel aisément.
★★★★★	Reliez vos données à celles d'autres personnes pour établir le contexte.

Illustration 1 Schéma de déploiement de données ouvertes 5 étoiles de Tim Berners-Lee

Les 5 étoiles de TBL restent un cadre de référence pour tous ceux qui travaillent sur des données ouvertes. Ils ont normalement été interprétés comme étant " cumulatifs ", en ce sens que " chaque étoile supplémentaire suppose que les données répondent aux critères de l'étape ou des étapes précédentes",[3] ce qui souligne l'importance de

[1] http://linkeddata.org/ and https://www.w3.org/wiki/LinkedData
[2] Tim Berners-Lee. Linked Data. https://www.w3.org/DesignIssues/LinkedData.html
[3] https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data

la première étape, une licence ouverte, en l'absence de laquelle même l'application des quatre autres étoiles ne vous conduirait pas à des données vraiment ouvertes. Toutes les autres étoiles concernent l'interopérabilité des données (nous le verrons mieux dans le module 2 et dans les leçons 4.2 et 4.3, ainsi que dans le cadre étroitement associé des données ouvertes liées), tandis que la première étoile concerne l'ouverture à la réutilisation. Les 5 étoiles de TBL constituent un véritable cadre pour l'ouverture et sont donc aussi appelées " les 5 étoiles de l'ouverture ".

[Pour être plus précis, pour faire la distinction entre les données liées et les données ouvertes liées, TBL a déclaré : " Les données liées ne sont pas nécessairement ouvertes - il y a beaucoup d'utilisations importantes des données liées en interne et pour les données personnelles et celles du groupe concerné. Vous pouvez avoir des données liées 5 étoiles sans qu'elles soient ouvertes. Cependant, si elles prétendent être des données ouvertes liées, elles doivent l'être pour obtenir n'importe quelle étoile"[4]]

Le concept "d'ouverture" est la pierre angulaire de toutes les initiatives en matière de partage des connaissances et des données depuis des décennies. La définition officielle de " données ouvertes " est que ce sont des données qui peuvent être librement utilisées, réutilisées (modifiées) et redistribuées (partagées) par quiconque[5]. Il s'agit d'un concept général qui s'applique à tout type de ressource (document, image, ensemble de données), il comporte quelques règles et est assez facile à appliquer.

Toutefois, ces dernières années, certaines des difficultés reliées à ce cadre ont été perçues comme pouvant décourager le partage des données, en particulier dans le contexte d'une recherche intensive et des données transmises à travers différentes étapes de la chaîne de transmission des données, car elles sont très strictes :

- La définition de " licence ouverte " est très stricte : dans son cours de base sur les données ouvertes, l'Open Data Institute (ODI) déclare que " cette licence doit permettre aux utilisateurs d'utiliser les données comme ils le souhaitent, y compris de les transformer, les combiner et les partager avec d'autres, même commercialement "[6].

[4] Tim Berners-Lee. Linked Data. https://www.w3.org/DesignIssues/LinkedData.html
[5] http://opendatahandbook.org/guide/en/what-is-open-data/
[6] http://training.theodi.org/InADay/#/id/co-01; see also http://opendefinition.org/od/2.1/en/

L'exigence d'une licence ouverte pourrait empêcher le partage de données qui ont certaines restrictions d'accès, mais qui peuvent être réutilisées dans de grandes communautés ou par le biais de simples accords et peuvent donc avoir un impact important.

- Les quatrième et cinquième étoiles, surtout dans les versions qui décrivent chaque étoile plus en détail, sont parfois considérées comme trop liées à l'approche technique du Cadre de description des ressources (voir les sections suivantes et la leçon 4.2), au lieu d'être des principes génériques qui peuvent être mis en œuvre avec toute technologie.

Cela ne signifie pas que ce cadre a été remplacé : il reste le cadre de référence pour une interopérabilité élevée des données et pour un réseau de données ouvertes, plus ou moins couplé et ascendant.

1.2. Les principes FAIR

Récemment, étant donné que la réutilisation des données est unanimement reconnue comme un moteur important de l'innovation et que la façon dont les données sont partagées est essentielle à leur réutilisation, un nouvel intérêt s'est manifesté pour la définition d'un cadre plus formel et mieux coordonné qui pourrait servir davantage à la recherche intensive de données et au partage des données à travers la chaîne de valeur.

En 2014, la nécessité de mieux définir les " règles " pour un partage plus efficace des données a conduit un groupe de représentants de différentes parties prenantes - universités, industrie, éditeurs, agences de financement - à se réunir à Leyde, aux Pays-Bas, pour discuter d'un " ensemble minimal de principes directeurs et pratiques acceptés par la communauté "[6].

Ces discussions ont débouché sur un ensemble de principes appelés les principes FAIR : selon ces principes, les données doivent l'être : Trouvables, accessibles, interopérables et réutilisables (FAIR: Findable, Accessible, Interoperable and Reusable).

[6] Force11. Guiding Principles For Findable, Accessible, Interoperable And Re-usable Data
Publishing Version B1.0. https://www.force11.org/fairprinciples

En se contentant de lire ces quatre principes, ce cadre semble tout à fait conforme au concept d'ouverture et au système 5 étoiles de TBL (et en fait, il n'y a pas de discontinuité entre les deux, et le cadre FAIR ne contredit ni ne remplace le cadre 5 étoiles de TBL) ; cependant, à la lecture des détails sur chacun des principes, il existe quelques différences clés et un niveau supérieur ou généralisation qui font du principe FAIR un véritable "cadre" formel :

A. L'accent est mis sur des droits d'accès clairs plutôt que sur l'ouverture, conformément à la nécessité d'une plus grande flexibilité (mais aussi d'une plus grande précision) dans le partage des données, afin de faciliter également le partage des données qui peuvent être soumises à certaines restrictions d'accès mais peuvent encore être réutilisées sous certaines conditions: Nous apprécions les exceptions au libre accès intégral aux données (par exemple pour des raisons liées à la vie privée des patients ou à la propriété intellectuelle). C'est pourquoi nous considérons qu'une licence adaptée aux objets des données est la clé de la publication de données FAIR".

B. Une attention particulière est accordée à la provenance, à l'attribution et à la persistance, conformément au fait que les principes FAIR ont été convenus par une communauté qui veut travailler ensemble et partager des données et qui a besoin d'un environnement de confiance avec quelques règles de base. Sous cet aspect, les principes FAIR exigent des règles plus rigoureuses en matière de confiance que le réseau de données ouvertes plus ascendant conçu par TBL.

C. Les principes ne veulent pas aller au niveau des spécifications techniques ; ils sont : un "guide général sur des données FAIR", et non une "spécification". Lors de l'élaboration des principes directeurs FAIR, les choix techniques de mise en œuvre ont été sciemment évités. Les [Principes directeurs FAIR] sont destinés à guider les responsables de la mise en œuvre des environnements de données FAIR pour vérifier si leurs choix de mise en œuvre particuliers rendent effectivement les données résultantes FAIR'.

D. Contrairement aux 5 étoiles de TBL, les principes FAIR ne sont pas " cumulatifs ": " Ces facettes FAIR sont évidemment liées, mais techniquement quelque peu indépendantes les unes des autres, et peuvent être mises en œuvre dans n'importe quelle combinaison, progressivement, à mesure que les fournisseurs de données (...) évoluent vers un degré croissant d'état FAIR (FAIR-ness)''

Voici quelques détails clés sur chaque principe qui illustrent les caractéristiques décrites ci-dessus :

1. Quelques détails du principe Trouvable :

a) Les objets de données doivent être persistants et mettre L'accent sur leurs métadonnées.
b) Les identificateurs de tout concept utilisé dans les objets de données doivent être uniques et persistants

2. Quelques détails du principe Accessible:

a) Sur autorisation appropriée' - Même les données qui nécessitent une autorisation spéciale sont considérées comme FAIR si elles appliquent les autres principes, de sorte que les données ne doivent pas nécessairement être utilisables "par n'importe qui".
b) Par le biais d'un protocole bien défini' - cela signifie que l'interopérabilité n'est pas liée à un seul protocole (par exemple HTTP et plus spécifiquement une API REST ou un terminal SPARQL, qui sous-tendent le cadre Données liées).

3. Quelques détails du principe Interopérable:

a) "Les (méta)données sont actionnables par la machine"
b) Les formats de (méta)données utilisent des vocabulaires et/ou ontologies partagés
Les (méta)données à l'intérieur de l'objet de données doivent donc être à la fois syntaxiquement analysables et sémantiquement accessibles par machine" - ces deux indications sont bien sûr parfaitement réalisées en RDF, mais aucune technologie spécifique n'est mentionnée. En général, l'ensemble du principe d'interopérabilité est tout à fait conforme aux étoiles de TBL et tous les principes reposent sur l'hypothèse clé que les données doivent être "FAIR pour les machines aussi bien que pour les personnes", et que "la lisibilité des métadonnées par machines est une condition sine qua non de la FAIRness".

4. Parmi les indications pratiques qui détaillent le principe Réutilisable, nous avons: Les objets de données publiés doivent faire référence à leurs sources avec des métadonnées et une provenance suffisamment riche pour permettre une citation appropriée' - ceci est conforme à l'objectif de construire une infrastructure de référentiels de données fiables où la paternité et l'attribution sont particulièrement importantes.

Les principes FAIR sont rapidement adoptés par de nombreux acteurs, en particulier les bailleurs de fonds de la recherche. Elles ont été récemment adoptées par les lignes directrices de la Commission européenne sur la gestion des données FAIR à l'horizon 2020.

1.3. Autres cadres généraux

Il y a d'autres cadres à prendre en compte lors de la publication de données ouvertes, un plus générique que les deux précédents (Accès libre), un plus strict sur les licences (Contenu libre) et un autre très détaillé sur les " meilleures pratiques " techniques avec des choix clairs en matière de mise en œuvre (Données W3C sur les bonnes pratiques Web).

Les mouvements Open Access et Open Science plaident respectivement pour la publication des résultats de la recherche (des articles de revues aux thèses en passant par les ensembles de données) de manière accessible ainsi que pour rendre " la recherche scientifique, les données et la diffusion accessibles à tous les niveaux de la société, amateur ou professionnel ". Suivre les indications sur la manière de parvenir à un accès ouvert en suivant soit la voie verte[7] (auto-archivage), soit la voie d'or[8] (publication dans des revues en accès libre, dont certaines publient maintenant aussi des ensembles de données) est déjà une étape essentielle du partage des données.

Dans les nouveaux projets financés par l'UE, le libre accès aux données de recherche est une exigence pilote qui sera obligatoire dans le prochain programme-cadre. Des exemples d'endroits où des ensembles de données pourraient être publiés sont également démontrés dans des catalogues publics comme OpenAIRE [9] ou Dryad[10].

Les données W3C sur les meilleures pratiques du web[11] sont des lignes directrices très détaillées concernant la publication et l'utilisation des données sur le Web conçues pour aider à soutenir " un écosystème autosuffisant ". Par rapport aux principes FAIR, ces meilleures pratiques vont beaucoup plus loin dans la mise en œuvre technique effective et recommandent des solutions spécifiques.

[7] https://en.wikipedia.org/wiki/Open_access#Self-archiving:_green_open_access
[8] https://en.wikipedia.org/wiki/Open_access#Journals:_gold_open_access
[9] https://www.openaire.eu
[10] http://datadryad.org/
[11] https://www.w3.org/TR/dwbp/

L'approche du partage des données est très similaire à celle des principes FAIR, soulignant la nécessité de prévoir également la publication de données à accès contrôlé, la nécessité de la fiabilité et de la persistance des données, et la nécessité de convenir d'un ensemble de règles communes.

Les meilleures pratiques sont toutes liées à un ensemble " d'avantages ": chaque avantage représente une amélioration dans la façon dont les ensembles de données sont rendues disponibles sur le Web: Compréhension (humaine), Traitabilité, Découvrabilité, Réutilisation, Confiance, Capacité de liaison, Accès, Interopérabilité.

Illustration 2 Les avantages du W3C pour l'application des données sur les meilleures pratiques du web

Comme vous pouvez le voir, les principes FAIR sont tous couverts, mais le cadre W3C couvre aussi un pas en arrière (compréhension humaine) et un pas en avant (traitabilité des données). Nous mentionnerons certaines des solutions techniques recommandées comme meilleures pratiques W3C dans la prochaine section.

Une définition encore plus ouverte de ce que l'on entend par " ouvert " est donnée par les 5R pour le contenu libre[12]: Le contenu libre est " toute œuvre pouvant faire l'objet d'un droit d'auteur (à l'exclusion traditionnellement des logiciels, qui sont décrits par d'autres termes tels que "open source") qui fait l'objet d'une licence qui donne aux utilisateurs la permission libre et perpétuelle de s'engager dans les activités des 5R : Retenir, Réutiliser, Réviser, Remixer, Redistribuer "

1.4. Un cadre de données ouvertes pour l'agriculture et la nutrition

Outre ces cadres généraux pour le partage des données, il y a eu des initiatives consacrées à la promotion des données ouvertes dans des secteurs spécifiques. Pour l'alimentation et l'agriculture, après un accord lors de la Conférence internationale du G8 sur les données ouvertes pour l'agriculture en 2012, l'initiative Global Open Data for

[12] http://opencontent.org/definition/

Agriculture and Nutrition [13] (GODAN) a été lancée lors de la Conférence sur le partenariat ouvert des gouvernements en octobre 2013. L'initiative se concentre sur l'obtention d'un soutien de haut niveau parmi les gouvernements, les décideurs politiques, les organisations internationales et les entreprises.

GODAN a une déclaration d'intention [14] à laquelle plus de 500 partenaires ont adhéré jusqu'à présent, et la déclaration est orientée vers des données totalement ouvertes:

''L'initiative Global Open Data for Agriculture and Nutrition (GODAN) vise à soutenir les efforts mondiaux pour rendre les données agricoles et nutritionnelles pertinentes disponibles, accessibles et utilisables sans restriction dans le monde entier.'

Toutefois, elle reconnaît les préoccupations légitimes quant à la l'ouverture complète :

'L'initiative plaide en faveur de "politiques de données ouvertes et d'accès libre par défaut, dans les secteurs public et privé, tout en respectant et en s'efforçant de concilier ouverture et préoccupations légitimes concernant la vie privée, la sécurité, les droits communautaires et les intérêts commerciaux".

1.5. Outils d'évaluation des données connexes

Certains outils ont été développés pour évaluer l'ouverture et/ou la FAIRness des données : deux exemples sont les certificats ODI [15] développés par l'Open Data Institute (un outil pour évaluer et reconnaître la publication durable de normes de données ouvertes de qualité, s'appuyant sur des cadres tels que opendefinition.org, 5star Open Data, Sunlight principles[16], DCAT) and the Data Seal of Approval[17] par les services néerlandais d'archivage et réseau (très conformes aux principes FAIR mais plus liés à la qualité des dépôts numériques, pas à des jeux spécifiques).

[13] http://godan.info
[14] http://www.godan.info/pages/statement-purpose
[15] https://certificates.theodi.org/en/
[16] This is a framework to assess the openness of government data: https://sunlightfoundation.com/policy/documents/ten-open-data-principles/
[17] https://www.datasealofapproval.org

2. Recommandations pratiques pour le partage des données

Certains des principes décrits précédemment ont des implications en termes de politique des données, mais la plupart d'entre eux ont de lourdes implications en termes d'options de mise en œuvre technique. En ce qui concerne en particulier le principe d'interopérabilité, de nombreuses implications sont très techniques et elles seront examinées plus en détail dans les leçons 4.2-4.4. Les détails sur la mise en œuvre d'un bon référentiel de données sont abordés dans la leçon 3.3 et vous pouvez les comparer avec les exigences énumérées ici. Nous ne fournissons ici que quelques recommandations générales qui peuvent aider le gestionnaire de données à choisir les outils appropriés ou à guider les développeurs vers des solutions technologiques pour appliquer les principes les plus importants du partage des données. Dans ce texte, nous mentionnerons comme exemples d'outils de référentiels de données CKAN[18] et Dataverse[19] (voir la leçon 3.3).

Certains des principes décrits précédemment ont des conséquences en termes de politique des données, mais la plupart d'entre eux ont de lourdes implications en termes de choix de mise en œuvre technique. En ce qui concerne en particulier le principe d'interopérabilité, de nombreuses implications sont très techniques et elles seront examinées plus en détail dans les leçons 4.2-4.4. Les détails sur la mise en œuvre d'un bon référentiel de données sont abordés dans la leçon 3.3 et vous pouvez les comparer avec les exigences énumérées ici. Nous ne fournissons ici que quelques recommandations générales qui peuvent aider le gestionnaire de données à choisir les outils appropriés ou à guider les développeurs vers des solutions technologiques pour appliquer les principes les plus importants en relation avec le partage des données. Nous mentionnerons comme exemples d'outils de référentiels de données CKAN et Dataverse dans la leçon 3.3.

Bien que les principes FAIR ne spécifient pas de choix explicites quant à la mise en œuvre technique, le système 5 étoiles de TBL fournit quelques exemples et propose des technologies spécifiques. Mais surtout, le document qui peut aider à identifier des solutions technologiques pour la publication de données est le W3C Data on the Web Best Practices (DWBP) mentionné précédemment, auquel nous ferons souvent référence.

[18] https://ckan.org/
[19] https://dataverse.org/

Il s'agit d'un document très technique, mais qui contient des conseils pour pratiquement tout ce qui est nécessaire à la publication de données sur le Web, et il est conforme à la fois aux 5 étoiles de TBL et aux principes FAIR.

2.1. Octroi de licences

Comme nous l'avons vu, les prescriptions pour l'octroi de licences sont différentes dans le schéma de TBL (licence ouverte) et dans le cadre de FAIR (toute licence qui clarifie les droits d'utilisation).

Bien sûr, il s'agit avant tout d'un choix de politique en matière de données - vous devez tenir compte de plusieurs facteurs, notamment : si votre recherche a été financée par l'État et que vous aviez le mandat de publier les résultats en libre accès ; si votre ensemble de données comprend des données sensibles ; si vos données ont une valeur commerciale (ou si elles pourraient entrer dans un espace précompétitif) ; et bien sûr, votre politique organisationnelle.

Vous pouvez même publier différentes parties de vos ensembles de données sous différentes licences ou appliquer des licences à des éléments de données spécifiques (selon le modèle FAIR' modulaire' et 'récurrent' Data Object).

Les deux approches en matière d'octroi de licences peuvent en tout état de cause être mises en œuvre conformément à la 4e meilleure pratique du DWBP : Fournir des informations sur les licences de données

''Dans le contexte des données sur le Web, la licence d'un ensemble de données peut être spécifiée dans les données, ou en dehors de celles-ci, dans un document séparé auquel il est lié. (...) Il devrait être possible pour les machines de détecter automatiquement la licence de données d'une distribution''

DWBP fournit également des exemples de mise en œuvre : un vocabulaire qui peut être utilisé pour exprimer les métadonnées de la licence (Dublin Core Terms, schema.org) et même des langages de droits lisibles par machine tels que le Creative Commons Rights Expression Language, l'Open Data Rights Language et l'Open Data Rights Statement Vocabulary.

Idéalement, pour être interprétées de manière univoque, les métadonnées de licence devraient pointer vers une URI ou au moins vers l'URL d'une licence publiée.

Pour en savoir plus sur l'octroi de licenses de données, voir le module 5 de ce cours.

2.2. Infrastructure du référentiel : protocoles, persistance et identificateurs

Nous regroupons ici un certain nombre d'exigences architecturales qui concernent le référentiel plus que l'ensemble des données en soi : ces exigences ont des conséquences profondes dans le choix des solutions techniques pour la réalisation de votre catalogue. Les détails techniques sur la façon de mettre en œuvre ceci nécessiteraient une leçon dédiée : à moins que vous ne développiez vous-même la plate-forme de référentiel, vous pouvez utiliser ces exigences comme critères d'évaluation pour la plate-forme de référentiel que vous allez choisir. Pour en savoir plus sur la mise en place d'un référentiel de données, lisez la leçon 3.3 et sur les identificateurs persistants, lisez la leçon 2.5.

La 4e étoile de TBL recommande simplement d'utiliser les URI pour indiquer les objets, afin que les gens puissent facilement les localiser.

Le cadre FAIR a des exigences beaucoup plus élevées pour le référentiel

- Les (méta)données se voient attribuer un identificateur globalement unique et éternellement persistant
- Les (méta)données sont récupérables par leur identificateur à l'aide d'un protocole de communication standardisé
- Le protocole est ouvert, libre et universellement applicable
- Le protocole prévoit une procédure d'authentification et d'autorisation, le cas échéant
- Les métadonnées sont accessibles, même lorsque les données ne sont plus disponibles.

En ce qui concerne les protocoles, DWBP indique comme Meilleure Pratique 21 : Utiliser des Interfaces Web Standardisées -'il est recommandé d'utiliser des URI, des verbes HTTP, des codes de réponse HTTP, des types MIME, des liens HTTP tapés et la négociation de contenu lors de la conception des API'. La meilleure pratique recommande explicitement les API RESTful.

En ce qui concerne les identificateurs persistants, les pratiques les plus courantes sont l'utilisation des :

- Identificateurs de ressources uniques (URI) qui résolvent les URL
- Identificateurs d'objets numériques (DOI).

La pratique 9 du DWBP **:** " Utiliser les URI persistantes comme identificateurs d'ensembles de données " est largement basée sur le cadre de données liées, et fournit des liens vers de nombreux documents techniques sur la façon de construire les URI et d'assurer la persistance des URI. Toutefois, il examine également les DOI et suggère une façon de fusionner les deux approches en ajoutant le DOI à un modèle URI : " Les identificateurs d'objets numériques (DOI) offrent une alternative similaire. Ces identificateurs sont définis indépendamment de toute technologie Web mais peuvent être ajoutés à une " URI stub ". Les DOI sont une partie importante de l'infrastructure numérique pour les données de recherche et les bibliothèques"

Bien que la mise en œuvre de tout ce qui précède soit techniquement difficile, la plupart des outils de référentiels de données existants créent des URI (par exemple CKAN) ou des DOI (par exemple Dataverse) pour les ensembles de données téléchargées et utilisent des protocoles ouverts (outre http, ils utilisent des API REST comme SPARQL ou OAI-PMH). (Si vous voulez être entièrement compatible avec Linked-Data, il est important que vous choisissiez un outil qui crée des URIs déréférençables, permettant la négociation de contenu[20] et l'utilisation du protocole SPARQL.)

Cependant, le domaine URI est sous la responsabilité du propriétaire du référentiel et l'application du principe de persistance exigera un certain engagement au niveau des politiques pour maintenir le domaine URI (ou les DOI). Pour les URI, DWBP propose une solution alternative : " Lorsqu'un éditeur de données ne peut ou ne veut pas gérer son espace URI directement pour la persistance, une autre approche consiste à utiliser un service de redirection tel que purl.org. On obtient ainsi des URI persistantes qui peuvent être redirigées au besoin pour que l'emplacement éventuel soit éphémère ‘’

2.3. Formats de données

En ce qui concerne les formats de données, il y a des recommandations faciles qui respecteront les principes de tous les cadres que nous avons décrits :

- La 3e étoile de TBL recommande: " il faut rendre les données disponibles dans un format ouvert non exclusif (par exemple,

[20] http://linkeddata.org/conneg-303-redirect-code-samples

CSV au lieu d'Excel) ".

- FAIR : Le point 1 d'interopérabilité recommande :
  'Les métadonnées doivent utiliser un langage formel, accessible, partagé et largement applicable pour la représentation des connaissances'.

Fondamentalement, les données doivent être " sérialisées ", présentées dans un format lisible et facile à analyser par machine, éventuellement sans avoir besoin d'un logiciel propriétaire. Il existe des formats techniquement lisibles par machine (HTML, Excel), mais ils ne sont pas nécessairement faciles à analyser, soit parce qu'ils ne sont pas rigoureusement structurés, soit parce que les algorithmes pour les analyser sont propriétaires.

Nous recommandons la meilleure pratique 12 du DWBP du W3C: Utiliser des formats de données standardisés et lisibles par machine : Rendre les données disponibles dans un format de données normalisé lisible par machine et facilement analysable, y compris, mais sans s'y limiter, CSV, XML, Turtle, NetCDF, JSON et RDF"

La plupart des outils de référentiels de données existants exposent les données en RDF (par exemple, CKAN) ou en JSON ou XML (par exemple, Dataverse). Pour plus d'informations à ce sujet, voir la leçon 4.3 sur l'interopérabilité structurelle.

2.4. Les métadonnées

Les 5 étoiles de TBL ne mentionnent pas les métadonnées.

Les principes FAIR définissent un rôle fondamental pour les métadonnées, pour leur répétabilité :

- Les données sont décrites à l'aide de métadonnées riches
- Les métadonnées spécifient l'identificateur de données

Et pour la réutilisabilité, les méta(données):

- ont une pluralité d'attributs précis et pertinents
- sont publiées avec une licence d'utilisation claire et accessible
- sont associées à leur provenance

- satisfont aux normes communautaires relatives au domaine.

Une remarque à propos des métadonnées : si vous utilisez un outil de référentiel de données, cela est normalement hors de votre contrôle. Les dépôts de données sont livrés avec leurs modèles de métadonnées et leur couche d'exposition aux métadonnées, et dans la plupart des cas, les modifier signifie pirater le code de programmation. Il est vrai que ces outils tendent à respecter les normes de métadonnées existantes (par exemple, CKAN expose les métadonnées à l'aide du vocabulaire de l'EDSC, mais pas entièrement), mais ce n'est pas toujours le cas, et dans certains cas vous pouvez vouloir ajouter des normes spécifiques au domaine (par exemple, pour vous conformer aux normes communautaires relatives au domaine).

Ainsi, d'une part, il est très important d'évaluer le modèle de métadonnées d'un outil de référentiel de données avant de l'adopter; d'autre part, il est souhaitable que les outils permettent aux gestionnaires de données de facilement adapter leur modèle de métadonnées.

Même si cela est géré par l'outil de référentiel de données, il peut être intéressant de consulter la meilleure pratique 1 du DWBP : " Fournir des métadonnées ", notamment en ce qui concerne les métadonnées lisibles par machine. Tout d'abord, conformément à la meilleure pratique consistant à utiliser des formats facilement analysables, il recommande de sérialiser les métadonnées dans des formats tels que Turtle ou JSON (ou de les intégrer dans la page HTML en utilisant RDFa ou Json-LD), puis d'utiliser les vocabulaires existants : " Lors de la définition de métadonnées lisibles par machine, il est fortement recommandé de réutiliser les termes standard existants et les vocabulaires populaires. Par exemple, les termes des métadonnées du Dublin Core (DCMI) et le Vocabulaire du catalogue de données devraient être utilisés pour fournir des métadonnées descriptives".

L'utilisation des normes de métadonnées existantes aide à mettre en œuvre l'exigence FAIR selon laquelle les métadonnées doivent satisfaire aux normes communautaires pertinentes au domaine et, espérons-le, aux exigences relatives aux métadonnées de licence et de provenance, en considérant que les vocabulaires appropriés des ensembles de données devraient inclure ces métadonnées.

La leçon 4.4 sur l'interopérabilité sémantique traite davantage des métadonnées.

2.4.1. Métadonnées et catalogues

Il est préférable que les métadonnées soient dans l'ensemble de données lui-même (de nombreux formats et vocabulaires structurés permettent des modèles hiérarchiques ou relationnels qui comprennent des métadonnées sur l'ensemble de données et les données réelles) pour la découverte de soi.

Cependant, l'auto-découverte supposerait une infrastructure existante d'ensembles de données distribués et de dépôts d'ensembles de données qui exposent des métadonnées standardisées et qui sont explorées ou fédérées par des recherches distribuées, ce qui n'est pas le cas. Les principes FAIR demandent que les métadonnées des ensembles de données soient enregistrées dans des catalogues d'ensembles de données où elles peuvent être trouvées plus facilement : les " (méta)données sont enregistrées ou répertoriées dans une ressource consultable ".

De nombreux outils de référentiels de données offrent normalement aussi de bonnes fonctions de catalogues de données, offrant des fonctionnalités de recherche de métadonnées et exposant toutes les métadonnées par le biais d'API.

Même Tim Berners-Lee a ajouté une note à ce sujet à la page de conception de Linked Data : Aujourd'hui, en 2010, on m'a demandé avec insistance, pour les données gouvernementales, d'ajouter une nouvelle exigence, à savoir qu'il devrait y avoir des métadonnées sur les données elles-mêmes, et que ces métadonnées devraient être disponibles dans un catalogue important. (...) Oui, il devrait y avoir des métadonnées sur votre ensemble de données'[21]

2.5. Les liens

La 5e étoile de TBL recommande : " reliez vos données à d'autres données pour fournir un contexte ".

Les principes FAIR recommandent, pour l'interopérabilité : 'les (méta)données incluent des références qualifiées à d'autres (méta)données'.

[21] Tim Berners-Lee. Linked Data. https://www.w3.org/DesignIssues/LinkedData.html

C'est le cœur de l'architecture Linked Data et le mécanisme de base du Web sémantique

La meilleure pratique 10 du DWBP : Utiliser les URI persistantes comme identificateurs dans les ensembles de données, préconise que " *les ensembles de données devraient utiliser et réutiliser les URI d'autres personnes comme identificateurs si possible* ".

C'est une autre exigence qui devrait être utilisée comme critère lors du choix d'un outil de référentiel de données (et comme exigence lors du développement d'un nouvel outil) : jusqu'à présent, il semble que les outils les plus populaires ne permettent ni de lier un concept interne (comme une catégorie) à un concept externe, ni d'utiliser une URI comme valeur pour un élément de métadonnées (sauf comme chaîne, sans considérer celui-ci comme une ressource). Par conséquent, il est actuellement difficile de se conformer à cette exigence à l'aide d'outils comme Dataverse ou CKAN.

En conclusion, il est important que les gestionnaires de données analysent les répercussions pratiques de la mise en œuvre des exigences des principaux cadres de partage des données et prennent des décisions éclairées au sujet de leur dépôt de données en conséquence.