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 : LE PARTAGE DE DONNÉES OUVERTES

LEÇON 4.2: Introduction à l'interopérabilité des données

Objectifs et résultats d'apprentissage

Cette leçon a pour objectif d'expliquer les bases de l'interopérabilité des données.

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

- Comprendre les bases de l'interopérabilité des données
- Déterminer les différents types et niveaux d'interopérabilité des données.

Sommaire

Module 4 : Le partage des données ouvertes..... 2

Leçon 4.2: Introduction à l'interopérabilité des données..... 2

Objectifs et résultats d'apprentissage..... 2

Liste des illustrations.....4

1.Cadres directeurs : De l'ouvert au FAIR.....5

2.Niveaux d'interopérabilité..... 6

3.L'nteropérabilité des données et des métadonnées.....7

4.L'interopérabilité des données et ensembles de données..... 8

Lecture complémentaire..... 11

..

Liste des illustrations

- Illustration 1 Les couches de métadonnées (Luiz Olavo Bonino et al, 2016,'Fair Data Technology Update)'..... 9
- Illustration 2 Données du W3C sur le modèle des meilleures pratiques Web..... 9
- Illustration 3 Modèle FAIRPort (Luiz Olavo Bonino et al, 2016,'Fair Data Technology Update')
  10

1. Cadres directeurs : De l'ouvert au FAIR

La définition la plus utilisée de l'"interopérabilité" sur le Web est : "la capacité d'un système ou d'un produit à fonctionner avec d'autres systèmes ou produits sans effort particulier de la part de l'utilisateur". Wikipedia le définit comme " une caractéristique d'un produit ou d'un système, dont les interfaces sont entièrement conçues pour interagir avec d'autres produits ou systèmes, présents ou futurs, dans leur mise en œuvre ou accès, sans aucune restriction".

Lorsque l'on parle d'interopérabilité des données, étant donné que les données sont sérialisées dans des ensembles de données, les définitions ii proposées peuvent être facilement appliquées à un ensemble de données en tant que produit.

L'interopérabilité des données est la capacité d'un ensemble de données à fonctionner avec d'autres systèmes ou ensembles de données sans effort particulier de la part de l'utilisateur.

Dans le cadre d'une conférence organisée par la communauté CIARD sur l'interopérabilité des données pour l'agriculture, l'interopérabilité des données a été définie comme "une caractéristique des ensembles de données permettant de récupérer, traiter, réutiliser et reconditionner facilement les données (" exploitées") par d'autres systèmes"[1]

Certaines définitions la désignent par le terme "interopérabilité entre deux systèmes", mais il est communément admis que quelque chose est réellement interopérable (ou plus interopérable) lorsque le plus grand nombre possible de systèmes peuvent l'inter-opérer. De plus, nous verrons qu'en utilisant certains formats de données et en appliquant certaines normes de données, les données peuvent être rendues "interopérables par conception" sans nécessairement savoir avec quel système elles seront interopérables : l'interopérabilité planifiée avec des systèmes spécifiques signifie que les données seront "étroitement couplées" avec ces systèmes, alors que l'interopérabilité maximale vise à un couplage en toute fluidité avec autant de systèmes que possible.

Cependant, il n'y aura jamais rien de tel qu'une interopérabilité universelle ou parfaite, un moyen de produire des données qui conviendra à tous les cas possibles. L'interopérabilité est toujours relative à un système de normes partagées et de façons communes d'utiliser des données qui sont dans certains cas très larges et polyvalentes (comme Dublin Core ou schema.org et un cas d'utilisation de moteur de recherche générique) et dans d'autres cas très spécifiques aux communautés scientifiques ou d'intérêts (comme les spécifications de données et la visualisation des séquences génétiques).

[1] Interim Proceedings of International Expert Consultation on "Building the CIARD Framework for Data and Information Sharing", CIARD (2011)
http://www.fao.org/docs/eims/upload/297074/IECProceedings-main-doc.pdf

En effet, les définitions ci-dessus désignent l'interopérabilité comme une caractéristique des seuls ensembles de données, ce qui est exact parce qu'ils font l'objet de l'interopérabilité, mais l'écosystème des acteurs et des produits qui doivent coopérer pour atteindre une interopérabilité totale est plus large: une définition intéressante de l'interopérabilité qui souligne l'importance des "attentes communes" est celle du Data Interoperability Standards Consortium (DISC): "l'interopérabilité des données concerne la possibilité pour les systèmes et services qui créent, échangent et consomment les données, de définir des attentes communes, claires quant à leur contenu, contexte et signification'[2]

2. Niveaux d'interopérabilité

L'interopérabilité peut être obtenue à différents niveaux. Alors que Wikipédia distingue l'interopérabilité syntaxique de l'interopérabilité sémantique, une distinction plus granulaire est faite dans la classification des types d'interopérabilité par la société des systèmes d'information et de gestion de la santé (HIMSS)[3]:

'L'interopérabilité "fondamentale" permet l'échange de données d'un système de technologie de l'information à un autre et ne nécessite pas la capacité du système de technologie de l'information récepteur pour interpréter les données.

Ce niveau concerne principalement les protocoles de transmission, que nous n'analyserons pas car il n'intéresse pas le gestionnaire de données, mais il couvre également certains protocoles d'échange de niveau supérieur, travaillant principalement sur le protocole HTTP commun (par exemple des API REST spéciales comme OAI-PMH, SPARQL, Linked Data API ou la négociation de contenu basée sur des requêtes de type contenu HTTP). Celles-ci peuvent intéresser les gestionnaires de données car, avant d'être lues et comprises, les données doivent être transmises, et il existe des moyens différents et plus pratiques de le faire en plus des téléchargements FTP. Nous examinerons ce niveau d'interopérabilité dans la leçon 4.3.

L'interopérabilité "structurelle" est un niveau intermédiaire qui définit la structure ou le format de l'échange de données (c'est-à-dire les normes de format de message). L'interopérabilité structurelle définit la syntaxe de l'échange de données. Il garantit que les échanges de données entre les systèmes informatiques peuvent être interprétés au niveau du champ de données.

C'est le niveau où les formats de fichiers et les formats de données jouent le rôle le plus important et c'est le niveau où les (méta)données deviennent lisibles et

[2] http://datainteroperability.org/

[3] http://www.himss.org/library/interoperability-standards/what-is-interoperability

peuvent être analysées par les machines: plus il est facile pour la machine d'analyser le format / la syntaxe des (méta)données (XML, Json, CSV...) plus les données sont structurellement compatibles. Nous examinerons ce niveau d'interopérabilité plus en profondeur dans la leçon 4.3.

L'interopérabilité " sémantique " assure l'interopérabilité au plus haut niveau, c'est-à-dire la capacité de deux ou plusieurs systèmes ou éléments à échanger des informations et à utiliser les informations qui ont été échangées. L'interopérabilité sémantique tire parti à la fois de la structuration de l'échange de données et de la codification des données, y compris le vocabulaire, afin que les systèmes informatiques récepteurs puissent interpréter les données.

Alors qu'avec l'interopérabilité structurelle, les machines comprennent ce que sont les différents éléments (et leur relation structurelle réciproque), avec l'interopérabilité sémantique, elles comprennent aussi la signification de ces éléments et peuvent les traiter avec des outils sémantiques et faire un raisonnement avancé. Les formats de données ne suffisent pas: les technologies sémantiques (voir leçon 4.4) permettent d'intégrer des éléments sémantiques lisibles par machine à partir de vocabulaires convenus dans des (méta)données sérialisées dans la plupart des formats de données existants, bien que les formats les plus appropriés à ce jour soient les différentes versions de RDF (RDF/XML, Turtle, N-Triples) et JSON-LD. Nous examinerons ce niveau d'interopérabilité dans la leçon 4.4.

Dans la leçon 4.3, nous verrons comment l'interopérabilité structurelle et architecturale peut être mise en œuvre. Dans la leçon 4.4, nous verrons comment l'interopérabilité sémantique peut être mise en œuvre, avec des exemples plus spécifiques dans la leçon 4.4.1.

Pour des recommandations plus détaillées sur la façon de mettre en œuvre l'interopérabilité des données, le document du W3C " Data on the Web Best Practices "[4] est probablement le meilleur document de référence. Elle est fortement basée sur l'approche des " données liées " (voir la leçon 4.3), mais de nombreuses recommandations aident à mettre en œuvre l'interopérabilité à différents niveaux, même si elles ne visent pas des données liées à 100%.

3. L'interopérabilité des données et métadonnées

Les données sans métadonnées ne peuvent pas être comprises par les machines. Les données sont normalement accompagnées de métadonnées.

La définition des données dans Wikipedia est : "**Les données sont des valeurs de variables qualitatives ou quantitatives, appartenant à un ensemble**

[4] W3c. Data on the Web Best Practices. https://www.w3.org/TR/dwbp/

d'éléments. Ainsi, en fin de compte, les données font toujours partie d'un ensemble d'éléments. Alors que dans l'élément individuel (ligne, enregistrement) de l'ensemble, les données sont les valeurs de certaines variables, elles seront donc toujours encapsulées dans une forme de paire clé-valeur où la clé (la variable) est finalement un élément de métadonnées qui donne sens aux données. Cette paire clé-valeur est ce qu'en termes FAIR[5] on appelle " une association unique entre deux concepts ", qui est " l'un des " éléments de données " les plus petits possibles ".

Les deux parties de la paire clé-valeur peuvent présenter des problèmes d'interopérabilité, de sorte que nous traitons des questions d'interopérabilité des métadonnées (par exemple, des schémas convenus pour les éléments de métadonnées, des noms de variables convenus) et des questions d'interopérabilité des données/valeurs (par exemple, des listes/gammes contrôlées convenues dont la valeur doit être prise, ou des questions syntaxiques). Toutefois, puisqu'il est également courant de considérer les valeurs contrôlées et la spécification de la syntaxe ou de l'unité de mesure comme des "métadonnées" (surtout parce qu'elles devraient idéalement être définies par des éléments de métadonnées séparés et non dans la valeur elle-même), nous pouvons également dire que l'interopérabilité concerne principalement les métadonnées. L'interopérabilité des données passe par l'interopérabilité des métadonnées; par exemple, nous voulons que les métadonnées "température de l'air" soient interopérables pour obtenir ensuite la valeur réelle (les données) dont nous avons besoin - le nombre qui exprime la valeur ne pourra jamais être trouvé ou compris en soi sans les métadonnées associées.

Nous suivrons la même convention que celle adoptée dans les principes directeurs FAIR mentionnés ci-dessus, en utilisant le terme (méta)données lorsque quelque chose s'applique indifféremment aux données et métadonnées.

4. L'interopérabilité des données et des ensembles de données

Il y a des métadonnées qui accompagnent les données individuelles et il y a des métadonnées générales sur l'ensemble de la collection à laquelle les données appartiennent.

Un ensemble de données peut être décrit et rendu interopérable comme un produit en soi, un " travail " (avec ses propres métadonnées) et comme les données contenues dans un ensemble de données (avec leurs propres métadonnées, les variables auxquelles les valeurs correspondent).

[5] We talked about the FAIR principles in lesson 4.1. See https://www.force11.org/fairprinciples

L'interopérabilité peut être réalisée au niveau des métadonnées de l'ensemble de données et au niveau des données : de bonnes métadonnées de l'ensemble de données peuvent permettre de découvrir un ensemble de données ; de bonnes métadonnées sur la structure des données peuvent rendre les données analysables (en outre, l'utilisation de sémantiques telles que des noms de variables normalisés liés à des vocabulaires convenus ou des valeurs contrôlées dans les enregistrements de données réelles permettra de comprendre et d'inter-opérer complètement des données ; voir la leçon 4.4 à ce propos).

Les principes FAIR sont très explicites en ce qui concerne les différentes couches de métadonnées/données : selon leur définition, un objet de données est " un élément de données avec éléments de données + métadonnées " et ils utilisent le terme (méta) données lorsque "le principe est vrai pour les métadonnées ainsi que pour les éléments de données réels recueillis dans l'objet de données ".

De plus, le même ensemble de données peut être distribué sous différentes formes (un fichier Excel, un CSV, un format spécifique à une application) et pour que les machines sachent quelle distribution obtenir, nous avons également besoin de métadonnées sur les distributions (par exemple format, licence et aussi métadonnées de structure).

Il est important de savoir qu'il existe différentes couches de métadonnées: les métadonnées décrivant l'ensemble de données (métadonnées de description/métadonnées de découverte), les métadonnées décrivant les distributions et les métadonnées décrivant les données (métadonnées structurelles décrivant les noms, types et plages des valeurs dans l'ensemble de données, par exemple, les noms et types des colonnes dans un ensemble de données tabulaire) qui peuvent être au niveau de l'ensemble de données ou à celui de l'élément de données.

Tous les grands modèles de métadonnées d'ensembles de données/données prévoient des métadonnées au moins à deux niveaux : ensemble de données et distribution ; la plupart d'entre eux prévoient également des métadonnées au niveau de l'enregistrement ou même (dans le cas de FAIR) de plus petits éléments de données, comme une paire clé-valeur.

Layer	Description	Example
FDP (Data repository)	Information about the FDP as a data repository	PID, title, description, license, owner, API version. etc.
Catalog	Information about the catalog of datasets offered	PID, title, description, publisher, etc.
Dataset	Information about each of the offered datasets	Publisher, issue date, theme, etc.
Distribution	Information about how the dataset is distributed	AccessURL, downloadURL, format, mediaType, etc.
Data record	Information about the actual data, types, identifiers, etc.	Data items types, identifiers, domain, range, etc.

Illustration 1 Couches de métadonnées (Luiz Olavo Bonino et al, 2016,'Fair Data Technology Update')[6]

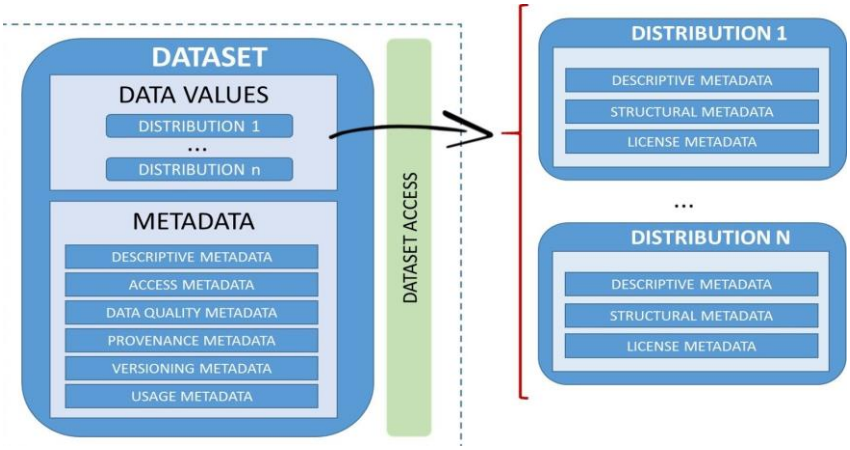

Illustration 2 Données du W3C sur le modèle des meilleures pratiques du Web [7]

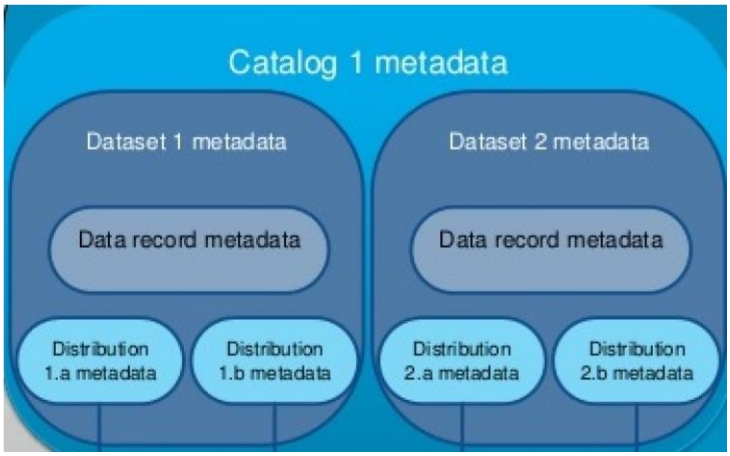

Illustration 3 Modèle FAIRPort (Luiz Olavo Bonino et al, 2016,'Fair Data Technology Update')[7]

[6] https://www.slideshare.net/lolavo/dtl-partners-event-fair-data-tech-overview-day-1
[7] https://www.w3.org/TR/dwbp/
[7] https://www.slideshare.net/lolavo/dtl-partners-event-fair-data-tech-overview-day-1

Traditionnellement (voir les formats de données de longue date comme NetCDF[8] ou HDF5[9], ou considérer le modèle de la norme ISO 19115: Information géographique - Norme des métadonnées, qui couvre des dizaines d'éléments de métadonnées), les métadonnées au niveau de l'ensemble de données comprennent des informations spatiales et temporelles sur la collecte des données, des informations sur les auteurs, certaines conditions environnementales qui peuvent être applicables à la base de données entière, et les noms des variables utilisées dans la séquence des données (qui définit la structure des enregistrements). Les noms des variables dans les enregistrements de données subséquents seraient les métadonnées des données individuelles.

Par exemple, les noms de colonnes d'un ensemble de données tabulaires sont des métadonnées sur les données (les métadonnées sur l'ensemble de données peuvent se trouver dans un fichier de documentation distinct). Dans les structures d'ensemble de données d'observation typiques, telles que NetCDF ou HDF5, les métadonnées sur l'ensemble de données sont au début et comprennent les " métadonnées de découverte " et les " métadonnées d'utilisation ", qui doivent appuyer l'utilisation des données et contiennent normalement toutes les dimensions et noms de variables utilisés dans l'ensemble de données ; puis une séquence tabulaire ou en tableau des enregistrements contenant les données réelles suit.

Cela est également important lorsqu'il s'agit de choisir les normes de données ou " vocabulaires " les plus appropriés (voir leçon 4.4) à utiliser pour ses propres données : il existe des vocabulaires pour décrire des ensembles de données et des vocabulaires pour coder les enregistrements réels de données (et il existe des vocabulaires qui font les deux mais utilisent différentes classes ayant différentes propriétés). Il est essentiel de choisir le vocabulaire/dictionnaire approprié pour les métadonnées de l'ensemble de données et pour les données.

Et cela est important lors du choix d'un outil de gestion des ensembles de données : jusqu'à présent, la plupart des outils ne prennent pas en charge un modèle de métadonnées qui prend en charge différentes couches de métadonnées. Les outils de référentiels de données et les vocabulaires des ensembles de données couvriront, nous l'espérons, cette question sous peu. Pour l'instant, il semble que le seul vocabulaire d'ensemble de données qui couvre cela soit le vocabulaire du W3C Data Cube.

En tout état de cause, en ce qui concerne les métadonnées des ensembles de données, l'absence de métadonnées de " structure de données " (par exemple, les noms des variables et les types, unités de mesure et codes des valeurs) peut sérieusement entraver l'interopérabilité réelle des données : les machines ne peuvent comprendre la signification des données que si les métadonnées relatives aux ensembles de données comprennent des métadonnées structurelles décrivant les dimensions et variables utilisées dans

[8] http://www.unidata.ucar.edu/software/netcdf/docs/
[9] https://support.hdfgroup.org/HDF5/doc/H5.format.html

ces données, à savoir les véritables métadonnées des données, qui donnent leur sens aux données.

En fait, avec une infrastructure distribuée prévue de triples (voir leçon 4.3) représentant les métadonnées à tous les niveaux différents possibles, certains triples concernant un ensemble de données ici, d'autres triples concernant le même ensemble de données ailleurs, et les triples pour les enregistrements contenus potentiellement partout (mais tous liés à l'URI de cet ensemble de données), le concept d'ensemble de données devient beaucoup moins une entité physique ou un fichier maintenant qu'une entité conceptuelle, une collection de documents (qui résident partout, liés à la collection correspondante) avec la même structure.

Les leçons suivantes expliqueront plus en détail comment mettre en œuvre l'interopérabilité structurelle et sémantique.

Lecture complémentaire

Tom Heath and Christian Bizer (2011) *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.
http://linkeddatabook.com/editions/1.0/