Open Data Management in Agriculture and Nutrition

This e-learning course is the result of a collaboration between ***GODAN Action*** *partners, including* ***Wageningen Environmental Research*** *(WUR),* ***AgroKnow****,* ***AidData****,* ***the Food and Agriculture Organization of the United Nations*** *(FAO),* ***the Global Forum on Agricultural Research*** *(GFAR), and* ***the Institute of Development Studies*** *(IDS),* ***the Land Portal****,* ***the Open Data Institute*** *(ODI) and* ***the Technical Centre for Agriculture and Rural Cooperation*** *(CTA).*

GODAN ACTION
Global Open Data for Agriculture & Nutrition

UKaid
from the British people

GODAN Action is a three-year project UK's Department for International Development to enable data users, producers and intermediaries to engage effectively with open data and maximise its potential for impact in the agriculture and nutrition sectors. In particular we work to strengthen capacity, to promote common standards and best practice and to improve how we measure impact. [www.godan.info]

UNIT 4: SHARING OPEN DATA

LESSON 4.2: INTRODUCTION TO DATA INTEROPERABILITY

Aims and learning outcomes

This lesson aims to explain the basics of data interoperability.

After studying this lesson, you should be able to;

- understand the basics of data interoperability
- identify the different types and layers of interoperability of data.

Contents

List of figures

1. Guiding frameworks for data: From open to FAIR

The most used definition of 'interoperability' on the web is: 'the ability of a system or a product to work with other systems or products without special effort on the part of the user'. Wikipedia defines it as 'a characteristic of a product or system, whose interfaces are completely understood, to work with other products or systems, at present or future, in either implementation or access, without any restrictions'.

When we talk about data interoperability, considering that data are serialised in datasets, the definitions above can be applied easily to a dataset as a product.

Data interoperability is the ability of a data set to work with other systems or datasets without special effort on the part

In the proceedings of a conference organised by the CIARD community on data interoperability for agriculture, data interoperability was defined as '**a feature of datasets … whereby data can be easily retrieved, processed, re-used, and re-packaged ("operated") by other systems**.'[1]

There are some definitions that define it as the interoperability 'between two systems', however it is a common view that something is really interoperable (or more interoperable) when as many systems as possible can interoperate it. Even more, we will see that by using certain data formats and applying certain data standards, data can be made 'interoperable by design' without necessarily knowing with which system they will be interoperable: planned interoperability with specific systems means that the data will be 'tightly coupled' with those systems, while maximised interoperability aims at loose coupling with as many systems as possible.

However, there will never be something like universal or perfect interoperability, a way of exposing data that will be suitable for all possible cases. Interoperability is always relative to a system of shared standards and common ways of using data that are in some cases very broad and all-purpose (like Dublin Core or schema.org and a generic search engine use case) and in other cases very specific to scientific or interest communities (like data specifications and visualisations of gene sequences).

[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

Indeed, the definitions above define interoperability as a feature of data(sets) alone, which is correct because they are the object of the interoperation, but the ecosystem of actors and products that have to co-operate for achieving full interoperability is broader: an interesting definition of interoperability that highlights the importance of 'shared expectations' is the one from the Data Interoperability Standards Consortium (DISC): 'Data interoperability addresses the ability of systems and services that create, exchange and consume data to have clear, shared expectations for the contents, context and meaning of that data.'[2]

2. Levels of interoperability

Interoperability can be achieved at different levels. While Wikipedia distinguishes between syntactic interoperability and semantic interoperability, a more granular distinction is made in the classification of types of interoperability by the Healthcare Information and Management Systems Society (HIMSS)[3]:

'***Foundational***' *interoperability allows data exchange from one information technology system to be received by another and does not require the ability for the receiving information technology system to interpret the data.*

This level is mostly about transmission protocols, which we will not analyze as it is not of interest to the data manager, but it also covers some higher-level exchange protocols, mostly working on top of the common HTTP protocol (for instance special REST APIs like OAI-PMH, SPARQL, Linked Data API or content negotiation based on HTTP content-type requests). These may be of interest to data managers because, before being read and understood, data has to be transmitted, and there are different and more convenient ways to do this besides FTP downloads. We will look at this level of interoperability in lesson 4.3.

'***Structural***' *interoperability is an intermediate level that defines the structure or format of data exchange (i.e., the message format standards). Structural interoperability defines the syntax of the data exchange. It ensures that data exchanges between information technology systems can be interpreted at the data field level.*

This is the level where file formats and data formats play the most important role and it is the level where (meta)data become 'machine-readable' and can be parsed by machines: the easier it is for machine to parse the (meta)data format / syntax (XML, Json, CSV…) the more structurally interoperable data are. We will look at this level of interoperability more in depth in lesson 4.3.

---

[2] http://datainteroperability.org/

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

*'**Semantic**' interoperability provides interoperability at the highest level, which is the ability of two or more systems or elements to exchange information and to use the information that has been exchanged. Semantic interoperability takes advantage of both the structuring of the data exchange and the codification of the data including vocabulary so that the receiving information technology systems can interpret the data.*

While with structural interoperability machines understand what the different elements are (and their reciprocal structural relation), with semantic interoperability they also understand the meaning of those elements and can process them with semantics-aware tools and do some advanced reasoning. Data formats are not enough for this: semantic technologies (see lesson 4.4) allow to embed machine-readable semantic elements from agreed vocabularies in (meta)data serialised in most of the existing data formats, although the most suitable formats for this so far are the various flavours of RDF (RDF/XML, Turtle, N-Triples) and JSON-LD. We will look at this level of interoperability in lesson 4.4.

In lesson 4.3 we will see how structural and architectural interoperability can be implemented. In lesson 4.4 we will see how semantic interoperability can be implemented, with more specific examples in lesson 4.4.1.

For more detailed recommendations on how to implement data interoperability, the W3C 'Data on the Web Best Practices'[4] is probably the best reference document. It is heavily based on the 'Linked Data' approach (see lesson 4.3), but many of the recommendations help implement interoperability at different levels, even if not aiming at 100% linked data.

3. Interoperability of data and metadata

Data without metadata cannot be understood by machines. Data normally come with metadata.

The definition of data in Wikipedia is: '**Data are values of qualitative or quantitative variables, belonging to a set of items**.' As such, in the end data are always part of a collection (a set of items) and in the individual item (row, record) in the set, data are the values of some variables, so they will always be encapsulated in some form of key-value pair where the key (the variable) is in the end a metadata element that gives meaning to the data. This key-value pair is what in FAIR[5] terms is called 'a single association between two concepts', which is 'one of the smallest possible 'Data Elements'.

Both parts in the key-value pair can present issues of interoperability, so we deal with issues of interoperability of the metadata (e.g. agreed schemas for

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

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

metadata elements, agreed variable names) and issues of interoperability of the data/value (e.g. agreed controlled lists/ranges from which the value has to be taken, or syntax issues). However, since it is also common to consider controlled values and specification of syntax or unit of measure as 'metadata' (especially because they should be ideally defined by separate metadata elements and not within the value itself), we can also say that interoperability is mostly about metadata. The interoperability of data is achieved through the interoperability of metadata, e.g. we want the metadata 'air temperature' to be interoperable to then get the actual value (the data) that we need – the number that expresses the value will never be findable or understandable *per se* without the associated metadata.

We will follow the same convention adopted in the FAIR guiding principles mentioned above, using the term (meta)data when something applies indifferently to data and metadata.

4. Interoperability of data and interoperability of datasets

There are metadata that accompany the individual collected data and there are general metadata about the whole collection to which the data belong.

A dataset can be described and made interoperable as a product *per se*, a 'work' (with its own metadata) and as the data that are contained in a dataset (with their own metadata, the variables to which the values correspond).

Interoperability can be achieved at the level of the dataset metadata and at the level of the data: good dataset metadata can make a **dataset discoverable**; good metadata about the data structure can make the **data parsable** (in addition, using semantics such as standardised variable names linked to agreed vocabularies or controlled values in the actual data records will make **data understandable** and fully interoperable; see lesson 4.4 for more on this).

The FAIR principles are very explicit regarding the different layers of metadata/data: according to their definition, a Data Object is '**a Data Item with Data elements + Metadata**' and they use the term **(Meta) data** when 'the principle is true for Metadata as well as for the actual, collected Data Elements in the Data Object'.

Furthermore, the same dataset can be distributed in different forms (an Excel file, a CSV, an application-specific format) and in order for machines to know which distribution to get, we also need metadata about the **distributions** (e.g. format, licensing, again structural metadata).

It is important to be aware of the fact that there are **different layers of metadata**: metadata describing the dataset (description metadata/discovery metadata), metadata describing the distributions, and metadata describing the data (structural metadata describing the names, types and ranges of the values in the dataset, e.g. column names and types in a tabular dataset) which can be at the level of the dataset or at the level of the data element.

All major dataset/data metadata models foresee metadata at least at two levels: dataset and distribution; most of them also foresee metadata at the level of either record or even (in the case of FAIR) smaller data elements such as a key-value pair.

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.

Figure 1 Metadata layers (Luiz Olavo Bonino et al., 2016, 'Fair Data Technology Update')[6]

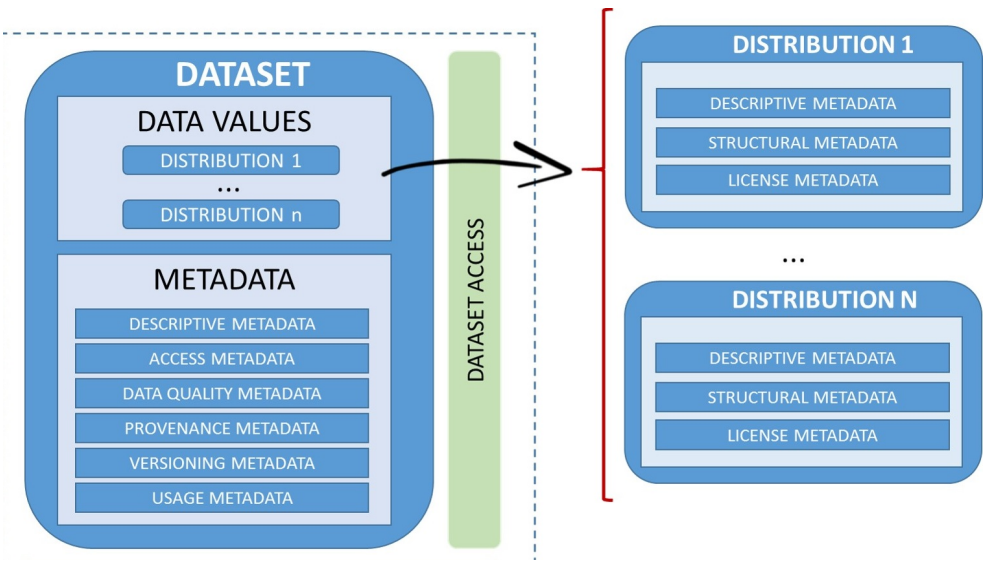

Figure 2 W3C Data on the Web Best Practices model[7]

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

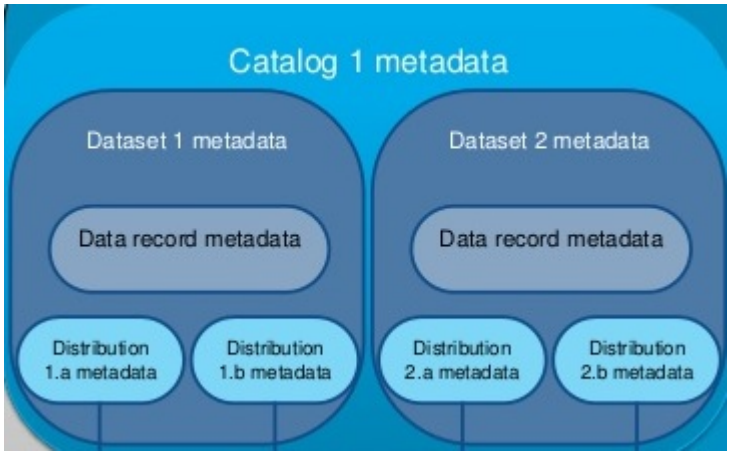

Figure 3 FAIRPort model (from Luiz Olavo Bonino et al., 2016, 'Fair Data Technology Update')[8]

Traditionally (see long-standing data formats like NetCDF[9] or HDF5[10], or consider the model of the 'ISO 19115: Geographic Information - Metadata standard', which covers dozens of metadata elements) metadata at the level of the dataset include spatial and temporal information about the collection of data, information about authorship, some environmental conditions that may apply to the whole database, and the names of the variables used in the data sequence (which define the structure of the data records). The names of the variables in the subsequent data records would be the metadata of the individual data.

For example, column names in a tabular dataset are metadata about the data (metadata about the dataset could be in a separate documentation file). In typical observation dataset structures, such as NetCDF or HDF5, metadata about the dataset is at the beginning and includes 'discovery metadata' and 'use metadata', which have to support the use of the data and normally contain all the dimensions and variable names used in the dataset; then a tabular or array-like sequence of records with the actual data follows.

This is also important when it comes to choosing the most suitable data standards or 'vocabularies' (see lesson 4.4) to use for one's own data: there are vocabularies to describe datasets and there are vocabularies to encode the actual records of data (and there are vocabularies that do both but using different classes with different properties). Choosing the right vocabulary/dictionary for both the dataset metadata and the data is essential.

And it is important when choosing a tool for dataset management: so far most tools do not support a metadata model that caters for different layers of

[8] https://www.slideshare.net/lolavo/dtl-partners-event-fair-data-tech-overview-day-1
[9] http://www.unidata.ucar.edu/software/netcdf/docs/
[10] https://support.hdfgroup.org/HDF5/doc/H5.format.html

metadata. Data repository tools and dataset vocabularies will hopefully cover this soon. At the moment, it seems that the only dataset vocabulary that covers this is the W3C Data Cube vocabulary.

In particular, regarding dataset metadata, the lack of 'data structure' metadata (e.g. the names of the variables and the types, units of measures and codes of the values) can seriously hinder real interoperability of the data: machines can understand the meaning of the data only if the metadata about the dataset include structural metadata that describe the dimensions and variables used in the data, i.e. the real 'metadata' about the data, which give meaning to the data.

Actually, with a foreseen distributed infrastructure of triples (see lesson 4.3) representing metadata at all possible different levels, some triples about one dataset here, some other triples about the same dataset somewhere else, and the triples for the contained records potentially anywhere (but all linked to the URI of the dataset), the concept of dataset becomes much less a physical entity or a file now than a conceptual entity, a collection of records (residing anywhere, linked to the belonging collection) with the same structure.

The following lessons will explain more in detail how to implement structural and semantic interoperability.

Further Readings

- 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/