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 4.4: SEMANTIC INTEROPERABILITY

Aims and learning outcomes

This lesson aims to;

- explain the basics of semantic interoperability
- introduce what 'vocabularies' are
- provide guidance on how to choose the most suitable vocabularies
- provide guidance on how to use vocabularies in the (meta)data.

After studying this lesson, you should be able to;

- understand the basics of semantic interoperability
- choose the vocabularies that better fit their needs
- (guide developers to) use vocabularies in the (meta)data.

Contents

List of figures

1. Introduction

In this lesson we discuss the basic principles and approaches to be used when seeking to improve semantic interoperability of data.

All the data formats seen in lesson 4.3 define only data structures, how to encode fields/variables and values, and the only thing a machine can do is parse the structure and extract variables and values, without knowing how to treat each of them. Variables and values have a meaning which in many cases can be only understood by humans (and in some cases only by humans who speak the same language and know the conventions of the same discipline).

Human beings can interpret data through human-readable semantics that have always been used in (meta)data in different ways: a string to identify the topic of something or the colour of a thing (e.g. in germplasm phenotypical descriptions), codes taken from a code list of authoritative values (e.g. type of soil) or conventional variable names. But as we said, interoperability is all about being understood by computer software: strings can be different in each dataset and in different languages, and even codes without a reference system behind them do not mean anything to computers and do not allow them to make decisions on how to treat the values.

If on the other hand the metadata contained information on the reference system (a 'semantic structure' like a thesaurus or a code list) from which each variable and each value came, and that semantic structure were machine-readable and provided some stable identifiers that computer programs could use as stable values to design their behaviour (e.g. using the values as common search values across different datasets), we would have achieved semantic interoperability.

So on the one hand the metadata have to embed information on the reference semantic structures and point to the exact elements they are using from that structure; on the other hand these semantic structures, like the data, have to be 'serialised' in such a way that machines can read and process them, and use them to interpret the data.

Details on how to publish a semantic structure, or a 'vocabulary', in machine-readable format are beyond the scope of this lesson. In short, for our purposes in this lesson, let us say that such vocabularies are published as datasets, with terms/concepts and their related descriptions, codes and ideally URIs, in a machine-readable format – for the moment let us assume XML or RDF/XML.

2. Semantic structures or 'vocabularies'

Vocabularies are agreed sets of terms, possibly with defined relationships between them. This includes both terms used for description metadata, like

metadata element names, properties, predicates (so terms in **description vocabularies**: metadata schemas, ontologies…) and terms used to categorise, annotate, classify (so terms in **value vocabularies**: thesauri, code lists, classifications, authority lists…).

Nowadays, following the path set by the W3C[1], the term that is most commonly used when referring to resources that define semantic elements is 'vocabularies':

> 'On the Semantic Web, vocabularies define the concepts and relationships (also referred to as "terms") used to describe and represent an area of concern. Vocabularies are used to classify the terms that can be used in a particular application, characterise possible relationships, and define possible constraints on using those terms.'

This includes both terms used for description metadata, like metadata element names, properties, predicates (so terms in description vocabularies: metadata schemas, ontologies…) and terms used to categorise, annotate, classify (so terms in 'value vocabularies', also sometimes called 'Knowledge Organisation Systems' (KOS): thesauri, code lists, classifications, authority lists…).

Other terms that are used for defining these resources are 'semantic resources' or 'semantic structures'. Semantic structures are closely inter-related with 'data standards', but the term data standards goes beyond the area of semantics because it also includes syntactic standards like data formats/data structures (see structural interoperability compared with semantic interoperability in lesson 4.2).

In this lesson, we shall use these terms almost interchangeably, tending to use 'vocabularies' or 'semantic structures' when we talk specifically about semantics and 'data standards' when we refer to a combination of semantics and data formats (some standards define both).

We will also avoid the expression 'semantic standards' and will rather say 'published semantics'. Strictly speaking, a standard should be a specification formally endorsed by different parties that need to use it to improve the compatibility of something. Normally, recognised standards are created by standardisation bodies. Then there are also de facto standards, specifications developed either unilaterally or with a small scope that become widely adopted because of popularity or industry dominance.

However, the area of semantics, especially as applied to agri-food data, is still largely experimental and in many cases standardisation bodies have not gone (or won't go) into disciplinary semantics, so institutions working in specific disciplines have started developing their own semantic

[1] W3C. Vocabularies. https://www.w3.org/standards/semanticweb/ontology

structures/vocabularies and very few of them have reached the status of 'standard'. Therefore, what is relevant for this lesson is semantic structures that are at least: (a) designed/ created with the objective of being broadly endorsed and used; and (b) publicly available and referenceable ('published'), and possibly machine-readable.

2.1. Types of vocabularies

There is no formal classification of types of vocabularies (which in itself could be a useful example of a value vocabulary).

The exercise of creating a vocabulary of vocabulary types has been partially done by the Dublin Core[2] initiative: their 'KOS Types Vocabulary' is limited to specific types of the more general type 'value vocabulary' or KOS. Their list is quite useful to give an idea of the great variety of KOS and of the mixture of features that are combined in their definition:

- **categorisation scheme**: loosely formed grouping scheme
- **classification scheme**: schedule of concepts and pre-coordinated combinations of concepts, arranged by classification
- **dictionary**: reference source containing words usually alphabetically arranged along with information about their forms, pronunciations, functions, etymologies, meanings, and syntactical and idiomatic uses
- **gazetteer**: geospatial dictionary of named and typed places
- **glossary**: collection of textual glosses or of specialised terms with their meanings
- **list**: a limited set of terms arranged as a simple alphabetical list or in some other logically evident way; containing no relationships of any kind
- **name authority list** or authority file: controlled vocabulary for use in naming particular entities consistently
- **ontology**: formal model that allows knowledge to be represented for a specific domain; an ontology describes the types of things that exist (classes), the relationships between them (properties) and the logical ways those classes and properties can be used together (axioms) *[see below a note on how an ontology can be seen as a KOS but also as a description vocabulary, an extended schema]*
- **semantic network**: set of terms representing concepts, modeled as the nodes in a network of variable relationship types
- **subject heading scheme**: structured vocabulary comprising terms available for subject indexing, plus rules for combining them into pre-coordinated strings of terms where necessary
- **synonym ring**: set of synonymous or almost synonymous terms, any of which can be used to refer to a particular concept
- **taxonomy**: scheme of categories and subcategories that can be used to sort and otherwise organise items of knowledge or information

[2] http://wiki.dublincore.org/index.php/NKOS_Vocabularies#KOS_Types_Vocabulary

- **terminology**: set of designations belonging to one special language
- **thesaurus**: controlled and structured vocabulary in which concepts are represented by terms, organised so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms.

There is no authoritative list of description/modelling vocabularies, but the most commonly used such types are:

- **schema** (or metadata element set): any set of metadata elements, like XML schemas, RDF schemas or less formalised set of descriptors
- **application profile**: a schema which consist of metadata elements drawn from one or more namespaces, combined together by implementors, and optimised for a particular local application
- **messaging standard**: standards which describe how to format syntactically (and sometimes semantically) a message usually describing some event- or time-related information; messages are triggered by an event and transmitted in some way
- **ontology**, seen as a more powerful form of schema.

As can be seen from the two lists above, ontologies are a special case: 'In computer science and information science, an **ontology** is a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse.'[3] As such, it can be used for multiple purposes: it can be used as a description vocabulary, using the relations or even the classes defined by the ontology as metadata elements/properties describing your data (e.g. 'extreme temperature resistance' or 'frost resistance' in the Wheat Trait Ontology[4]), or as a value vocabulary, using classes or entities as terms for controlled values (e.g. wheat illnesses like *Puccinia striiformis* from the Wheat Trait Ontology, or countries from the FAO Geopolitical Ontology[5]).

Sometimes the boundaries between a schema and an ontology are blurred, but perhaps what can be considered typical of an ontology is the 'functional' more than descriptive design: classes, properties and especially relationships are designed as a model that is 'actionable' and can be used for reasoning. However, the tendency nowadays is to use just the word 'vocabulary' and not delve too much into the definition of the different types[6].

Lesson 4.4.1 will provide more specific examples on how to identify the most suitable vocabularies.

---

[3] From Wikipedia: https://en.wikipedia.org/wiki/Ontology_(information_science)

[4] http://vest.agrisemantics.org/content/wheat-trait-ontology

[5] http://vest.agrisemantics.org/content/geopolitical-ontology

[6] See on the W3C page on ontologies: https://www.w3.org/standards/semanticweb/ontology

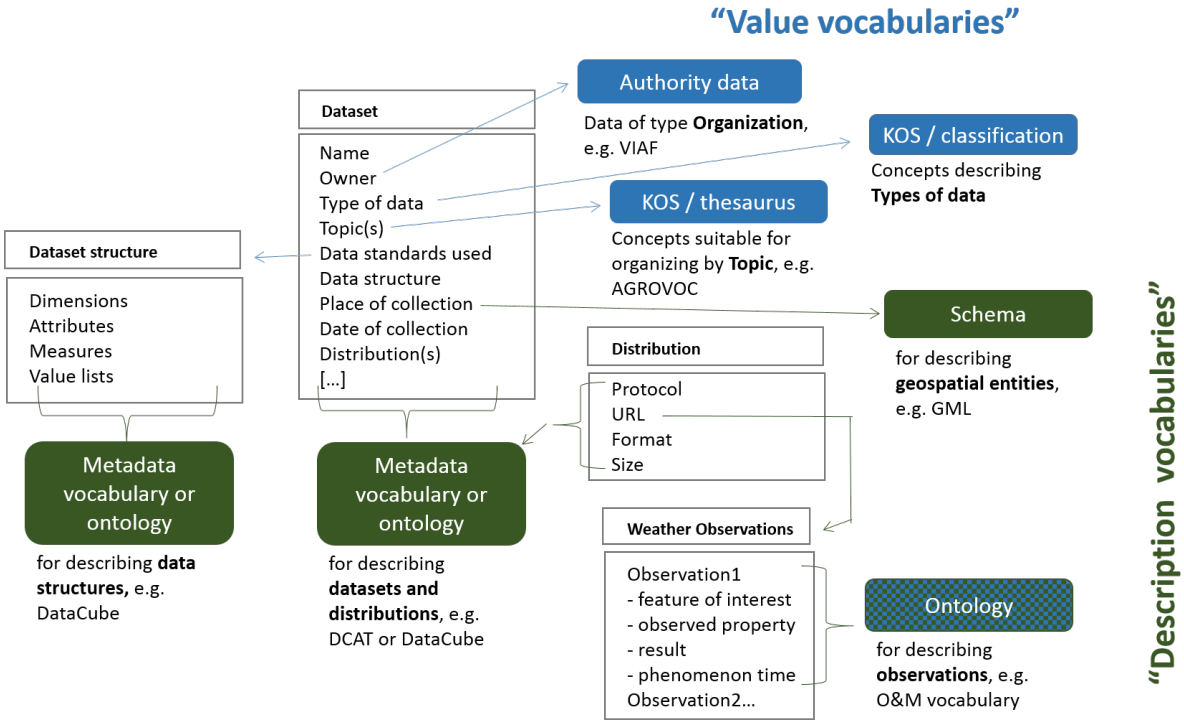

Figure 1 Example of use of different types of vocabularies to add semantics to (meta)data

2.2. How to identify the most suitable published semantic structures

The most useful source of information are of course catalogues that are dedicated to the agri-food domain, but general catalogues searchable by domain can also be of help. An important distinction is to be drawn between catalogues/directories/registries on the one hand and repositories on the other: registries are conceived as metadata catalogues, providing descriptions and categorisation of vocabularies and linking to the original website and original serialisation of the standard, while repositories host the full content of the vocabulary, so that the terms themselves can be browsed. Below is an overview of some existing catalogues/repositories of data standards and vocabularies.

Agri-food domain

- GODAN Action Map of data standards - http://vest.agrisemantics.org
  A catalogue of data standards of different types and formats for the agri-food domain, categorised according to sub-domain, types of data, format and other criteria.
- AgroPortal - http://agroportal.lirmm.fr/
  A repository of ontologies and value vocabularies, specialised in agronomy and food.
- Planteome - http://browser.planteome.org/amigo
  A repository of ontologies for plant biology.

General

- FAIRsharing - https://fairsharing.org/
  Evolved from the Biosharing directory of standards for life sciences, it is now a general directory of data standards of different types. It has a good tagging system but the coverage of agri-food standards is still poor.
- Linked Open Vocabularies (LOV) - https://lov.okfn.org/dataset/lov
  Directory of RDF vocabularies (see lesson 4.3 for a description of RDF) spanning across all disciplines. It is not organised by domain or discipline and vocabularies can only be browsed through a small number of free tags.
- The Basel Register of Thesauri, Ontologies and Classifications (BARTOC) - http://bartoc.org/
  BARTOC includes all types of KOS in any format, across all subject areas. The categorisation of vocabularies is quite generic (food and agriculture would fall partly under Pure Science and partly under Technology without further sub-categorisations).

3. Embedding semantics in the (meta)data

3.1. Using a schema for your data

If you identify a metadata vocabulary/schema/ontology that has the classes and properties you need to describe your data, you can reuse it to model and represent your data. An important thing to note about using an existing published schema is that by doing this alone your data will already be more semantically interoperable, because instead of local arbitrary metadata element names which are meaningless for a computer, you will use element names from a published vocabulary, and software tools that are aware of that vocabulary will be able to do something with it, e.g. match the values with values from other datasets that use the same schema.

The adopted schema becomes the 'language' of your data structure. For instance, instead of using a custom XML structure with local arbitrary element names, by adopting an existing XML schema you declare that you are using elements from it and for each element you will use the element name of the selected schema instead of a local one, with a prefix that indicates from which schema the element comes.

The example below (Figure 4.4.2), from the CSML user manual[7], declares in the root XML element that it uses elements from the GML (http://www.opengis.net/gml) and O&M (http://www.opengis.net/om) schemas from the Open Geospatial Consortium (OGC) as prescribed in the Climate Science Modelling Language (CSML) schema and assigns prefixes to

[7] http://http//proj.badc.rl.ac.uk/csml/browser/Documentation/trunk/CSMLUsersManual.pdf

these schemas, and then the schema prefixes are used in the document to indicate that the metadata element name used comes from the corresponding schema (where its meaning is defined). For a computer, reading <om:Phenomenon> and knowing that 'om:' stands for the O&M schema located at http://www.opengis.net/om, instead of just reading <phenomenon>, makes a big difference: as a minimum, it knows that om:Phenomenon in any dataset means the same thing and has the same nested elements, but specialised software tools aware of the meaning of that term in the schema can build advanced functionalities, visualisations, modelling, etc.

```
<?xml version="1.0" encoding="UTF-8"?>
<Dataset xmlns="http://ndg.nerc.ac.uk/csml" xmlns:gml="http://www.opengis.net/gml"
xmlns:om="http://www.opengis.net/om" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://ndg.nerc.ac.uk/csml XMLSchemas/CSMLAppSchema.xsd" gml:id="TestDataset">
 <gml:metaDataProperty xlink:href="http://ndg.nerc.ac.uk/Metadata/TestDataset"/>
 <gml:description>Test CSML Dataset.</gml:description>
 <UnitDefinitions gml:id="UnitDefs">
   <gml:name codeSpace="http://ndg.nerc.ac.uk">TestDatasetUnitDefs</gml:name>
   <gml:definitionMember>
     <gml:UnitDefinition gml:id="psu">
       <gml:description>Conventional practical salinity units.</gml:description>
       <gml:name>practical salinity units</gml:name>
       <gml:quantityType>sea water salinity</gml:quantityType>
       <gml:catalogSymbol codeSpace="http://ndg.nerc.ac.uk/units">psu</gml:catalogSymbol>
     </gml:UnitDefinition>
   </gml:definitionMember>
 </UnitDefinitions>
 <PhenomenonDefinitions gml:id="PhenDefs">
   <gml:name>TestDatasetPhenDefs</gml:name>
   <gml:definitionMember>
     <om:Phenomenon gml:id="rainfall">
       <gml:description>Liquid precipitation measured with raingauge.</gml:description>
       <gml:name codeSpace="http://ndg.nerc.ac.uk">rainfall</gml:name>
     </om:Phenomenon>
   </gml:definitionMember>
 </PhenomenonDefinitions>
```

Figure 2 Example of XML using very well known published schemas

The technique of declaring the external schema and using its element names with a prefix is the same also in the case of RDF documents. The higher level of interoperability of using an RDF vocabulary (and therefore an RDF dataset), especially if following the Linked Data pattern (see lesson 4.2) is due to two facts: (a) metadata elements (classes or properties in RDF) are identified by URIs and those URIs are dereferenced to web pages that contain machine-readable information about the class/property, so a computer program can follow the link and read more data, and if there are other links to other external entities it can continue following them and getting more and more data; and (b) as noted previously, an RDF file has fewer potential ambiguities than a non-RDF XML file and is interpreted more reliably. You can use an ontology in the same way.

A similar use of existing schemas can be applied if using the CSV or JSON format, with some limitations. A simple flat schema can be applied to a CSV file by using the element/property names as the column names. This might be

exploited by some software product that is aware of the vocabulary and can therefore do something more with the values than it would be able to do with just custom strings. XML and RDF schemas are also normally easily represented as a JSON structure, using the prefixed metadata element/property names as JSON labels. An XML or RDF schema can be codified as a JSON schema and JSON files based on the schema can be created.

There is also a variant of JSON called JSON-LD (JSON for Linking Data), based on a specification that provides a method of encoding Linked Data and therefore RDF using JSON.

3.2. Using value vocabularies for annotating/categorising your data

A slightly different case is when you want to use values from an existing vocabulary as values of some of the metadata, for instance if you want to use the AGROVOC term for *Oryza sativa* or the term identifying a country from the FAO Geopolitical Ontology.

As a minimum, once some suitable vocabulary has been identified, if the vocabulary does not use URIs and/or the URIs cannot be used in the dataset, at least the literal values of the terms can be used in the dataset. Systems that are aware of the vocabulary and can match the literal values against the URI can already do something with this. Ideally, you should use the URI of the term you want to refer to.

Again, this can be done in different ways depending on the data format you're using. In non-RDF XML, in CSV or in JSON you can use the URI as the value of the element/column/label (e.g. in XML you can use the URI of the Geopolitical Ontology country as the value of the dc:spatial element, perhaps specifying the scheme='URI' attribute to make it clear it's a URI). However: (a) parsers for these formats do not usually follow the URI and get additional data, like the label of the term, so it may be useful to have a separate element for the label of the term for human-readable displays; (b) even if the URI contains the vocabulary base URI, parsers for these formats would not normally fetch additional information about it, so additional metadata about the vocabulary from which the term was taken can be useful.

Ideally, semantic interoperability is fully achieved using an RDF-enabled serialisation format (XML/RDF, Turtle, N3, JSON-LD) and using the URI of the selected term as the object of a triple. The advantage of using RDF is that RDF parsers and crawlers would normally look up the label from the URI address so you may not need to also add a triple where the object is the string.

The example below uses URIs of terms from the AGROVOC thesaurus (for describing the theme of the dataset: weather forecasting) and from the FAO Geopolitical ontology (for identifying the country: Argentina)

```
<dcat:Dataset rdf:about="#item-d277fa36-beb2-48fc-afdc-e9d887d8329f">
    <dct:title>Example dataset #1</dct:title>
    <dct:description>Example description</dct:description>
    <dcat:theme rdf:resource="http://aims.fao.org/aos/agrovoc/c_8340"></dcat:theme>
    <dct:issued>2014-01-01</dct:issued>
    <dct:spatial rdf:resource="http://www.fao.org/countryprofiles/geoinfo/geopolitical/resource/Argentina"></dct:spatial>
    <dct:temporal>2014</dct:temporal>
    <dct:source rdf:resource="#item-0ce19e3c-80e5-4e77-a369-ff1a93e37281"></dct:source>
    <dcat:distribution rdf:resource="#item-19832665-75c8-4706-abcf-2f5f2009446a"></dcat:distribution>
</dcat:Dataset>
```

Figure 3 Example of RDF triples with URIs of terms from other vocabularies

Even if you do not find an ideal vocabulary that meets your needs and resort to using your own terms, you can link your local term to some similar or broader term in existing vocabularies. This technique is called 'mapping' local values to external values and, besides being the cornerstone of the Linked Data architecture, it multiplies the possibilities of giving meaning to local data, as computer programs can crawl all the mapped URIs and collect more metadata about the term from other vocabularies.

The example below is similar to the previous one, but for the theme of the dataset a more precise local term is used ('weather forecasts'), with a local URI, mapped to the external more authoritative URI of the 'weather forecasting' term in the AGROVOC thesaurus.

```
<dcat:Dataset rdf:about="#item-d277fa36-beb2-48fc-afdc-e9d887d8329f">
    <dct:title>Example dataset #1</dct:title>
    <dct:description>Example description</dct:description>
    <dcat:theme rdf:resource="#local-6789"></dcat:theme>
    <dct:temporal>2014</dct:temporal>
</dcat:Dataset>
<skos:Concept rdf:about="#local-6789">
    <skos:preferredLabel>weather forecasts</skos:preferredLabel>
    <skos:closeMatch rdf:resource="http://aims.fao.org/aos/agrovoc/c_8340"></skos:closeMatch>
</skos:Concept>
```

Figure 4 Example of RDF triples mapping a local term to a term in an external vocabulary

In lesson 4.4.1 you will find more specific examples on the selection and use of data standards and semantics for interoperability.