A Map of Agri-food Data Standards

Valeria Pesce, Jeni Tennison, Lisette Mey, Clement Jonquet, Anne Toulet, Sophie Aubin, Zervas Panagiotis

GODAN ACTION LEARNING PAPER

A map of agri-food data standards

Valeria Pesce
Global Forum on Agricultural Research (GFAR)

Jeni Tennison
Open Data Institute (ODI)

Lisette Mey
Land Portal Foundation

Clement Jonquet
Center for Biomedical Informatics Research (BMIR) Stanford University School of Medicine

Anne Toulet
Laboratory of Informatics, Robotics and Microelectronics of Montpellier (LIRMM)

Sophie Aubin
French National Institute for Agricultural Research (INRA)

Panagiotis Zervas
Agroknow

25 November 2016

GODAN ACTION
Global Open Data for Agriculture & Nutrition

Executive summary

The VEST/AgroPortal online map of standards is a deliverable of the GODAN Action project.

GODAN Action supports data users, producers and intermediaries to effectively engage 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.

The project is part of the GODAN programme that promotes the proactive sharing of open data to make information about agriculture and nutrition available, accessible and usable. The project has been initially funded for 3.5 years by the UK Department for International Development.

Specifically the map of standards supports the GODAN Action task of mapping relevant standards and identifying areas where the lack of standards is inhibiting. The objectives of this task are:

- To map currently available open and proprietary standards in use for the exchange of key data on agriculture and nutrition;
- To identify where a lack of standards is inhibiting the effective use of agricultural and nutritional data and the best methods for promoting open data standards.

The map of standards fulfils the first objective and is also designed to support the gap analysis exercise described in the second objective. The full report on the gap analysis is also available (Pesce, Kayumbi, Tennison, Mey and Zervas 2016).

Paraphrasing what the Dublin Core Metadata Initiative says about their DCMI Registry[1], the main purpose of a global map of data standards in a specific field is to promote the discovery and re-use of vocabularies and their properties, classes and controlled values. The re-use of existing vocabularies or classes/properties therein is essential to standardisation, and promotes greater interoperability between vocabularies and data sets. The discovery of existing vocabularies is an essential, and prerequisite, step in this process. This map promotes the wider adoption, standardisation and interoperability of vocabularies by facilitating their discovery and re-use across diverse communities of practice.

In addition, it provides a useful overview of what exists and helps to identify overlaps, duplication, gaps and limits to adoption, hopefully encouraging practitioners not to duplicate efforts and to collaborate both to develop and use common standards.

The first version of the map is available at http://vest.agrisemantics.org

This report is an accompanying document to the map of standards that describes the approach to the implementation of the map, the categorisation of standards, and gives an overview of the initial content. It also gives details on the coverage and organisation of the map, and a summary of how we conducted a call to action to experts to contribute to its improvement.

Our approach to creating the map of standards was based on the principles of building on what already exists, providing for sustainability and defining standards in terms of use.

To decide how to shape the coverage of the map, we first identified the needs of the users:

- data managers (looking for the best way to standardise their data and make them interoperable)
- researchers (looking for the best way to standardise their data)
- developers (looking for technical information on data standards to inform how they consume data in their applications)
- other data consumers, such as journalists (looking for documentation on the data standards used in

1 http://dcmi.kc.tsukuba.ac.jp/dcregistry/

Scope and definitions

the data sets they are trying to understand)

As the map is designed to support work on open data, we limited the scope to data standards, and chose not to include other types of standards. It is also important to clarify that while data standards may include the notion of data formats, we do not include mere data formats in the map (such as CSV, MDB, XML). This is for two reasons. Firstly, formats are domain-agnostic and are already documented in general registries (Internet Assigned Numbers Authority[2], World Wide Web Consortium (W3C))[3]. Secondly, data formats do not support the semantic interoperability of data per se.

Within the scope of the map, by 'data standards' we mean 'vocabularies' without addition or qualification. This is the broad sense in which vocabularies are defined by W3C (Isaac et al. 2011), which includes metadata element sets (schemas or definitions of description models, more general 'description vocabularies') and value vocabularies (sets of controlled values): see section 2.3.1 for more on this.

This is why we use the terms 'data standards' and 'vocabularies' interchangeably, and they range from description/modelling standards (XML schemas, RDFS schemas, ontologies, application profiles, even UML models) to knowledge organisation systems of different types (classifications, thesauri, even certain types of International Organization of Standardization (ISO) standards as controlled lists of values).

The domains of food and agriculture span across several disciplines (including plant sciences, farming systems, natural resources management, forestry, all disciplines involved in the food supply chain), and are also closely interlinked with neighbouring disciplines (such as climate, environment, geospatial, biology).

We decided to include standards covering all of these disciplines. In the standards map, we also included generic standards that are universally used to describe any resource (such as Dublin Core) or to describe or provide values for generic properties (geographic, ownership, provenance), as these are useful in any domain.

However, we may decide in the future to refer to other existing vocabulary registries, like Linked Open Vocabularies (LOV)[4] or the Basel Register of Thesauri, Ontologies and Classifications (BARTOC) directory[5], for these non domain-specific vocabularies.

2 http://www.iana.org/
3 https://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/
4 https://lov.okfn.org/dataset/lov/
5 http://bartoc.org/

Contents

2 Approach

Our approach to building the map of standards was based on the following principles:

Building on what exists

This is also in line with UK Digital Design principle 'Do less: if someone else is doing it – link to it'.
There were two existing portals collecting information on vocabularies for food and agriculture: the FAO VEST Registry (http://aims.fao.org/vest-registry) and the AgroPortal (http://agroportal.lirmm.fr) and we built on those. (See 2.1.)

Providing for sustainability

We are relying on the synergies between different partners and initiatives as a basis for the sustainability of the platform. (See 2.2.)

We want this map to be maintained well beyond the end of the project, as a common asset of the community working with agricultural and nutrition data.

Looking at standards in terms of use

As indicated in the description of GODAN Action Focal Areas, work on standards (focal area 1) is led by the needs of those who have to use the standards (in their data sets, their data management tools, their information systems). This guided us in defining the coverage of the database (which 'standards'?) and the elements (core metadata, 'assessment' and gap analysis criteria) that we needed to analyse for each standard. (See 2.3.)

2.1 Building on what exists

We did some research on existing vocabulary registries, both to see if we could link to them even if not specific to our domain and to see how they were structured and learn from them.

The best known directory of vocabularies is the Linked Open Vocabularies (LOV) directory[7], with 581 vocabularies spanning across all disciplines. While this is the most promising place to register open vocabularies, it is not organised by domain or discipline, and vocabularies can only be browsed through a small number of free tags. In addition, it only includes Linked Open Vocabularies, so only vocabularies formalised in some Resource Description Framework (RDF) 'dialects'. So we could use it to search for vocabularies that we may want to include and to learn from their metadata model. However, the information there was not organised in a way that could help us in our aim of providing a good overview of the status and current availability/accessibility of data standards of all types for food and agriculture, especially to identify gaps, overlaps, duplication and so on.

Another interesting directory is the Basel Register of Thesauri, Ontologies and Classifications (BARTOC).[8] Its main goal is to list as many Knowledge Organization Systems as possible in one place in order to achieve greater visibility, highlight their features, make them searchable and comparable, and foster knowledge sharing. BARTOC includes any kind of KOS from any subject area, in any language, any publication format, and any form of accessibility.

The scope of this registry was closer to what we wanted to do but again the categorisation of vocabularies was quite generic (food and agriculture would fall partly under Pure Science and partly under Technology without further sub-categorisations) and the metadata was too limited for our purposes.

What remains to be decided – and will be one of the objects of discussion in year 1 – is if and how to implement some form of exchange between our domain-specific map of data standards and these two broader registries.

On the one hand, we cannot import from these registries because of a) lack of a granular categorisation of vocabularies that would allow us to filter vocabularies relevant to our domain; and b) insufficient metadata for our purposes.

On the other hand, we could liaise with their managers to see if they may want to harvest metadata from our map, in order to spare vocabulary managers the double effort of registering their vocabularies on our map and on the two other broader ones. However, many of the vocabularies in our map are already in the other registries, even if with limited metadata.

More relevant to us was what had already been done to map and/or put together the data standards used in the area of food and agriculture. We knew that there were two main efforts on which we could build:

6 http://www.godan.info/godan-action/about
7 https://lov.okfn.org/dataset/lov
8 http://bartoc.org/

- **The VEST Registry**[9]
  This registry, managed by FAO in the Agricultural Information Management Standards (AIMS) website, from the outset covered all types of vocabularies in any format, ranging from description metadata sets (XML schemas, RDFS schemas, application profiles) to KOS of different types (classifications, thesauri, ontologies).

  Also the domain coverage was quite broad: besides vocabularies used specifically for food and agriculture data, the directory included generic vocabularies used for any type of data (from bibliographic resources to time series to soil data, germplasm data etc.) and vocabularies used in neighbouring disciplines (climate, environment, geospatial).
  The registry was conceived as a metadata catalogue, providing descriptions and categorisation of standards and linking to the original website and original serialisation of the standard.

- **The AgroPortal**[10]
  The AgroPortal was designed as a repository of ontologies used in the area of food and agriculture. As a repository, the portal stores the actual RDF content of the ontologies and offers advanced functionalities for linking and using the vocabularies. It enables ontology search, versioning, visualisation, commenting, recommendation, semantic annotation, as well as storing and exploiting of ontology alignments.

  Therefore, the scope of the AgroPortal is limited to RDF vocabularies, including strictly OWL or OBO ontologies as well as SKOS concept schemes.

These two platforms were not communicating with each other and few people were aware of both.

Since the two portals had different audiences and different objectives, merging them was not an option. Instead, we decided to synchronise the essential metadata so that the resulting global map included all the ontologies uploaded to the AgroPortal.

The result of this work is the VEST/AgroPortal global map of data standards used in food and agriculture: http://vest.agrisemantics.org.

This VEST/AgroPortal map is the continuation of the VEST Registry started on the FAO AIMS website and it includes metadata from the AgroPortal ontology repository managed by the University of Montpellier and Stanford University.
In order to maintain some synchronisation between the two platforms and also to highlight and exploit the differences, we implemented some interaction:

- The VEST/AgroPortal map is the landing page for those looking for all the standards. For vocabularies that also have a corresponding record in the AgroPortal, a link to the AgroPortal page with the full content is provided. New metadata from the AgroPortal are imported at regular intervals.

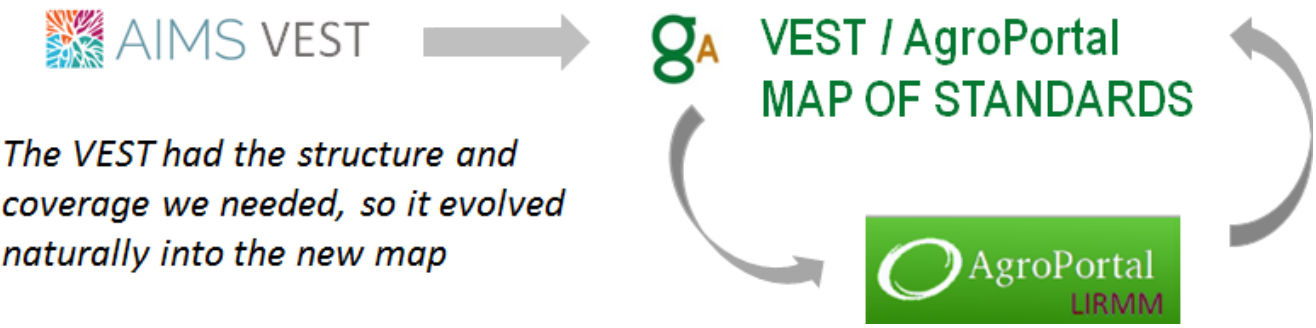
- If someone navigating this map wants to register or edit the description of a vocabulary:

  - the metadata are added/edited in the new VEST map.
  - if the vocabulary is already in the AgroPortal repository, the system will also provide a link to edit the description or upload a new version to the AgroPortal.
  - if the vocabulary is not yet in the AgroPortal and if some of the metadata indicate a potential RDF vocabulary (e.g. format = RDF or OWL; SPARQL endpoint not empty, vocabulary type = ontology), the system will display an invitation to also upload the vocabulary to the AgroPortal, and will point the user to the relevant page on the AgroPortal website.

In conclusion, we are building on two existing portals - the FAO VEST Registry and the AgroPortal – and we're making sure that the specificity of each is preserved while ensuring that they communicate with each other and don't become silos.

9 http://aims.fao.org/vest-registry
10 http://AgroPortal.lirmm.fr

Figure 1 GODAN Action map of standards building on VEST Registry and AgroPortal



2.2 Providing for sustainability

The partnership between GODAN and the Research Data Alliance (RDA), especially through the Interest Group on Agricultural Data Interoperability (IGAD), allowed us to flag the map of standards in discussions leading to the creation of the 'AgriSemantics' Working Group. The map will therefore be one of the resources used in that group and one of the components of the semantic infrastructure under discussion in the group.

We also leveraged our connection with the Institut National de la Recherche Agronomique (INRA) to involve the actors behind the AgroPortal (see below): the University of Montpellier and Stanford University. The map of standards is now part of a common strategy among these partners and will hopefully be included in the roadmap for a common infrastructure that will be designed by the eROSA H2020 project.

Besides, the experts participating in the FAO AIMS and Coherence in Information for Agricultural Research for Development (CIARD) communities were already aware of the VEST Registry, now evolved into the new map of standards. They will continue using it as the reference place for data standards for food and agriculture.

2.3 Looking at standards in terms of use

The types of users who we see as benefiting from the map of standards are:

- data managers looking for the best way to standardise their data and make them interoperable;
- researchers looking for the best way to standardise their data;
- developers looking for technical information on data standards in order to consume data using those standards in their applications;
- other data consumers, such as journalists, looking for documentation on the data standards used in the data sets that they are trying to understand.

Consideration of the possible needs of these users guided us in defining the coverage of the database (which 'data standards'?) and the elements (core metadata, 'assessment' and gap analysis criteria) that we needed to analyse for each standard.

2.3.1 Coverage: types of data standards

In the project proposal, we gave an initial overview of what we thought the coverage of the 'map of standards' would be. For us, data standards initially included:

- metadata standards, especially semantic metadata vocabularies like RDFs vocabularies.
- ontologies (which may range from something similar to a simple metadata model to a complex ontological model for reasoning or decision-making).
  'value vocabularies' or Knowledge Organisation Systems (KOS) that formalise classifications, thesauri, subject lists (normally finite and rarely extended lists of values in a domain).
- name authorities, somehow similar to value vocabularies but more functional to the use of unambiguous names and the disambiguation of synonyms and variants, especially in non-finite lists (authors, organisations, some organisms).
- and everything in between: e.g. there are ontologies that are also used as name authorities (like the Crop Ontology or the Gene Ontology or the Geopolitical Ontology).

The list above clearly identifies data standards as different forms of 'vocabulary', and indeed in the map of standards and in the documentation we use the two terms interchangeably.

In the inception phase, we fine-tuned this list based on our initial survey on the quality of the map of standards. Experts gave us feedback on additional types of data standard that should be also included because the categories of users we identified would be looking for them. Based on this feedback, we added types of data standards such as ISO data standards, UML models and messaging standards.

These types of data standards are not normally associated with the concept of 'vocabulary', but can be considered as vocabularies in a broad sense, as they are all used to represent, describe, categorise or provide standard values for some type of data.

So, to clarify the coverage, we can refer to these two types of vocabularies as defined by the W3C (Isaac et al. 2011):

- **Value vocabularies:** A value vocabulary defines resources (such as instances of topics, art styles, or authors) that are used as values for elements in metadata records. Examples include: thesauri, code lists, term lists, classification schemes, subject heading lists, taxonomies, authority files, digital gazetteers, concept schemes, and other types of knowledge organisation systems.

- **Metadata element sets (or element sets):** A metadata element set defines classes and attributes used to describe entities of interest. In Linked Data terminology, such element sets are generally made concrete through RDF Schemas or OWL Web Ontology Language ontologies, with the term 'RDF vocabulary' often being used as an umbrella for these. (Outside the linked data domain, any set of entity properties or entity-relationship model qualifies.)

A better term to encompass element sets, schemas and ontologies can be 'description vocabularies', though sometimes in literature they are simply called 'vocabularies', while value vocabularies are often called by their specific type (thesaurus, classification, taxonomy and so on) or sometimes collectively as Knowledge Organisation Systems (KOS).

The reason why our scope is so broad in terms of types of vocabularies compared to that of the AgroPortal, or even to that of the LOV registry, is that our map is not a technical platform (where ,for example, the RDF content of the ontologies is stored and various analyses and manipulations can be made). Rather it is an overview of the status and current availability/accessibility of standards for food and agriculture, especially in order to identify gaps, overlaps, duplication and so on.

There are several types of standards that are useful for sharing agri-food data besides life sciences ontologies and models: there are important taxonomic classifications, international descriptors standards, industry data exchange standards or UML models like INSPIRE that have never been converted into RDF.

See section 2.3.4.2 of this report for all the types of data standards, with definitions.

2.3.2 Coverage: domain

Since the domains of food and agriculture span across several disciplines (including plant sciences, farming systems, natural resources management, fisheries, all disciplines involved in the food supply chain) but are also closely interlinked with neighbouring disciplines (such as climate, environment, geospatial, biology), we included standards covering all of these disciplines. However, we may decide in the future to refer to other existing vocabulary registries, like LOV or BARTOC, for these non domain-specific vocabularies.

See section 2.3.4.1 of this report for all domains (plant sciences, natural resources, law, nutrition etc.) and types of data (agronomic data, soil data, plant phenotypic data, market data etc.).

2.3.3 Metadata model

The core properties used in the map are drawn from existing standards: though not designed specifically in the project around users' needs, all the properties provided by major standards cover all the essential information that different types of users may need about the standard. Besides those essential properties, we added properties that will help to evaluate how suitable the featured standards are for users.

The map of standards uses properties from the most widely used vocabularies to describe vocabularies:

Table 1 Vocabularies used to describe vocabularies and related entities

Name	Prefix	URL / more information
Dublin Core	dc	http://purl.org/dc/elements/1.1/
Dublin Core Terms	dct	http://purl.org/dc/terms/
Simple Knowledge Organisation System	skos	http://www.w3.org/2004/02/skos/core#
Ontology Metadata Vocabulary	omv	http://omv.ontoware.org/2005/05/ontology#
Friend of a Friend	foaf	http://xmlns.com/foaf/0.1/
Data Catalog Vocabulary	dcat	http://www.w3.org/ns/dcat#
schema.org	schema	http://schema.org/
Vocabulary of Interlinked Datasets	void	http://rdfs.org/ns/void#
SPARQL Service Description	sd	http://www.w3.org/ns/sparql-service-description#
Vocabulary for Annotating Vocabulary Descriptions	vann	http://purl.org/vocab/vann/
Vocabulary of a Friend	voaf	http://purl.org/vocommons/voaf#
Descriptive Ontology of Ontology Relations	door	http://kannel.kmi.open.ac.uk/ontology#

Below is the list of core properties as mapped between the VEST Registry and the AgroPortal.

Table 2 Main properties of a vocabulary and corresponding RDF properties

VEST field	VEST RDF property	AgroPortal (additional) RDF property
Identifier	dct:identifier	dct:identifier
Name	dct:title skos:ConceptScheme > rdfs:label	omv:name
Alternative name	dct:alternative	omv:acronym
Acronym		
Logo		foaf:logo
Description	dct:description	omv:description
Languages	dct:language	omv:naturalLanguage
Creators	dct:creator	omv:hasCreator
Contact e-mail	dcat:contactPoint	dcat:contactPoint
URL	dcat:landingPage	schema:url
Contributor	dct:contributor	omv:hasContributor
Publisher	dct:publisher	dct:publisher
Domain	dct:subject	omv:hasDomain
Type	dct:type	omv:hasFormalityLevel
Format	dcat:mediaType	omv:hasOntologySyntax
License/Copyright	dct:license	omv:hasLicense
Description	dct:description	omv:description
SPARQL Endpoint	void:sparqlEndpoint	sd:endpoint
Dump	void:dataDump	void:dataDump void:csvDump
Namespace	void:uriSpace vann:preferredNamespaceUri	vann:preferredNamespaceUri
Aligned to	voaf:hasEquivalencesWith	door:isAlignedTo
Uses (extends, generalizes, specializes)	voaf:reliesOn voaf:extends voaf:specializes voaf:generalizes	omv:useImports door:explanationEvolution voaf:generalizes
Used by	voaf:usedBy	voaf:usedBy
Overlaps with	voaf:similar	door:similarTo door:ontologyRelatedTo
Link to machine-readable description	dcat:Distribution > dcat:accessURL dct:conformsTo	omv:conformsToKnowledgeRepresentationParadigm
Creation date	dct:created	dct:created
Modification date	dcat:modified	omv:modificationDate
Example Resource	void:exampleResource	omv:keyClasses
Number of triples	void:triples	omv:numberOfAxioms
Number of entities	void:entities	void:entities
Classes	void:classes	omv:numberOfClasses
Properties	void:properties	omv:numberOfProperties

In addition to these core metadata, new metadata to evaluate the standards were included.

The evaluation metadata were based on two existing frameworks (the assessment process used by the UK Government's Open Standards Board and the Open Data Institute Open Data Certificates criteria). They are grouped under three main categories as qualities for which the values are mostly Yes/No, with (in some cases) more shades of Yes:

- **Fitness for purpose:** Complete, Authoritative, Largely compatible;
- **Adoption/reliability:** Known, Discoverable, Used in software, Used in data sets, Endorsed, Regulatory, Long-term, Sustainable, Participatory, Collaborative, Maintained;
- **Usability/openness:** Available on the web, Versatile, Served by APIs, Manageable, Documented, Supported, Testable, Machine-readable, Meaningful, Referenceable, Linked, Annotated, Clearly licensed, Openly licensed.

Most of these properties were included to support the gap analysis report and are therefore better described in Pesce et al. (2016).

2.3.4 Categorisation of standards

The following categorisations are used to organise records in the map, each using concepts mapped whenever possible to external URIs or URLs with definitions provided by authoritative bodies.

The full categorisations can be viewed here: http://vest.agrisemantics.org/about/structure

2.3.4.1 Domains and types of data

Domains

Having agreed that the map would cover a broad range of disciplines related to food and agriculture, developing the actual list of domains or disciplines under which standards would be classified was a difficult task; there seems to be no agreed classifications of food- and agriculture-related disciplines.

Universal classifications like the Dewey Decimal Classification (DDC) and the Library of Congress Classification (LCC) did not seem to be a good starting point, as food and agriculture disciplines are fragmented under different main headings (in DDC mostly under 500 for Science and 600 for Technology, but partly under Social Sciences; in LCC mostly under S for Agriculture and partly under T for Technology and Q for Science) and at different levels of granularity. Looking at domain-specific classifications designed by domain experts, the two major ones we could find (though it is not clear if they are currently in use) were:

- the FAO 'AGRIS/CARIS Classification' (FAO 1998)
- the Subject Category Codes of the US Department of Agriculture (USDA)[11]

The two classifications are not perfectly aligned and have some disciplines at a different level in the hierarchy, but a general alignment of the first level is possible.

Table 3. Rough alignment of FAO AGRIS/CARIS classification and USDA category codes

FAO AGRIS/CARIS	USDA mapped
A. AGRICULTURE IN GENERAL	Agricultural (General) (A000)
B. GEOGRAPHY AND HISTORY	Geography (B100) – Agricultural History and Biography (B500)
C. EDUCATION, EXTENSION AND INFORMATION	Agricultural Education and Training (not Extension) (C100) – Extension and Advisory Work (Non U.S.) (C200) – U.S. Extension Services (C210)
D. ADMINISTRATION AND LEGISLATION	Administration of Agricultural Agencies and Organizations (D100) – Laws and Regulations (D500)
E. ECONOMICS, DEVELOPMENT AND RURAL SOCIOLOGY	Agricultural Economics (General) (E100) – Home Economics and Human Ecology (U000) (+ Auxiliary)
F. PLANT SCIENCE AND PRODUCTION	Plant Science (General) (F000) – ~Food Science and Food Products (Q000)
H. PLANT PROTECTION	Pesticides (General) (H000)
J. POSTHARVEST TECHNOLOGY	~Agricultural Products (General) (S000) – ~Agricultural Engineering and Safety (N000)
K. FORESTRY	Forestry (K000)
L. ANIMAL SCIENCE, PRODUCTION AND PROTECTION	Animal Science (L000) – ~Feed Products (R000)
M. FISHERIES AND AQUACULTURE	Aquatic Sciences (M000)
N. AGRICULTURAL MACHINERY AND ENGINEERING	Agricultural Engineering and Safety (N000)
P. NATURAL RESOURCES AND ENVIRONMENT	Natural Resources (P000) – ~Soil Sciences (J000)
Q. PROCESSING OF AGRICULTURAL PRODUCTS	Agricultural Products (General) (S000) – Food Science and Food Products (Q000)
S. HUMAN NUTRITION	Human Nutrition (T000)
T. POLLUTION	Pollution (W000)
U. METHODOLOGY	Agricultural Research and Methodology (A500)
-	Meteorology and Climatology (B200)

11 http://agricola.nal.usda.gov/help/categorycodes.html

The VEST Registry already had a classification of domains which had been directly derived from the FAO AGRIS/CARIS categories, although using a different terminology. Since this classification aligned well with the FAO one and was not too far from the USDA one, we decided to continue using it in the new map of standards.

Table 4. Alignment between VEST domains and FAO AGRIS/CARIS classification

VEST domains	FAO AGRIS/CARIS classification
Agricultural Research, Technology and Engineering	N. AGRICULTURAL MACHINERY AND ENGINEERING – U. METHODOLOGY
Animal Science and Animal Products	L. ANIMAL SCIENCE, PRODUCTION AND PROTECTION
Economics, Business and Industry	E. ECONOMICS, DEVELOPMENT AND RURAL SOCIOLOGY
Education and Extension	C. EDUCATION, EXTENSION AND INFORMATION
Farms and Farming Systems	J. POSTHARVEST TECHNOLOGY
Fisheries and Aquaculture	M. FISHERIES AND AQUACULTURE
Food and Human Nutrition	Q. PROCESSING OF AGRICULTURAL PRODUCTS – S. HUMAN NUTRITION
Forest Science and Forest Products	K. FORESTRY
Government, Agricultural Law and Regulations	D. ADMINISTRATION AND LEGISLATION
Health and Pathology	H. PLANT PROTECTION – T. POLLUTION
Natural Resources, Earth and Environment	P. NATURAL RESOURCES AND ENVIRONMENT
Plant Science and Plant Products	F. PLANT SCIENCE AND PRODUCTION
Rural and Agricultural Sociology	E. ECONOMICS, DEVELOPMENT AND RURAL SOCIOLOGY

In order to cover the neighbouring and cross disciplines that we also wanted to include, we added a 'General/peripheral' category under which we grouped:

- Breeding and Genetic Improvement
- Geopolitical domain
- Information Management
- Language domain
- Meteorology and Climatology
- Organisms
- Physical and Chemical Sciences

Types of data

A granular classification of what data standards cover should go beyond just the domain (natural resources, plant sciences, fisheries, nutrition etc.) and indicate the type of data (e.g. plant phenotypic data, soil chemical properties, food prices, food nutrients etc.) that the standard wants to formalise.

As indicated in the project proposal, the map of standards and the gap analysis should cover 'standards in use for the exchange of key data on agriculture and nutrition'. Which types of data are key for food and agriculture? Before even evaluating which ones are key, it's difficult to come up with a classification of data set types on which all experts would agree.

In order to adopt an approach to types of data that was not too isolated, we aligned our work with that being done for the Open Data Charter Agricultural Sector Package (http://agpack.info/), which included interviews with several experts and also addressed the challenge of classifying data types. The audience of the Ag Sector Package is governments, so the selection of types of data is limited to that scope. The map of standards meanwhile has to cover all types of data; but we have roughly aligned the classification and shared some definitions between the Sector Package and the map of standards for the first-level categories and for data types that belong to each category.

Below is the current classification of types of data.

- **Administration and legislation data**
    - Government finance data: *Data on agronomic subsidies, taxes and fines.*
    - Land tenure data: *Data including the location, dimensions, boundaries and ownership of land parcels, which may also include details such as titles and specified land use.*
    - Official records: *Government data regarding official records, which may include relevant licences and permits, safety inspection data, tariffs, rates and pesticide usage data.*
- **General** (any type of entity in the domain)
- **Information resource metadata**
  This refers to metadata describing information resources. It refers to the type of information resource described and in that sense this category is domain-independent: the resource described can pertain to any domain. These metadata data sets don't contain the data directly, but describe and give access to either unstructured information resources like documents, journals and multimedia or structured resources like data sets or catalogues.
    - Sub-types are: Audio resources, Bibliographies, Blog posts, Catalogues, Data sets, Documents, Journals, Learning resources, Links, Material for agricultural extension, News, Photos, Presentations, Tweets, Videos
- **Natural resources, earth and environment data**
    - Geospatial data/objects: *Data including topographic and physical maps (and satellite or laser imagery) of natural features such as mountains, rivers and forests, and transport and building infrastructure.*
    - Hydrological data: *Data on the state of water, such as rivers, lakes and oceans, which may include real-time river and sea levels and flow data, flood zone locations, real-time and historical flood data, and water quality and temperature data.*
    - Land use data: *Data regarding agricultural land usage and changes in usage, such as different crop or vegetation cover.*
    - Soil data
    - Weather/meteorological data: *Real-time and historic observational and forecast data, which may include weather states, temperatures, rainfall, radiation, moisture, humidity, evaporation and climate maps.*
- **Research and agronomic data**
  Data generated from agricultural research (observations, field experiments, accessions) and scientific data (taxonomies, chemical elements, molecule composition) plus data on agronomic practices and agricultural technologies.
    - Agronomic data, agricultural technologies: *Data related to crop selection, agricultural technologies, treatment of diseases etc.*
        - Pest data: *Data on occurrence and treatment of diseases.*
    - Food nutrients: *Data on nutrients and other food ingredients.*
    - Livestock research data
        - Animal diseases
    - Models
    - Organisms
    - Plants/germplasm: *Data on plants and crops, which may include repository information,*

cultivars, landraces, farmers' varieties, breeding lines, accessions, genetic stocks and other related material. May also relate to non-cultivated flora and fauna, such as invasive or threatened species.

  - Crops
  - General Germplasm
  - Germplasm accessions
  - Location and Environmental
  - Phenotype and Trait
  - Plant Anatomy and Development
  - Structural and Functional Genomic

- Soil research data: *Data on the state of soil, including soil maps, expected soil conditions and nutrients, and outlined suitability for different uses, which may include pollutant data, such as emissions of land pollutants and contaminated land.*
  - Soil climate regimes
  - Soil maps
  - Soil profiles

- **Social/institutional, management, collaboration and coordination**
  - Events
  - Experts/People
  - Institutions
  - Policies
  - Projects
  - Publisher Policies
  - Vacancies

- **Socioeconomic data**
  Includes market data, demographic data, value chain data (food product data, farm data), infrastructure data, agriculture production aggregates.
  - Demographic data
  - Market data: *Data such as the locations of commodities and commonly traded goods markets, and trade price data, which may include daily wholesale maximum, minimum and modal prices at regional to national level.*
    - Food prices
  - Statistical data: *Various populations and agricultural-related data, which may include employment statistics, food security indicators, amounts of land dedicated to/value added by different crops, crop yields, fertiliser and irrigation usage, and other agricultural production details.*
  - Value chain data: *Data describing the actors in the agricultural value chain and the quality of their activities. Data related to the suppliers/ growers of crops and all related value chain data, which may include operational data like location, types of crops grown, supply chains, inventory data, farm or company level production and efficiency data, product data and transactional details.*
    - Farm management data
    - Food product data: *Data related to consumable products, such as menu and recipe information and ingredients, and may include a breakdown of treatments and pesticide usage.*
    - Public infrastructure data: *Roads, transport, market infrastructure.*

2.3.4.2 Types of data standards/ vocabularies

As explained above, we consider all types of 'vocabularies' – those that formalise descriptions of things and those that formalise sets of concepts – even if they are not encoded as XML or RDF vocabularies.

In very broad terms, vocabularies can be divided between description vocabularies and value vocabularies. Description vocabularies are those that prescribe the properties to be used to describe an entity as well as the relations between the entities described, from UML models to metadata schemas to ontologies, though ontologies are often a combination of description and values. Value vocabularies range from code lists to classifications to

thesauri.
However, there are so many different types of vocabularies within these two broad categories and the differences affect their use so much, that we decided to adopt a very detailed list of types to categorise the data standards. We also decided to include types that are not traditionally included in these categories: ISO data standards (which in some cases are models and in several cases are standardised value lists to be used in data sets) and 'messaging standards' (basically syntactic rules for event-driven messages, usually describing some time-related information such as an invoice or a certificate), because they are very much used in the industry and are in some cases regulatory.

One existing list of vocabulary types (only value vocabularies) is the Dublin Core list of KOS types (http://wiki.dublincore.org/index.php/NKOS_Vocabularies#KOS_Types_Vocabulary), which we re-used in our classification. We could not find other existing classification of vocabulary types, so the types we added have no correspondence in other systems.

Table 5. Types of vocabularies covered

Name	Description	Links and URLS
Application profile	A 'schema' which consists of metadata elements drawn from one or more namespaces, combined together by implementors, and optimised for a particular local application. See: http://www.ariadne.ac.uk/issue25/app-profiles	
Classification scheme	Schedule of concepts and pre-coordinated combinations of concepts, arranged by classification. Source: http://wiki.dublincore.org/index.php/NKOS_Vocabularies#KOS_Types_Vocabulary	http://w3id.org/nkos/nkostype#classification_schema
Code list	A code list is a predefined list from which some [statistical] coded concepts take their values. Source: 'GESMES/TS User Guide', Release 3.00, February, 2003. ISO standards that specify lists of scientific/industry-controlled/coded values can be categorised under this type.	
Dictionary	A reference source containing words usually alphabetically arranged, along with information about their forms, pronunciations, functions, etymologies, meanings and syntactical and idiomatic uses. Source: http://wiki.dublincore.org/index.php/NKOS_Vocabularies#KOS_Types_Vocabulary Based on: Merriam Webster online	http://w3id.org/nkos/nkostype#dictionary
Encyclopedia		
Gazetteer	Geospatial dictionary of named and typed places. Source: http://wiki.dublincore.org/index.php/NKOS_Vocabularies#KOS_Types_Vocabulary Based on: NKOS definitions. Based on KOS Taxonomy http://nkos.slis.kent.edu/KOS_taxonomy.htm	http://w3id.org/nkos/nkostype#gazetteer

Glossary	A collection of textual glosses of specialised terms with their meanings. Source: http://wiki.dublincore.org/index.php/NKOS_Vocabularies#KOS_Types_Vocabulary Based on: Merriam Webster online	http://w3id.org/nkos/nkostype#glossary
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. Source: http://wiki.dublincore.org/index.php/NKOS_Vocabularies#KOS_Types_Vocabulary. Based on: ANSI/NISO Z39.19-2005 (R2010). Guidelines for the construction, format, and management of monolingual controlled vocabularies. ISBN: 1-880124-65-3	http://w3id.org/nkos/nkostype#list
Messaging standard	'Messaging standards are standards which describe how to format syntactically (and sometimes semantically) a message usually describing some event or time related information such as an invoice or a certificate. Messages are sent because an event has occurred (a product has arrived or is ready for collection), or because an action needs to be taken (e.g. payment for an invoice). A good example is EDIFACT.' (Christopher Brewster) ISO standards that specify messaging standards can be categorised under this type.	
Metadata element set	Any set of metadata elements, like XML schemas, RDF schemas or less formalised set of descriptors. In the VEST Registry, we distinguish between a pure metadata element set (defining only basic classes and properties, like an XML schema or an RDF schema or a set of descriptors) and a real ontology defining also relationships and axioms and expressing complex models. Real ontologies are categorised under Ontology. 'A metadata element set defines classes and attributes used to describe entities of interest. In Linked Data terminology, such element sets are generally made concrete through RDF Schemas or OWL Web Ontology Language ontologies, the term "RDF vocabulary" often being used as an umbrella for these.' Source: W3C: https://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/ URL: https://www.w3.org/2001/sw/wiki/Library_terminology_informally_explained#Metadata_element_set_or_element_set	
Model	An abstract model that organises elements of data and standardises how they relate to one another and to properties of the real world entities. Source: Wikipedia. Unless they are formalised as ontologies or XML schemas, they can be categorised under this broader type, which includes UML models and entity-relationship models.	

Name authority list	Controlled vocabulary for use in naming particular entities consistently. Source: http://wiki.dublincore.org/index.php/NKOS_Vocabularies#KOS_Types_Vocabulary Based on: ISO25964-2:2013. Information and documentation – Thesauri and interoperability with other vocabularies. Part 2: Interoperability with other vocabularies	http://w3id.org/nkos/nkostype#name_authority_list
Ontology	A 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). Source: http://wiki.dublincore.org/index.php/NKOS_Vocabularies#KOS_Types_Vocabulary Based on: W3C Linked Data Glossary. W3C Working Group Note 27 June 2013	http://w3id.org/nkos/nkostype#ontology
Subject heading scheme	Structured vocabulary comprising terms available for subject indexing, plus rules for combining them into pre-coordinated strings of terms where necessary. Source: http://wiki.dublincore.org/index.php/NKOS_Vocabularies#KOS_Types_Vocabulary Based on: ISO25964-2:2013. Information and documentation – Thesauri and interoperability with other vocabularies. Part 2: Interoperability with other vocabularies	http://w3id.org/nkosnkostype#subject_heading_scheme
Taxonomy	Scheme of categories and subcategories that can be used to sort and otherwise organise items of knowledge or information. Source: http://wiki.dublincore.org/index.php/NKOS_Vocabularies#KOS_Types_Vocabulary Based on: ISO25964-2:2013. Information and documentation – Thesauri and interoperability with other vocabularies. Part 2: Interoperability with other vocabularies	http://w3id.org/nkos/nkostype#taxonomy
Terminology	Set of designations belonging to one special language [ISO 1087-1:2000, definition 3.5.1]. Source: http://wiki.dublincore.org/index.php/NKOS_Vocabularies#KOS_Types_Vocabulary Based on: ISO25964-2:2013. Information and documentation – Thesauri and interoperability with other vocabularies. Part 2: Interoperability with other vocabularies	http://w3id.org/nkos/nkostype#terminology
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. Source: http://wiki.dublincore.org/index.php/NKOS_Vocabularies#KOS_Types_Vocabulary Based on: ISO25964-2:2013. Information and documentation – Thesauri and interoperability with other vocabularies. Part 2: Interoperability with other vocabularies	http://w3id.org/nkos/nkostype#thesaurus

2.3.4.3 Other categorisations

Other controlled lists (taxonomies) in the map of standards are used as values of some key metadata, like format and licence. Formats in the taxonomy come partly from IANA (http://www.iana.org/assignments/media-types) and from the W3C formats (http://www.w3.org/ns/formats) with the addition of formats that are not in any standard list. The list of licences was compiled from the Open Definition Licenses Services (http://licenses.opendefinition.org/).

All classifications can be found here: http://vest.agrisemantics.org/about/structure

2.4 Call to action to experts

In order to make this platform a reliable asset, we called (and we'll continue calling) on our networks of experts to help us improve the map of standards and make it grow. Those who have already contributed are acknowledged here: http://vest.agrisemantics.org/content/credits.

The call to action was sent by email and published on several websites with the following text:

We need your help especially if you:

- *are the owner/manager of a standard/vocabulary relevant for food and agriculture-related data;*
- *know of any standard/vocabulary relevant for food and agriculture-related data that is not currently in the map;*
- *are an expert in open data standards.*

If so, please check the current map and if you can, help us in any of the following ways:

- *Add new standards that are not in the map yet. Click on Contribute: you can register/login (it's just an access account, we don't ask for additional information) and once logged in click on 'Add vocabulary' under Contribute.*
- *'Claim' a standard that is already there and improve its description. Once logged in you can click on 'Claim' on the standard page and explain what your responsibility is for that standard. We will grant you access to edit the description of that standard.*
- *Share and link the content of your vocabulary. If your standard is defined using a standard vocabulary (RDFS, SKOS) or ontology (OBO, OWL) language, take one step further and upload it to the AgroPortal, a repository of ontologies and vocabularies where advanced functionalities are included.*
- *Complete the short survey highlighted on the website. This will help us improve the way the map is organised and offer better functionalities.*

Although the call was widely disseminated through several channels (the FAO AIMS website and related social channels, the GFAR website, the GODAN website, the Land Portal community, the Agroknow's mailing list and more precisely to 707 data experts, Agroknow's Facebook page and Twitter account), the response was rather poor. This was due partly to the holiday season (the call was disseminated in August) and partly to the very technical nature of the exercise and the difficulty of targeting exactly the experts with the necessary knowledge.

According to the Land Portal Foundation, who contacted partners individually about their use of vocabularies, many organisations when reading about 'standards' in the Call to Action labelled this as a technical survey and deemed their knowledge too limited to participate.

What worked better was a) individually contacting experts whom we knew were interested in data standards as well as competent in one of the domains in our scope; b) surveying relevant websites and contacting the responsible organisations.

Besides partners directly involved in GODAN Action and those who had already contributed to the VEST Registry and the AgroPortal in the past, other organisations who provided input to the map in this phase are Institut National de la Recherche Agronomique (INRA), TNO Netherlands, National Institute of Informatics Japan, Commonwealth Scientific and Industrial Research Organisation (CSIRO) Australia. Others (International Food Policy Research Institute (IFPRI), Wageningen University, Tom Baker from DCMI) have promised to provide input soon.

An example of how work on specific areas can be conducted in the future is the work conducted by the Land Portal Foundation to survey a number of interesting vocabularies used in the land sector.

Since not many from the land governance community responded to the call to action, the Land Portal team approached the mapping exercise by visiting key websites of land-focused organisations and mapping the use of standards themselves. The Land Portal targeted websites of a sample of organisations, representing the diversity of organisations that publish information on land with which the Land Portal collaborates. This sample varied from global organisations and networks to very local organisations working on land governance issues. The standards added to the map of standards are an indication of the types used, and are intended to display the range of their sophistication.

Another approach that we will take in the next phase towards the discovery of relevant data standards to be included is to look at existing data sets, and check whether the data set is expressed in a standard data format and/or whether it contains terms from a standard vocabulary.

3 Current coverage

At the time of writing, the map contains 281 vocabularies, of which 149 are within the scope of food and agriculture. In general, what is worth noting about the coverage of the map at this stage is:

- As well as the vocabularies inherited from the VEST and the AgroPortal (246 of which 113 for food and agriculture), 35 new vocabularies have been added by experts since the launch of the platform (1 August 2016), all domain-specific.
- Coverage by domain is not balanced and depends highly on contingent situations:
    - The domain that is best represented is that of plant sciences: the reason for this is that the AgroPortal is specialised in that domain and the 46 ontologies managed in that portal and imported into the map account for a good percentage of the vocabularies covered.
    - The good coverage of the natural resources domain is mostly the result of the involvement of experts in the areas of land (the Land Portal Foundation, partner in the project) and soil (partners in other projects).
    - The relatively good coverage of the food area (though still far from comprehensive) is a consequence of the participation in the IC-FOODS conference (7–9 November 16, Davis, USA), where experts in food data and semantics met to discuss food ontologies and where the map of standards was presented.
- The predominance of ontologies over all other vocabulary types is also a consequence of the inclusion of all the ontologies from the AgroPortal, although the development of ontologies seems to be a clear trend in recent years, especially in fields like plant sciences and food.

Below are two charts representing the relative coverage by domain and by type of vocabulary.

An analysis of the current content of the map, especially on the level of usability and openness of the standards, can be found in deliverable D.1.1.2 (Pesce et al. 2016).

Figure 2. Number of data standards by domain

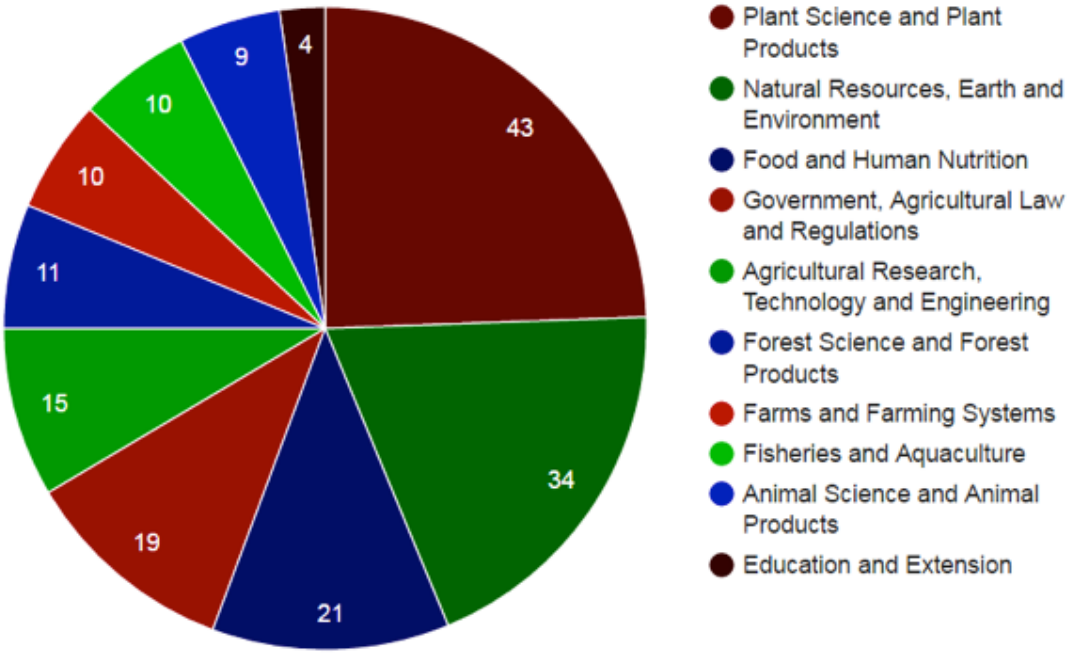
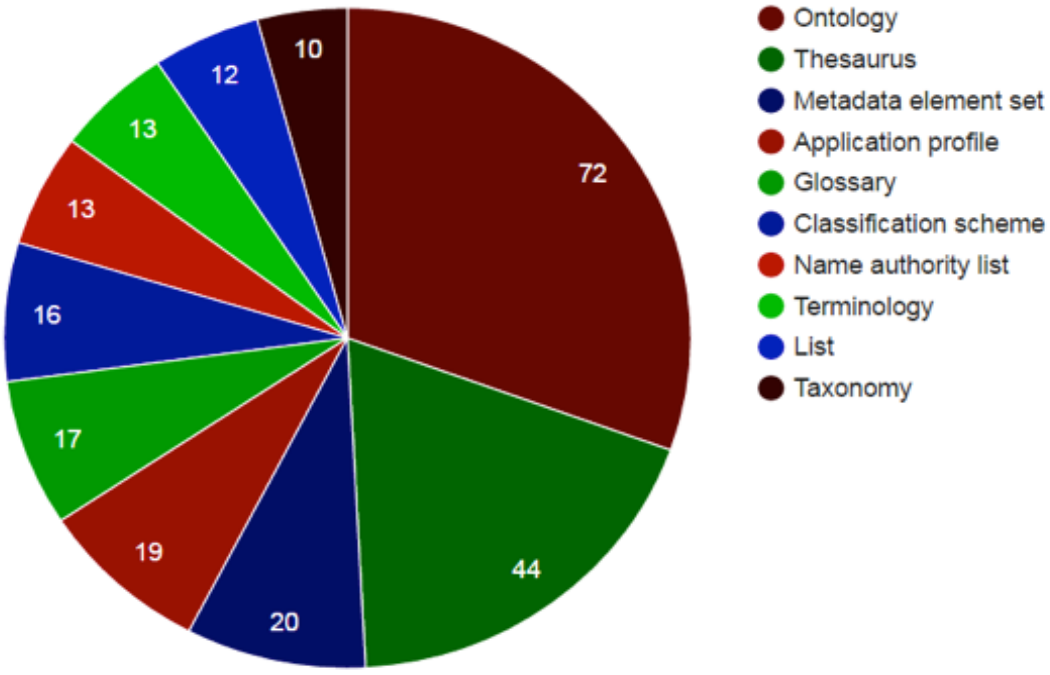


Figure 3. Number of data standards by vocabulary type



4 Conclusions

As this documentation shows, the map of standards was implemented following the principles of building on what exists, providing for sustainability and looking at standards in terms of use: principles that we put forward in the description of the whole project.

It is our intention to make sure that this map can be maintained well beyond the end of the project, as a common asset of the community working with agricultural and nutrition data. We have already illustrated some examples of collaborations in this sense and we will work on more in the next two years of the project.

We believe that having a one-stop shop where all domain-relevant data standards can be found will help to promote the wider adoption, standardisation, interoperability and re-use of vocabularies as well as of the data shared using these vocabularies. In addition, it helps to identify overlaps, duplication, gaps and limits to adoption, hopefully encouraging practitioners not to duplicate efforts and to collaborate to both develop and use common standards. In the next phase of developing the map, we will focus as planned on the first thematic topic that will be selected for the project. We will identify the types of data relevant for that topic and will survey and and better analyse the standards used in that domain.

At the same time, one of our main objectives for the next versions of this map, in years 1 and 2 of the project, is to work with more partners to cover all domains in a more balanced way.

References

Food and Agriculture Organization (1998) AGRIS/CARIS: Categorization Scheme, www.fao.org/docrep/003/u1808e/u1808e00.htm (accessed 20 November 2017)

Isaac, A.; Waites, W.; Young, J.; Zeng, M. (2011) Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets, W3C Incubator Group Report 25 October, https://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/ (accessed 20 November 2017)

Pesce, V.; Kayumbi, G.; Tennison, J.; Mey, L.; and Zervas, P. (2016) 'Agri-food data standards: a gap exploration report' GODAN Action Learning Paper http://www.godan.info/documents/gap-exploration-report-0 (accessed 22 November 2017)

GODAN Action brings together agriculture and nutrition specialists and open data experts to support individuals, organisations and communities to engage with open data.

The project is supported by the UK Department for International Development (DFID), led by Wageningen Environmental Research with international partners Agroknow, AidData, CTA, FAO, GFAR, IDS, Land Portal, and ODI.

Follow GODAN on Twitter: @godanSec

ORCID Identifiers:

Valeria Pesce https://orcid.org/0000-0003-3860-4304
Jeni Tennison https://orcid.org/0000-0002-3250-7530
Clement Jonquet https://orcid.org/0000-0002-2404-1582
Anne Toulet https://orcid.org/0000-0003-0463-0854
Sophie Aubin https://orcid.org/0000-0003-4805-8220
Panagiotis Zervas https://orcid.org/0000-0002-6531-4022
Lisette Mey - not available

GODAN ACTION
Global Open Data for Agriculture & Nutrition

UKaid
from the British people