Landscaping the use of semantics to enhance the interoperability of agricultural data

Sophie Aubin, Caterina Caracciolo, Panagiotis Zervas, Pascal Aventurier, Patrice Buche, Andres Ferreyra, Guojian Xian, Clement Jonquet, Antonis Koukourikos, Valeria Pesce, et al.

▶ To cite this version:

Sophie Aubin, Caterina Caracciolo, Panagiotis Zervas, Pascal Aventurier, Patrice Buche, et al.. Landscaping the use of semantics to enhance the interoperability of agricultural data: First deliverable of the RDA Agrisemantics Working Group. [Research Report] Research Data Alliance. 2017, pp.59. hal-02607815

HAL Id: hal-02607815
https://hal.inrae.fr/hal-02607815

Submitted on 9 Feb 2021

Landscaping the Use of Semantics to Enhance the Interoperability of Agricultural Data

RDA Agrisemantics Working Group - September 2017

Leading authors: Sophie Aubin, Caterina Caracciolo, Panagiotis Zervas
Contributors: Pascal Aventurier, Patrice Buche, Andres Ferreyra, Xian Guojian, Clement Jonquet, Antonis Koukourikos, Valeria Pesce, Ivo Pierozzi Jr, Catherine Roussey, Anne Toulet, Ferdinando Villa, Brandon Whitehead, Xuefu Zhang, Erick Antezana

Executive Summary

In this document we aim at composing a high level overview of the **use of semantics in the production, exchange and use of agricultural data**. In particular, we focus on the use of semantics for data management, and the resources, tools and services available for its production. We also look at the current research trends in the area.

Semantics refers to the description of the meaning of data, made possible by "semantic resources" (aka "semantic structures") aiming at making explicit the information that may help to find, understand, and reuse data(sets). Semantics may also make explicit the entities and relations the data embody. Granted that no data is ever produced or distributed without some attempts to describe its meaning (all databases have column names, all documents have a title and often some ways to indicate what topics they are about), the level of semantic richness, accuracy, shareability and reusability of the resources used vary greatly.

The goal of our this document is to indicate the main tasks where semantics is used or could be used for the treatment of agricultural data and highlight current bottlenecks, limitations and impact on interoperability of the current situation. The intended readers of this document are managers, project coordinators, data scientists, and researchers interesting in getting the big picture of semantics in agriculture. In particular, it aims at being readable and useful to the various communities involved, touching on both the data management and the agricultural side.

We use the phrase "semantic resources" to collectively refer to structures of varying nature, complexity and formats used for the purpose of expressing the "meaning" of data. However, we acknowledge the fact that not all semantic resources equally contribute to the achieving effective data interoperability, and wherever possible we highlight the current use of them and identify limitations. The concept of *agricultural data* is generic for data produced or used in agriculture, including data on agricultural production, or data relative to lab and field

experiments, environmental conditions or climate, just to mention a few relevant areas of data productions). Being aware of the width of the sector, we make no claims on comprehensiveness. In terms of formats, we are interested in both textual/semi-structured documents and structured data, including georeferenced data.

This landscaping exercise is based on (1) expertise of the group members (2) previous/ongoing initiatives and (3) a bibliometric analysis of the scientific literature. It lays the basis for the two following activities of the RDA Agrisemantics Working Group - the collection of real-life use cases where semantics is useful or needed, and the compilation of a set of recommendations for future infrastructural component supporting data management and semantics.

This document is organized in the following way. The Introduction (Sec.1) sets the context and introduces the terminology adopted in the course of the document. Sec. 2 (Semantics and Data Management) presents a high level account of possible different users of semantics in agriculture, with a discussion on state-of-the-art interoperability related work. Sec. 3 (Research Trends) analyzes the research trends in semantics for agriculture and nutrition in the past 10 years. Sec. 4 (Semantics Structures in the Agricultural Domain) describes the landscape of existing vocabularies for data specification in agriculture. Sec. 5 (The Semantic Expert Toolkit) describes the tools and services currently available for the creation of maintenance of semantic resources. As they are mostly generic tools, we look specifically at practices in the agriculture and food community. Finally, we draw our conclusions in Sec 6 (Conclusions and Next Steps).

Table of Contents

1. Introduction

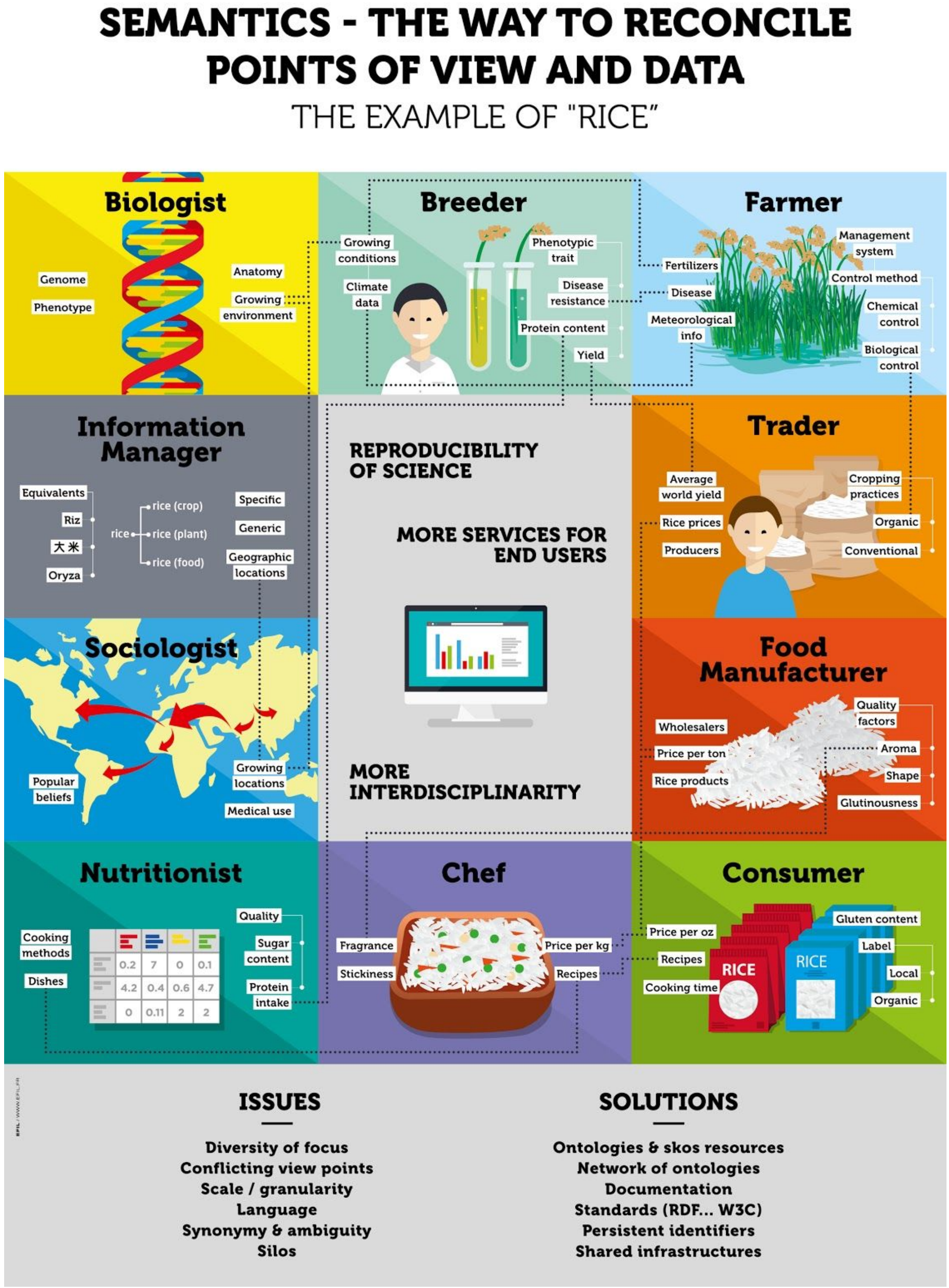
In agriculture, as well as in all other domains, data is produced in increasing amounts as well as by an increasing number of sources. This fact opens up a great deal of opportunities - To extract and analyse trends, discover new patterns, collect real-data information, test hypothesis, and understand impact of events possibly based on alternative theories, to mention only a few. However, these opportunities carry burning issues - How to describe the data produced so that its meaning is accurately described, preserved over time, and operable automatically also in conjunction with data produced by others? The principles of FAIR data (Wilkinson 2016),[1] namely data the are Findable, Accessible, Interoperable, and Reusable, express a widely shared concern, more and more relevant to agriculture, too.

Semantics is a well known area of study focussing on the “meaning” of languages, be those natural (e.g., in linguistics) or formal (e.g., in computer science, applied to programming

[1] https://www.nature.com/articles/sdata201618

languages), and possibly also to non-verbal communication (as in the case of semiotics). With the outburst of the web, semantics has gained a new area of application, as the increasing amount of data available in electronic formats calls for programmatic ways to find and manipulate data - hence, the need to “understand” their meaning. There is simply too much data for human experts to inspect each piece of data or datasets individually and decide, say, what its actual content is, to what geographical region it refers to, whether the information it contains is compatible with other pieces of information, or even simply judge which of the many lines in a text the title.

In the area of data management, it is common practice to use “semantics” to refer to any information that allow a system to (semi-) automatically identify the “meaning” of data. This includes the use of “traditional” metadata describing entire datasets or information items such as publications, but also more fine-grained, shared and machine readable description of individual pieces of data - not only serving the purpose of making a datasets findable on the web, but connect their contents in meaningful ways. This latter vision is graphically sketched in Figure 1 below.



SEMANTICS - THE WAY TO RECONCILE POINTS OF VIEW AND DATA

THE EXAMPLE OF "RICE"

ISSUES

Diversity of focus
Conflicting view points
Scale / granularity
Language
Synonymy & ambiguity
Silos

SOLUTIONS

Ontologies & skos resources
Network of ontologies
Documentation
Standards (RDF... W3C)
Persistent identifiers
Shared infrastructures

Figure 1: A graphical representation of "semantics" applied to data related to rice, devised by INRA with the collaboration of the Agrisemantics WG.[2]

[2] Also available at: https://www.rd-alliance.org/system/files/documents/SEMANTICS-RICE_poster_LD.jpg

Each block in Figure 1 represents an area of interest, and consequently data production, titled by the name of the professional typically involved in making (or using) the observations that are encoded in the data. Dotted lines connect “entities”, i.e., data that could and should be related and that currently remain isolated. Semantics is what allows one to draw those lines and exploit them for the purpose of achieving interdisciplinarity, reproducibility of science, and producing more and better services for end users.

A few issues that semantics must face are identified at the bottom of Figure 1. The different languages used to express the data are an obvious barrier, but even within the same language, the existence of synonyms and ambiguous expressions hamper the (re)use of data. Moreover, the difference in focus between disciplines may make difficult to find points of contact between otherwise obviously related data (e.g., a trader focusses on the amount of rice passing the douane, not in the various different species commercialized as “rice”, the focus of a biologist is typically independent of a specific location, while a farmer is as much interested in the biology of the plant she grows as in the climate of his farm). Moreover, with the same “object” passing from one focus to another, as in the case of “rice” along a production line, conflicting viewpoints may arise (e.g. all along the value chain, the price of rice successively includes the farmer’s work, transportation fees, then packaging,). Finally, it is often the case that observations made at different scale are hard to connect, as in the case of genotypes, anatomy and ecology. Next to the issues though, we also have solutions, developed in the broad area that we call semantics. Producing good documentation and appropriate metadata for any datasets is always the basis. Then adopting standards as much as possible, publishing data vocabularies online and with persistent identifiers, and adopting shared infrastructures whenever possible. Finally, ontologies should be used to formally describe the features of the entities observed and their relations formats.

To summarize, possibly the main message of Fig. 1 is that semantics and data interoperability go hand in hand. **Interoperability** is to be understood as the property of systems of being able to process information originally generated by and for third party applications. While syntactic interoperability (Wegner 1996) is in principle easy to achieve, **semantic interoperability** (based on the interpretation of the meaning of the data, beyond and above the syntax in which they are expressed) is generally more difficult as it implies that one is programmatically able to know if similar items from distinct datasets or information systems actually refer to the same objects of the world; and if not, in what they differ, and how they relate to each other. A correct and unambiguous description of the semantics of the data and the observable is the key element to pass data from one system to another and link data across systems in various applications and scientific endeavours. For example, data containing data about “corn height” cannot be automatically exchanged and reused unless it is unambiguously defined what “corn” and “height” are. Note that “corn” is a common name that actually collectively refers to different species, varieties and cultivars, while “height” implies establishing exactly what is measured and how, e.g., “height of the first leave at week 6th of development”.

Interoperability is to be actively sought in order to support the reuse of data, and so avoid duplicating effort. This approach is to be encouraged, as agriculture is a widely interdisciplinary field and the data needed in a given study or analysis may come from different sources and communities, and at different scale of observation. Our stand is that semantics is key to achieve interoperability, in that only by providing an explicit and machine readable meaning of data it is possible to programmatically reuse data.

As mentioned above, **metadata** (literally, "data on data") is a fundamental part of semantics. Metadata is used to describe salient features of data or datasets, and it is useful to distinguish at least two basic components of metadata - metadata elements (usually grouped together in a metadata schemas), and value vocabularies. A **metadata schema, or set of metadata elements**, defines classes and attributes used to describe entities of interest. For example, Dublin Core is generic for document description, Darwin Core for the description of specimens, DCAT for data catalogs, etc. In the web and especially within the Linked Data approach, metadata element sets are generally expressed by means of schemas in RDFs or OWL.[3] When a metadata element may only assume a limited set of values, **value vocabularies** are needed**.** Those are lists of possible values for elements in a metadata element set. For examples, a list of "topics" would provide the values for corresponding properties in a metadata schema. These lists may be organized in various structures, from flat lists to complex hierarchies, and expressed in various formats. Depending on the use or the domain of development, they may be known as thesauri, code lists, term lists, classification schemes, subject heading lists, taxonomies, authority files, digital gazetteers, concept schemes, or more generically, as Knowledge Organization Systems (KOS). Often, the term **vocabulary** is used to refer both to metadata element sets and value vocabularies (sets of controlled values)."*[4]* In this document we often apply this use. In the GODAN Action project[5], the term "data standard" is used to collectively indicate both types of vocabularies, also independently of the format used to express them. We prefer to avoid the expression "standard" in that it mixes the notion of status of originators and the actual adoption of the standard. However, it is sometimes used in Sec. 4.

In this document, we follow the common practice to use the expressions **semantic resources or semantics structures** interchangeably, to refer to any resource, be that metadata schema, value vocabulary, or ontology used to characterize in some ways the meaning of a piece of data or dataset.

This said, a question raises naturally - Is there a specific semantics for agriculture? And a specific type of interoperability for agricultural data? Despite agriculture being a highly interdisciplinary domain (it encompasses anything related to plant breeding and production as well agricultural practices, but it also includes diverse disciplines as rural sociology, agricultural economics, and soil, water and climate sciences to mention only a few), and the term "agriculture" often being used as a short label for *anything* that has to do with food production[6], one may argue that the basic notions on which semantics should be defined should be common to all domains. We share this argument but we also acknowledge the

[3] https://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025

[4] Note that W3C also states that this is why we use the terms "data standards" and "vocabularies" interchangeably, but we do not adopt this terminology here.

[5] GODAN Action Deliverable 1.1.2, "Gap exploration report" https://docs.google.com/document/d/1J2l_CUG56Ibd0PiO6hN1OuS-z0Qd0FQGV_vTOJP82Nw/edit

[6] Cf. the mandate of the Food and Agriculture Organization of the UN.

social nature of science, and the importance of having a community that shares the same vision and promotes actions to achieve common goals. For this reason, the scope of this document is to frame semantics in agriculture-related data management.

In this document, the data produced in any domain related to agriculture would collectively be called **agricultural data**. The data producers and users may then include researchers, practitioners and policy makers in agriculture. The definition of what is a domain, or subdomain, is a notoriously difficult task, often more related to institutional organization than to objective borders between them. Still, it is often convenient to attempt some organizations by domain. A recent exercise in this respect was done within a GODAN Action project, which produced a list of domains to categorize the records kept in the registry as well as in AgroPortal[7]. The list include 14 main categories, such as "Plant science and Plant Products", "Farms and Farming Systems", "Natural Resources, Earth and Environment", "Forest Science and Forest Products" etc - each obviously having strong links with a number of disciplines (climate, environment, geospatial, biology...). In this document, we sometimes refer to this categorization (especially in Sec. 4), mainly as a starting point to illustrate the general situation in agricultural data.

Within the broad area of agriculture (and agricultural data), it is also common to talk about statistical data, weather data, climate data, bibliographic data, genebank / germplasm data, earth observation and remote sensing data, as well as genetic data, georeferenced data, sensor data and so on. Sometimes these are considered as subdomain of the main broader area of agricultural data, sometimes as almost orthogonal classifications, to highlight different features such as their origin (e.g., sensors), typical contents (e.g., bibliographic data), level of generalization (c.f., weather data and climate data) and so on. The actual format of serialization (e.g., whether as relational database or .csv, or XML document or PDF) may also be expressed or not, and so any other higher-level classifications (e.g., structured/unstructured, georeferenced, ...). In this document, we do not aim to address specifically all these different types of data that are relevant to agriculture, although we let those distinction enter our discussion whenever possible.

Finally, the notion of Open Data refers to both legal and technical characteristics of datasets that may be reused by third parties for purposes that include selling for profit, as long as credits are attributed. In this document we make no assumptions on the openness of the data, although we strongly encourage the publication and reuse of Open Data.

[7] VEST / AgroPortal map of standards. Organization by domains: http://vest.agrisemantics.org/about/structure?qt-content_organization_tabs=3#qt-content_organization_tabs. The list is the result of alignment and merge of two major classifications: the "AGRIS/CARIS Classification" of FAO, and the Subject Category Codes of the US Department of Agriculture.

2. Semantics and Data Management

In this section, we give an overview of the main areas where some “semantics” is involved. The areas identified are the result of a group discussion held during the 9th RDA Plenary Meeting held in Barcelona in April 2017. The goal of this section is to show the width of application of semantics, discuss what type of semantic resources are used or preferred, their advantages and limitations.

It should be clear that not all users need to be aware of the semantic structures adopted for the creation, aggregation and search of the data they are inspecting, and typically they are not. However, some sort of semantic resources are often in use behind the scene, transparent to the end user.

Independently of the actual use of semantics, we would like to highlight three different groups of users, with respect to their level of involvement and awareness of the semantic structures used in any given application:

1. **Data end users.** These are the daily users of search engines on websites, online databases, or content management systems. They may include agriculture practitioners and researchers, as well as the general public. They might ignore all technical aspects of the use of the semantic structures in the dataset they are using (for example, for indexing, categorisation, query expansion, or reasoning) and even their existence all together.
2. **Developers of data-oriented applications (using semantic resources).** They may be considered intermediate users as they interact with the semantic resources, but generally through a tool, not directly. They could be data managers and librarians, application/infrastructure designers and developers, data aggregators, text mining practitioners, to mention only a few.
3. **Developers and maintainers of semantic resources**. These are technical users, with a variety set of skills ranging from modelling to programming, often with a touch of system administration. They may be in charge of either or both of the creation of semantic resources from scratch, or their maintenance and may or may not interact closely with developers using semantic resources.

The uses of semantics identified in this section are meant to meet the perspective of data end users, while we expect Sec. 4 and 5 to be of interest mostly to the other two groups of users.

2.1 Search for information

Search is the activity in which users express an information need and receive as results a set of documents of data items, usually ranked in order of relevance. Search may be done against virtually any types of data and datasets, in any language or combination of (when the languages of the query and the result are different, that is called a multilingual search), but it is always based on some sorts of indexing, consisting of associating a selection of keywords to each information resources. Indexing may be done either manually or automatically, in either case with a critical roles played by controlled vocabularies, to ensure that preference of spellings and synonyms, homonymy, or even errors do not jeopardize the quality of the results retrieved. Multilingual controlled vocabularies are used to allow for matching of queries and documents in different languages, while relations and hierarchies are usually exploited to expand and (dis)aggregate results.

As indexing is a widespread technique, possibly applied within the whole spectrum of information retrieval and information management, controlled vocabularies (be those flat lists or more complex thesauri) are also widespread and largely used (AGROVOC[8], CAB Thesaurus[9] and NAL Thesaurus[10] to mention some of the best internationally known thesauri in the area of agriculture). However, controlled vocabularies tend to be **locally defined**, not accessible outside the application of choice, and so hardly reusable to index other datasets. The consequence of this is a **proliferation of resources**, siloed from one application and the other. According to the discussions held within the RDA Agrisemantics Working Group, the causes for these phenomena are that existing resources are scarcely findable, and the difficulty of extending the coverage of the existing one in case so as to meet the needs of specific application. It should also be noted that these locally defined resources often overlap in coverage, exactly for their being isolated and not accessing other resources (consider for the example the three above mentioned thesauri, sharing a large set of terminology, or "concept" according to the SKOS parlance). This was a normal and necessary condition before the web but now semantic resources, just like data, can be shared online to limit duplication of effort and minimize proliferation. It should also be expected that the resources no longer allocated to produce duplicated data may become available to improve the quality of the data published.

Linked Data techniques prescribe that vocabularies are published online[11], endowed with URIs and linked to one another (primarily by means of SKOS properties for semantic matching). In this way, they may be reused more easily (because of the URIs) and also allow for searches beyond one's own dataset (because of the links). However, controlled vocabularies are typically oriented to capture terminologies and language uses rather than to express logical definitions (semantics) that can be operated automatically to check identities, allow for programmatic integration of data or to draw inferences. For example, data containing data about "corn height" cannot be automatically exchanged unless it is unambiguously defined what "corn" and "height" are, and when and how exactly the height of the corn is measured, e.g., height of the first leave at week 6 of development. In this sense controlled vocabularies only ensure a very first level of interoperability.

From the example given above, on the height of corn, it is clear that online LOD vocabularies address at least an important basic issues of semantic interoperability, namely the possibility of publicly define objects' identities. However, they also introduce the need for new techniques of indexing, such as indexing with URIs instead of terms (i.e., strings) or local codes, as it is normally done. However, the need exists for semantic-web compliant out-of-the-box indexing solutions that are ready for inclusion in other applications (Maui, Agrotagger, AgriDrupal).

[8] http://aims.fao.org/agrovoc

[9] http://www.cabi.org/cabthesaurus/

[10] https://agclass.nal.usda.gov/

[11] Best Practices for Publishing Linked Data https://www.w3.org/TR/ld-bp/#VOCABULARIES "Publish your vocabulary on the Web at a stable URI using an open license. One of the goals is to contribute to the community by sharing the new vocabulary.

Examples of use of linking technologies to smooth out the interoperability problems related to the use of local vocabularies already exist, see AGRIS[12] (Celli, 2015) indexed with the AGROVOC thesaurus (Caracciolo, 2012), a LOD vocabulary now expressed as a linked dataset. Another example is the Land Portal mashup[13], a semantic-mashup type of application reusing and repackaging data produced by third parties on the basis of semantic resources such as AGROVOC, the FAO Geopolitical Ontology[14] and World Bank Indicators[15] . The Land Portal addresses the general public as well as decision makers, and deals with a variety of data on land use, including demographic and economic data.

[12] http://agris.fao.org/agris-search/index.do

[13] http://landportal.info/

[14] http://www.fao.org/countryprofiles/geoinfo/en/

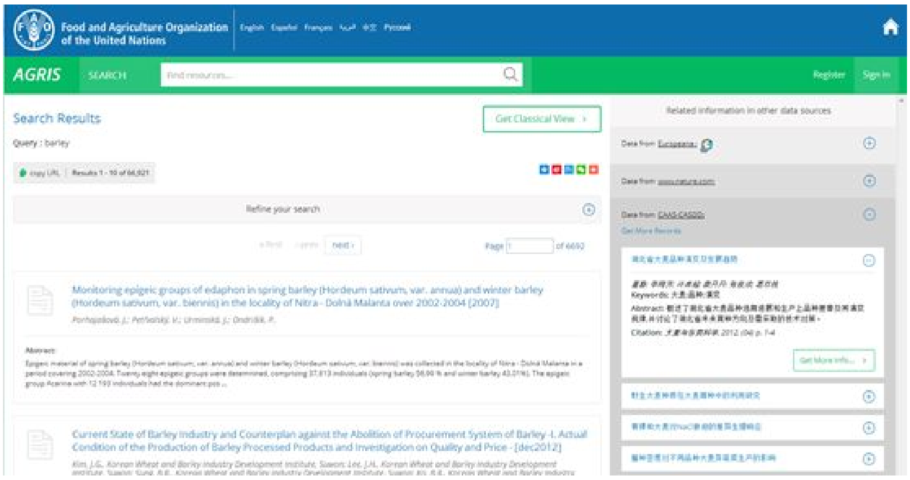
[15] https://data.worldbank.org/indicator

Connecting the Chinese Agricultural SciTech Documents Database with AGRIS

Chinese Agricultural SciTech Documents Database (CASDD) is an agricultural bibliographic database developed by the Agricultural Information Institute of the Chinese Academy of Agricultural Sciences (AII of CAAS). CASDD relies on the National Agricultural Library's content, one of the most comprehensive (over 10 million bibliographic records!), reliable and accessible Chinese literature resources of agricultural science and technology in the world (agronomy, horticulture, plant protection, soil sciences, animal husbandry, veterinary, agricultural engineering, agricultural products processing, agricultural economic and so on). So, making this data shareable and open is very meaningful and helpful to other countries.

The CASDD RESTful API provides a light-weighted and high performance solution for the third party applications (e.g. AGRIS) to access the records, such as query with Chinese and English keywords, AGROVOC and Chinese Agricultural Thesaurus (CAT) concepts and their HTTP URIs, authors, publication year and so on. The output formats include RDF/XML following the metadata models (e.g. the AGRIS AP) and just plain JSON.
The key point here is relying on the AGROVOC formal alignment with CAT and other KOSs, which acts as semantic bridges. Following this way, we can design and realize more applications to open and consume multilingual data internationally.

A flagship application of this CAT/AGROVOC mapping is the possibility for AGRIS users to benefit from the multilingual search feature for Chinese, searching the bibliographic database with Chinese keywords and get results in many languages, and conversely. In order to broaden the data usage, we interlinked with AGRIS's multilingual records, through creating mashup widget (Data from CAAS-CASDD) for the API in lower right of the AGRIS website pages (Fig.2). The mechanism is that the query terms user input (over 20 languages supported by AGROVOC) in the AGRIS search field will be mapped to AGROVOC concepts' URIs, and then the URIs will be sent to the CASDD RESTful API, the API will align the AGROVOC URIs to CAT URIs based on the KOSs mapping results. CAT URIs will finally be used to query the CASDD records with the help of lucene-skos in the URI-based term expansion way.



Monitoring epigeic groups of edaphon in spring barley (Hordeum sativum, var. annua) and winter barley (Hordeum sativum, var. biennis) in the locality of Nitra - Dolná Malanta over 2002-2004 [2007]

Current State of Barley Industry and Counterplan against the Abolition of Procurement System of Barley -I. Actual Condition of the Production of Barley Processed Products and Investigation on Quality and Price - [dec2012]

Fig2. Mashup Widget Interoperate with CASDD RESTful API in AGRIS

2.2 Information extraction

Techniques of information extraction (or text mining) are used to extract structured information from unstructured data, for example for reasoning. For instance, this is useful when data collected or processed in past research works have not been preserved in databases and are only recorded in the literature either in tables or in text form.

Information extraction tasks may be done either fully automatically, or in a semi-automatic fashion and the extracted information may still be in textual form, but more structured and concise. The extraction process can be applied to a variety of text, such as scientific papers, books, reports, free text fields of databases, collections, blog posts, tweets, surveys, etc.

Information extraction and text mining in general offer a first level of interoperability, from text to text, in that they allow for a common representation of concepts or objects expressed in various ways (multilingualism, syntactic or lexical variation, language level, etc.) through normalization and categorization. The second level of interoperability, from text to data, can be achieved by using common identifiers, e.g., URIs, for entities extracted from texts and objects in databases. While text mining tools still make little use of Linked Data approaches, we note a growing interest in this field.

Mostly, semantic resources are relevant in information extraction tasks when the tasks deal with the extraction of entities and their relations. They range from simple reference lists, hierarchically organised or not, to taxonomies and highly specialised application ontologies.

Named entity recognition and relation extraction involve resources which development can be (very) costly as it requires domain expertise and time. In the best cases, they can be derived from existing, shared resources like taxonomies and gazetteers. Still, as the resources grow in complexity and domain specialization, typically ontologies, the possibility of recycling goes down. In addition, as such engineered semantic resources are generally dependent on a given text mining technology. As a consequence, they lack standardisation and thus, are little shared and reused themselves.

The @Web[16] platform allows domain experts, guided by a domain ontology, to annotate and then extract data found in tables of scientific documents.

The OpenMinTeD[17] project for instance shows a constant concern for technical and semantic interoperability of tools and annotation resources. A couple of use cases developed in OpenMinTeD aim at linking data elements with texts, for instance gene marker records in the gnpIS platform to publications describing related research works.

[16] http://www6.inra.fr/cati-icat-atweb/

[17] Open Mining Infrastructure for Text and Data: http://openminted.eu/

2.3 Data organization

Data models organize elements of data and define how they relate to each other. They are fundamental to build and run information systems that capture, store, manage, query and analyse the data.

> *"If data models are developed on a system by system basis, then not only is the same analysis repeated in overlapping areas, but further analysis must be performed to create the interfaces between them."* (West & Fowler 1999)

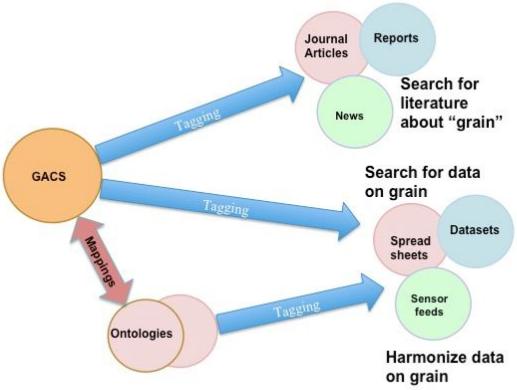
Mainstream data models, namely object serialization, relational, and hierarchical (XML) models, offer no means to internally record semantics or meaning, which often results in having it, when such a documentation exists, not up-to-date or difficult to access. In addition, the identifiers for the objects recorded in the databases are locally defined. These two features lead to poor sharing capability, ending in siloed databases and information systems and repeated development of similar models. In the world of mainstream data models, merging data from two databases developed independently require humans to understand the meaning of the data and agree on common formats to collaborate the two databases appropriately.

Using semantic data models (with RDFS and OWL) and the semantic web makes data management and reuse much simpler and efficient. Following good practices and guidelines, the meaning can be encoded in the model through standard formats and vocabularies for the structure and appropriate annotation properties. Possible values for those properties can also be encoded as controlled lists publicly available (and ideally published using standard, web-oriented formats, e.g., SKOS). This approach alleviates the burden of understanding the data and ensuring its consistency and capacity to evolve across time. In this, having external, community shared unique identifiers for common objects is also key. This is exactly what the GACS project aims to provide as a common infrastructure for agriculture and nutrition: unique identifiers for commonly used concepts and shared value lists.

Towards an open, persistent vocabulary for agriculture data and services: GACS

When data is published on the Web, using URIs as identifiers for their semantics, it is easier to interlink related elements among multiple online datasets ("Linked Open Data"). The techniques of Linked Open Data overcome many of the limitations of traditional information technology by expressing mappings and data semantics with global identifiers that can be looked up in schemas published on the Web, turning the Web into a vast, distributed dictionary for knowledge organization.

Over the years, FAO coordinated with the USDA National Agricultural Library (NAL) and the Centre for Agriculture and Biosciences International (CABI) on the improvement of their respective thesauri. In 2014-2016, the three partners have created the Global Agricultural Concept Scheme (GACS), a smaller concept scheme mapped to the 15,000 most frequently-used concepts in AGROVOC, NAL Thesaurus, and CAB Thesaurus. Concept mappings were computed with AgreementMakerLight and manually evaluated by staff of partner organizations. qSKOS and Skosify tools were used to check the quality of the resulting network. GACS beta is accessible for browsing in SKOSMOS[18] and is accessible in AgroPortal[19]. It is already usable for tagging information and datasets for discovery (semantic annotation), and also as building blocks for constructing other, more detailed knowledge organization systems such as ontologies. The GACS partners have committed to the long-term persistence of its URIs.



The project of consolidating the current GACS aims at improving its content and developing services on it to make it a key component of the Agrisemantics landscape offering solutions to data interoperability in the agriculture and food domains. GACS will be developed as a hub of concepts. By mapping the generic concepts of GACS to more granular, domains-specific concepts in ontologies, taxonomies, and specialized vocabularies, GACS can function as a switching language, glueing together a diversity of loosely compatible domain languages.

[18] http://browser.agrisemantics.org/gacs/en
[19] http://agroportal.lirmm.fr/ontologies/GACS

Ontologies are the semantic resources “corresponding” to database structures. Ontologies define object types, some constraints on them and how they interact. They can also provide labels for the objects and relations. Unlike other data models, ontologies are sharable and actually shared in general domain or community portals like AgroPortal. The aims of such portals are to reduce duplication of work and leverage the interoperability of information systems and datasets by the reuse and combination of small semantic models focused on specific domains of knowledge, e.g. crop/pests/farming technique.

2.4 Reasoning on data

Data are typically collected to gather information on a certain topic, event or object, to be further processed and analyzed to allow for informed decisions, make new findings, build arguments in policy and politics, among other uses. All these actions imply some sort of reasoning on the collected data, performed either by humans or machines. Machine-based reasoning requires that the data be logically described (often, called “annotated”) and that logical descriptions of the domain at hand are given, by means of ontologies.

Applications of machine-based reasoning include Decision Support Systems (DSS)[20], supporting analysis of complex situations and often producing a ranking of alternative solution for a given problem. Typical DSS also aim at supporting different types of users, to help stakeholders (e.g., farmers, advisors,retailers, funders, etc.) address complex tasks which imply to consider many parameters be they agronomical, biological, meteorological, environmental, economical, or social. DSS are applied to risks and uncertainties assessment and management linked to agricultural production like disease and pest control, crop rotation, nutrient management, water and drought management, food processing, and tasks automation. More recently, precision agriculture requires systems that combine different data to indicate the exact task or treatment to apply to each plant individually.

Such systems may be rather complex, as they may operate with a variety of parameters, e.g., on soil, weather, water, plants, animal, and micro-organisms, also on practices and know-hows. Also, the data used is experimental/observation data and data formalized based on knowledge elicited from expert knowledge. In fact, DSS typically include a database of data and a repository of models or ontologies that allow the system to perform the reasoning required. Both of them are typically produced and consumed by the same data producer, which ensure preservation of meaning through the operations performed by the system, but not necessarily the interoperability of the data. When heterogeneous formalisms are to be integrated, we are in the case described in the following section.

The semantic resources that are used for reasoning are typically ontologies. They aim to describe and control the properties and relations of different types of data in order to be able to 1) select and extract data with precise and meaningful queries from a database, and/or 2) reason on that data. To do so, the ontologies must provide formal (unambiguous) definitions

[20] Decision Support System have been studied and used in many areas, including agriculture. The interested reader can look at two reviews in the area, (Mir, 1970) and a (Lindblom, 2017).

of concepts and their relations, and axioms that allow an inference engine or any other simulation algorithm to produce new knowledge from the combination of data.

Reasoning systems usually utilize data produced on purpose, or rather, they are built around the data available, and when the need arises to integrate data produced by third parties, or for other purposes, some conversion/transformation procedure is needed. Therefore, although conceptually rather similar, the practice of reasoning on datasets built on purpose, and on integrating datasets with different origin, and possibly containing observation at different scales, are in practice quite different, that is why we dedicate a separate section to this latter application.

Sensors

Sensor technologies has improved and become more and more accessible for farmers. Wireless sensor networks and sensor miniaturization allow the deployment of sensing infrastructures in any place in an easy way. For example in the domain of smart city, Internet of Thing infrastructure has emerged in order to manage heterogeneous streams of sensor data. Similarly in agriculture, a new trend has emerged with precision agriculture based on Internet of Thing technologies. But, as mention in [Kamilaris et al 2016] no IoT platform exists for outdoor agriculture deployment, and smart city platforms should be adapted to reach agriculture needs.

With the help of advanced wireless sensor infrastructures, farmers can get ***real-time, highly accurate data*** from their fields and take the appropriate decision using Decision Support System (DSS). A large range of sensor types can be used in farms. Semantic web technologies will help handle heterogeneous sensor data stream in order to propose new DSS for farmers.

Sensors provide a basic measurement (e.g. temperature) called raw data. This raw data is provided without context by the sensor. Thus it has little meaning and its context should be expressed by adding metadata in order to interpret raw data correctly (e.g. air temperature or plant temperature or soil temperature). According to [Abowd et al., 1999] *“Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves”*.

For example [Goumopoulos et al, 2014] propose a context aware system dedicated to the management of strawberry irrigation system in a greenhouse environment. Wireless Sensor Network observes the environment and provides raw data stream (precipitation, soil moisture, air temperature, etc.). All the streams have to be enriched in order to build the context of the strawberry plant. The data stream enrichment task is called data stream annotation. The annotation process associates some metadata to the raw data to build what is called the low level context: the interpretation of the raw data stream. Some reasoning or machine learning techniques may be applied on the low level context to deduce more informative data (e.g. plant drought stress state, plant heat stress state). These new informations compose the high level context.

Ontology, vocabulary and semantic web techniques are used first to supervise the annotation process: define the context schema, integrate the heterogeneous raw data streams into the context schema and query the low level context. The context schema is composed of a network of ontologies. Semantic Sensor Network ontology [Compton et al, 2012] may be the core of this network of ontologies. Secondly, if high level context is required by the application, ontologies and inference mechanisms may be applied to deduce the high level context from the low level one. The goal is to extract from the low level context events that trigger the creation of a high level context (e.g. if the plant temperature and the air temperature is above a threshold during a time window the plant has drought stress). This extraction process may be done by a rule engine.

Machine learning techniques may further be applied on an archive of low level context in order to extract more rules or to adapt existing rules to a new plant.

2.5 Integrating an repurposing third party data

As hinted in the previous section, reasoning on data usually means that the data and ontologies used for the task are specific to some needs, or domain perspective, and they tend to reuse the terminology known and accepted in their domain. Data originated from different sources will then likely embody different perspectives and terminologies, e.g., different units of measurements or different methodologies of measurements altogether, different time frames, aggregation criteria, or scale. Consider for example the combination of phenotypic with genotypic data, or meteorological data with sensor data from the field. This implies that for data originated by third parties to be included in the applications some conversion processes has to be applied - not only to convert formats but to ensure that meaning is preserved from one system into the other.

Here is where most interoperability problem arise. Interoperability at the data level - so that data are comparable and combinable regardless of their provenance and scale - is still very challenging in the absence of common semantics. The two major issues are:

1. the assessment of the “identity” of the subject of the observation - do the entities from two separate systems refer to the same reality? Do they share the same properties that are important to the target application?
2. The specifications of relations among entities. One type of relations worth attention derive from the different scales the same object may be observed. Consider for example living organisms, such as plants. One the one hand, data may be produced about its morphology, or organs, cells, genes, and sequences. On the other hand, data may also be produced about an entire field, ecosystems or their interaction.

In order to address these issues, ontologies can serve as pivot models to define the features of the entities observed, state equivalence between identities and establish relations among them. In other words, ontologies can provide rules to transform and harmonize data in a meaningful manner.

Planteome & Crop Ontology

What genes are associated with a plant trait? What are the common adaptive traits in plants and how are genes network and expression affecting them. These are some questions that researchers in comparative plant biology address by analysing data of various types and backgrounds. To support their work, the Planteome project (http://planteome.org) builds a database of searchable and browsable annotations for plant traits, phenotypes, diseases, genomes, and gene expression data across a wide range of plant species. This data comes from various information systems, and is annotated using different vocabularies and keywords.

As data grows in size and diversity (20 database sources, 80 taxa, and 2 million bioentities including genes, germplasm, QTL) their integration and analysis becomes more difficult. To overcome this obstacle, Planteome has adopted an integrated approach, relying on common annotation standards and a set of reference ontologies for Plants on which applied ontologies are mapped. More than 17 million annotations actually link bioentites to terms from ontologies including reference ontologies (PO, TO, GO PATO) and application species specific ontologies (Crop ontologies). Specifically, the Crop Ontology (CO) (www.cropontology.org/) supports the development of crop specific ontologies by providing a specific trait templates for CO development which is then converted to an ontology format using the CO API. In the trait template, a CO term is defined by a unique combination of a Trait, a Method and a Scale. This very pragmatic approach has been widely adopted by biologists from CGIARs centers or from European infrastructures such as Elixir, INRA, Wageningen University among others.The strength of this approach relies in both its simplicity and its alignment to biologist everyday practices and concepts thus easing both the ontology curation and adoption. Currently, 21 crop specific ontologies are available on the CO website.

In order to allow the harmonized query of data annotated with terms from species-specific ontologies with data annotated with terms from the Planteome reference ontologies, a mapping of Crop Ontology to the Plant Trait Ontology (TO) was started recently (Laporte 2016).

To our knowledge, the system that most explicitly and effectively addresses these points is the k.LAB platform, a software package based on the Integrated Modelling approach, and exemplified in the ARIES application[21]. k.LAB consists of a set of generic ontologies (aka foundational, i.e., domain independent), and a set of domain specific ontologies, all defined according to compelling semantic principles that also allow for the expression of scale (Villa et al. 2017) and actionable by a software platform that allows users to easily extend the coverage of the system both in terms of domain ontologies and the corresponding datasets, so that new data sets may be seamlessly added to the repository without ad-hoc conversions and laborious process by the users. In fact, the key notion of the system is to consider data and domain ontologies as two sides of the same coin. The problem of identities is solved in k.LAB by referring to a few authoritative resources where

[21] http://aries.integratedmodelling.org/

comprehensive lists of identities are given, such as IUPAC[22] for chemical compounds and AGROVOC for common-sense taxonomies, GBIF[23] for scientific taxonomies. It is then possible to link those identities (typically, these are large sets of pairs, URI and labels) to the ontologies describing them (OWL2). The first large-scale application of this approach is the ARIES system (ARtificial Intelligence for Ecosystem Services: http://aries.integratedmodelling.org), a distributed infrastructure for the rapid assessment of ecosystem service values.

2.6 Conclusions

In this section, we have described the major areas of use of semantics in agricultural data management, with a specific look at data and system interoperability. The perspective adopted, and so the structure of the section, is the result of discussions which happened within the RDA Agrisemantics Working Group discussions, elaborated during the regular group calls and several private discussions and approved and consolidated during RDA Plenary 9[24]. The purpose of this section is to show that the use of semantics, or rather, the use of semantic resources, is ubiquitous in data management and use, although with different degrees of expressivity of the semantic resources used, and different support to the involved users. We also tried to highlight a few issues that are worth attention by the community.

Thesauri and controlled vocabularies have been originally devised for the purposes of indexing and retrieval of information resources, while ontologies are needed every time an advanced knowledge application is needed (e.g. reasoning). In either case, we have collected evidence of a tendency to proliferation of resources, both for vocabularies and for ontologies - cf. discussion on search. What has been indicated as reasons for such proliferation is the fact that resources are often not easily findable, or rather that the knowledge about them tends to be limited to specific communities, despite their coverage could possibly span more topics and communities. Moreover, resources are often not available online and therefore scarcely reusable. Finally, it was highlighted that governance of the editorial control is often a problem. In other words, people report the fact that often it is easier to create a new resource from scratch, or starting from reusable fragments, instead of requesting that an existing resource is improved and expanded.

Some of these problems may be addressed by adopting Linked Open Data (LOD) approaches, for example, the online publishing of resources makes them more easily findable, while the adoption of open formats and the creation of public APIs makes them accessible. However, the point of proliferation of resources becomes especially interesting when considering that thesauri and controlled vocabularies keep playing their main role of resources for indexing and retrieval, but we note a tendency to use them also to “tag” the identity of the subject measured in datasets of entities - observations, as per the scientific jargon. In other words, thesauri and controlled vocabularies are extending out of their traditional area of use, i.e., metadata schemas and information retrieval, and are more and

[22] https://iupac.org/
[23] https://www.gbif.org/
[24] https://www.rd-alliance.org/plenaries/rda-ninth-plenary-meeting-barcelona

more used as repositories of identities for the semantic web. Consider for example all those applications where it is important to distinguish and unambiguously refer to different species of plants and animals, or chemical compounds, or even agricultural practices or land uses. The suggestion we gathered from this landscaping exercise is that it is time to start carefully considering the difference between *keywords* and *codes* on one side, as normally dealt with in metadata schemes, indexing and information retrieval applications, and entities and URIs on the other side, as needed by reasoning and information integration in a world of abundant, distributed and related data. It is dubious that the Linked Data approach alone may solve the two issues here highlighted, namely the proliferation of semantic resources and the practical confusion between tagging and reasoning. Linking data and vocabularies helps, but ontologies are still needed, especially when the need is to integrate data on different object, at different scale and produced by different users.

3. Research trends

Having looked at how semantics contribute to applications for information and data management and analysis, we want to provide a short overview of current trends in research at the crossroads between semantics and agriculture/food/nutrition.

This section is based on the results of the bibliometric study that was run early 2017 which completes the one run by the e-ROSA project[25] by focusing on semantic aspects while e-ROSA's more generally tackles approaches to data interoperability in agriculture. The methodology, source choice, and analysis tools used are the same for both studies. The scope is the publications of the last 10 years in the form of articles, books, book chapters, proceedings papers, and reviews references in the Web of Science. The documents answer to the query made of keywords proposed and discussed by the members of the RDA Agrisemantics Working Group. A first set of semantics-related keywords plus a list of semantic resources known to be used in the domain were crossed with a second list of keywords denoting agriculture and nutrition. Some manual filtering was applied to exclude out of scope references (OWL as the bird, RDF as "recommended dose of fertilization", behavioral or medical related with no link to agriculture, environment nor nutrition, etc.). The detailed query is provided in Annex 1.
With this bibliometric study, we show the evolution of the domain and of its particular topics, how they are interlinked. We identify sub-communities of interest who collaborate on issues as varied as information management, sensors or big data. In a second time, we examine publication habits through journals and events.

3.1 Topics of interest

The interest of the scientific community in semantics applied to agriculture and nutrition issues is ever growing, particularly in the past couple of years as shown in Figure 2.

[25] see D1.1 on http://www.erosa.aginfra.eu/deliverables

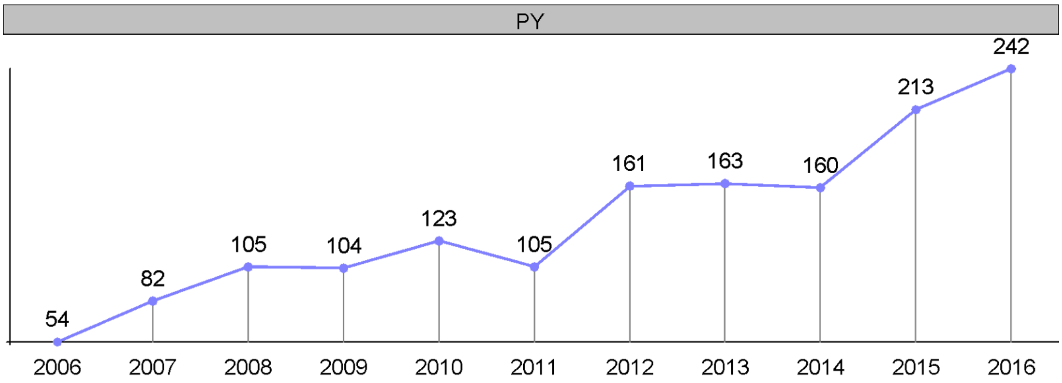


Figure 2: Evolution of the publications on semantics for agriculture between 2006 and 2016

The number of publications per year in the field has more than quadrupled in the last ten years with a clear acceleration in the past 3 years. One reason is the recent conquest by semantic technologies of new territories like sensor and big data as shown in Figure 3 which tends to represent the emergence of semantic topics across time[26]. The "time" at which nodes are positioned corresponds to the date when their number of occurrences reaches 20% of their total frequency over the whole dataset. The graphic clearly shows two distinct trends. The first, represented by the green and red nodes, is centered on information and knowledge representation, with applications in information retrieval and extraction. It tends to stabilization with less creation of new terms in recent years. A second trend, appearing in the early 2010s and represented by yellow and light blue nodes, focuses on data with integration and interoperability goals. The common denominator between both trends are metadata and controlled vocabularies, which are actually the subject of renewed interest.

[26] JPG file: https://www.rd-alliance.org/system/files/documents/historicalmap75-RADAR-2.jpg

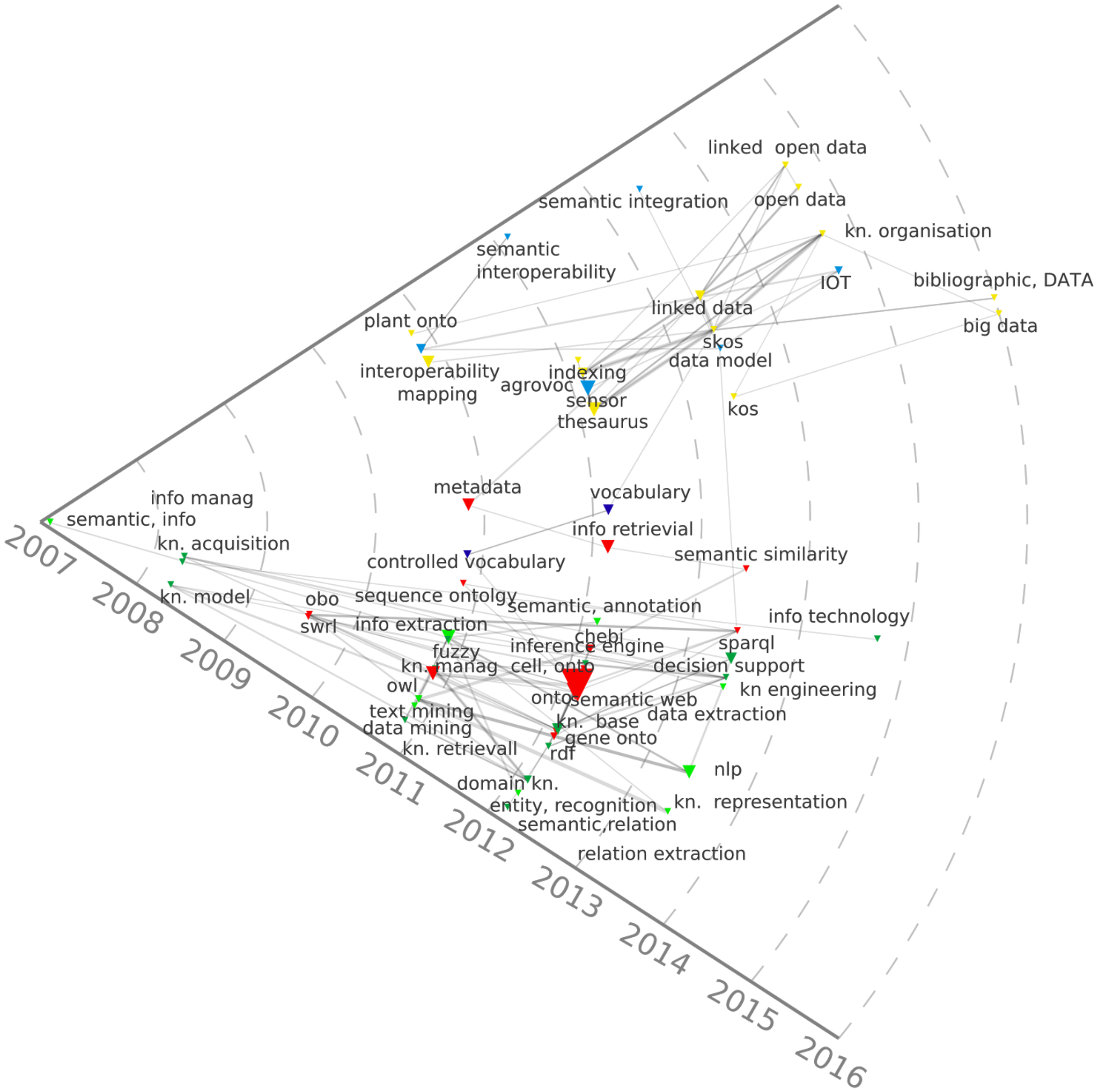


Figure 3: Historical graph of semantic topics

In Fig. 4, 5 and 6 thematic clusters[27] are presented in two dimensions, mixing semantic resources and topics (round marks, upper case terms) and agriculture and food topics (triangle marks, lower case terms). Groups of actors like universities and institutes are linked to the thematic clusters, giving an overview of who works on what, with whom.

The analysis reveals that the actors are linked according to various logics: an agricultural object of interest, e.g. proteins or crops; a type of data, e.g. sensors; a kind of application, e.g. information extraction; semantic objects, e.g. ontologies or metadata; also geographical.

[27] JPG files: https://www.rd-alliance.org/landscaping-material

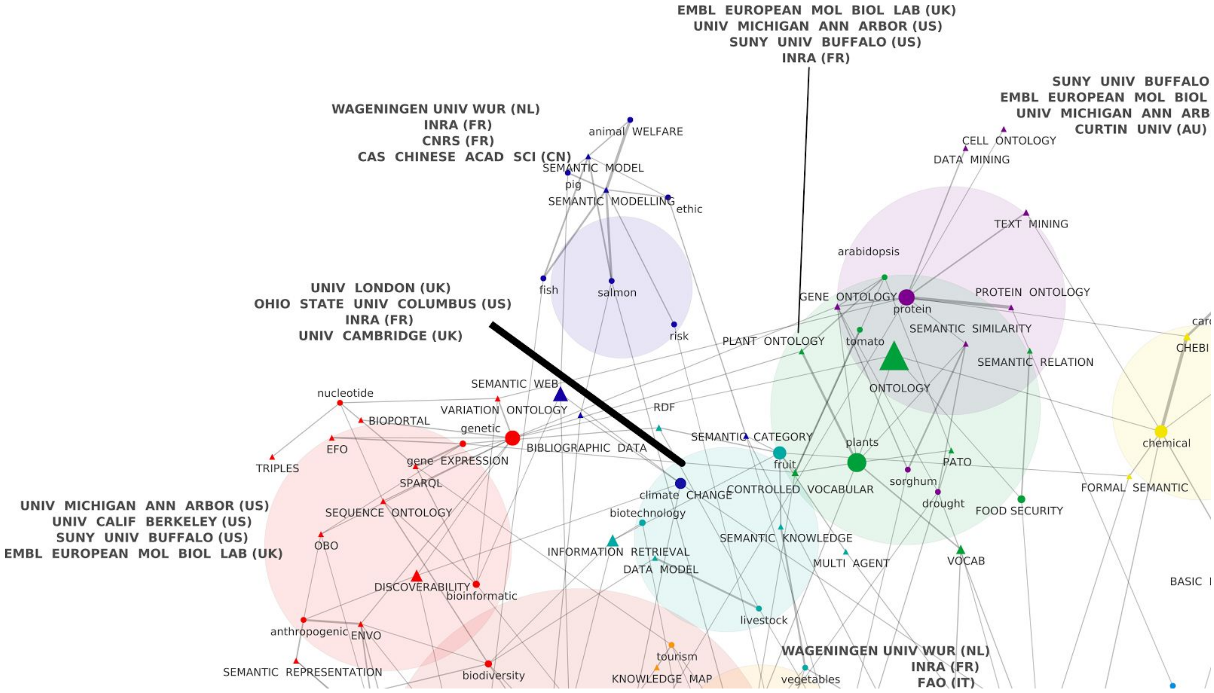


Figure 4: Thematic graph - the plant/biotech clusters

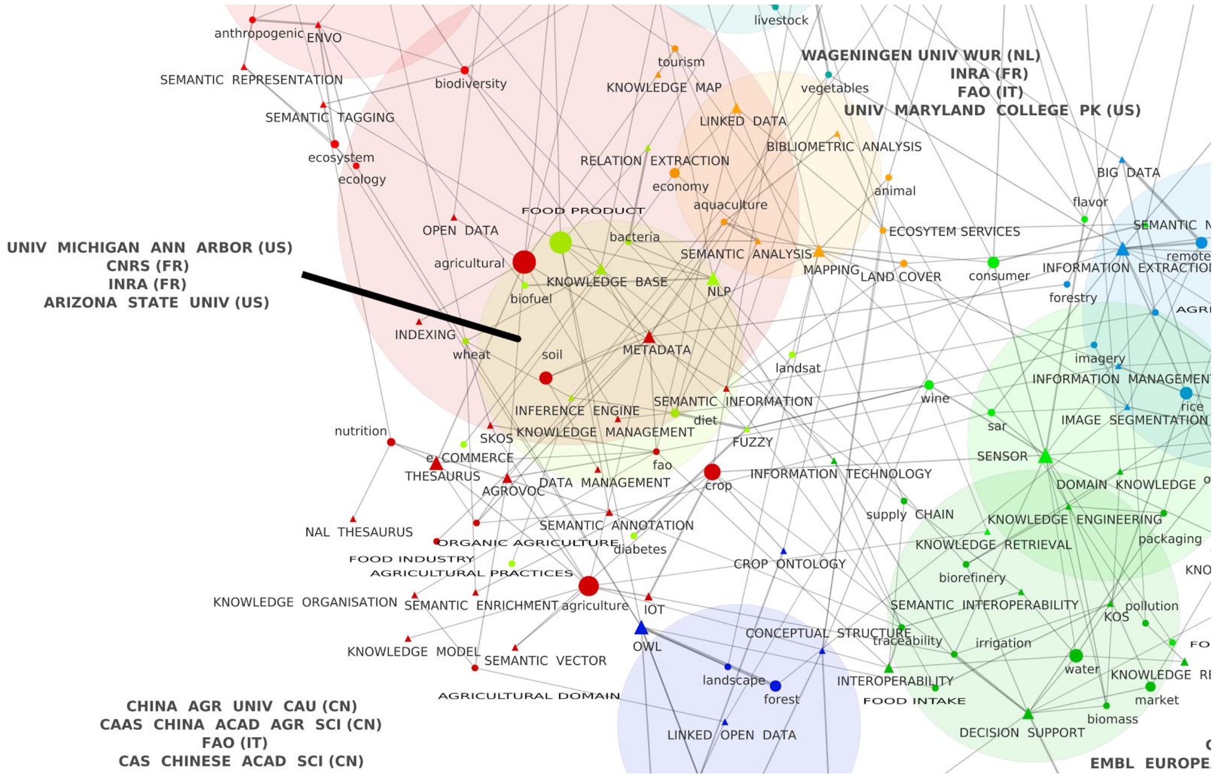


Figure 5: Thematic graph - the information management clusters

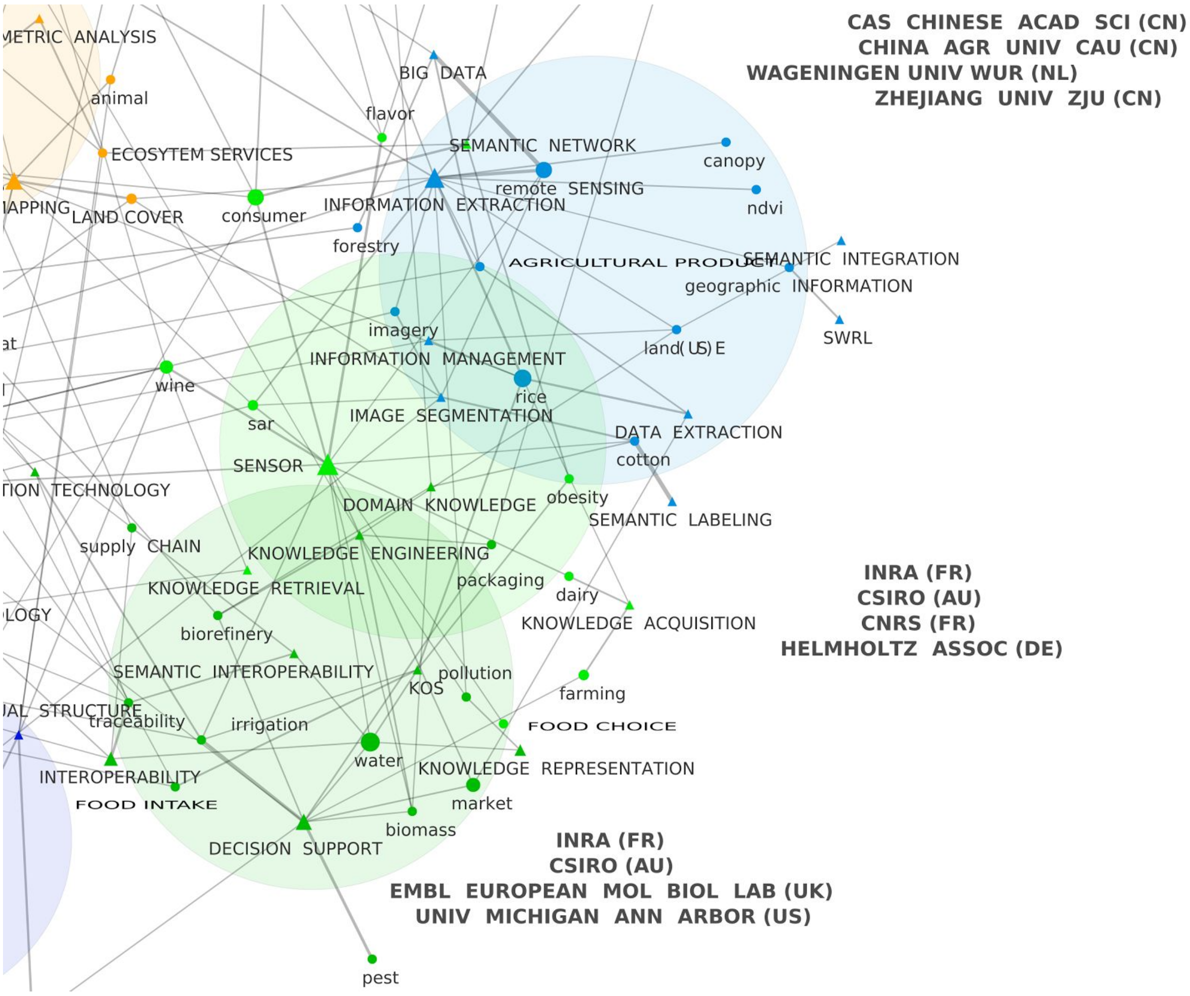


Figure 6: Thematic graph - the sensors / big data clusters

The evolution of topics as well as the variety of actors and types of their relations denote the existence of multifaceted and dynamic community which is linked to and benefits from the influence of many disciplines.

3.2 Publishing research results

3.2.1 Journals

Looking at publishing habits, this study shows no existence of a journal dedicated to semantics for agriculture and food comparable to the "Journal of Biomedical Semantics" which appears first in the ranking as quite a number of semantic resources used in agriculture come from the biomedical community.

Interestingly, open access is an important tendency in the community with 12 of the 14 most publishing journals in the area as green journals. A green journal allows self-archiving by the authors in either institutional or disciplinary open access repositories. This means that papers are potentially accessible for free in either their post-print (noted G in Table 1), preprint (noted G) or publisher's (noted G) version immediately after their publication. The real amount of open access papers depends on whether authors actually use this self-archiving mechanism or not.

Rank	Openness	Name of journal	N. of publications over the last 10 years	Percentage in corpus
1	G	Journal Of Biomedical Semantics	33	2.26%
2	G	Nucleic Acids Research	31	2.13%
3	G	BMC Bioinformatics	29	1.99%
4	G	Plos One	24	1.65%
5	G	Journal Of Integrative Agriculture	14	0.96%
6		Bioinformatics	13	0.89%
7	G	Computers And Electronics In Agriculture	12	0.82%
7	G	Food Quality And Preference	12	0.82%
7		Spectroscopy And Spectral Analysis	12	0.82%
8	G	Database-The Journal Of Biological Databases And Curation	11	0,75%
8	G	Communications in Computer and Information Science	11	0.75%
9	G	Environmental Modelling & Software	10	0.69%
9	G	Journal Of Biomedical Informatics	10	0.69%
10	G	Ecological Informatics	9	0,62%

Table 1: Main journals with Open Access status according to Sherpa/Romeo (http://www.sherpa.ac.uk/romeo/search.php)
G : green ; G : green BUT author cannot archive post-print; G : green BUT author cannot archive publisher's version/PDF

3.2.2 Conferences and workshops

Conferences and workshops are opportunities for scientists and engineers to meet and exchange important information, ideas and good practices. They are often at the origin of fruitful collaboration and innovation.

Table 2 that shows the most popular conferences and workshops could suggest that there is no major event explicitly dedicated to semantics in agriculture. But having a closer look at the MTSR conference ranked first and described as “an annual international interdisciplinary conference which brings together academics, researchers and practitioners in the specialized fields of metadata, ontologies and semantics research.” we see that it includes a special track on Metadata and Semantics for Agriculture, Food & Environment (AgroSEM'17). This event has become an important forum for researchers and other experts to meet and discuss.

Rank	Name of event	N. of publications	Percentage in corpus
1	Conference on Metadata and Semantic Research (MTSR)	37	6.13%
2	International Conference on Computer and Computing Technologies in Agriculture (CCTA)	19	3.15%
3	IEEE International Geoscience and Remote Sensing Symposium (IGARSS)	13	2.15%
4	International Conference on Computational Science and Its Applications (ICCSA)	9	1.49%
5	International Conference on Agro-Geoinformatics (Agro-Geoinformatics)	8	1.32%
6	International Conference on Computing for Sustainable Global Development (INDIACom)	7	1.16%
6	International Conference on Geoinformatics (Geoinformatics)	7	1.16%
6	International Semantic Web Conference (ISWC)	7	1.18%
7	International Conference on Ambient Systems, Networks and Technologies (ANT) / International Conference on Sustainable Energy Information Technology (SEIT)	6	0.99%
7	International Conference on Formal Ontology in Information Systems (FOIS)	6	0.99%
8	European Semantic Web Conference (ESWC)	5	0.83%
8	International Conference on Semantic Computing (ICSC)	5	0.83%
8	International ISKO Conference	5	0.83%

Table 2: Main conferences (604 papers published in conferences and workshops)

3.2.3 Data article for semantic resources?

With a particular interest in how semantic resources are spread and advertised, we looked for publications that specifically aim at describing a semantic resource[28] at the time it is released or updated (similarly to an article would describe a dataset), namely a data paper or data article. That kind of article is rare in the corpus, with some examples like (Ilic et al. 2006), (Rodriguez-Iglesias et al. 2013) and (Çağdaş & Stubkjær 2015). Other articles include a description of the resource but also tackle other issues like applications of the ontology, or

[28] We looked at publications that have either “ontology”, “thesaurus”, or “vocabulary” in their title.

research issues like *“The Gene Ontology project in 2008”* which informs on the Gene Ontology and Sequence Ontology last improvements but also on the annotation of reference genes and related tools. Such publications provide valuable information on the motivations, approach, construction methodology and the resulting content and structure.

It appears that, in the corpus, 1) only a small number of publications focusses on the resource itself independently of its application or any knowledge engineering issue 2) when they do, they are mostly in or close to the biological and medical domains 3) their structure and length vary greatly 4) they are not written in a resource reuse perspective.
We found no article of “data paper” type among them but this may be a bias of the resource we use. A rapid checking by querying the Web of Science Core Collection specifically on the “data paper” type with either “ontology” or “thesaurus” or “semantic” gave no results and “standard” only 13. Do people only publish semantics resources in data papers?

3.2 Conclusions

In this section, we reported on a bibliographic analysis of research articles in the field of semantics for agricultural and food data, covering a time span of 10 years (2006-2016). Goal of the study was to identify trends, reveal partnerships and know better our community practices.

This exercise presents challenges first because the field is difficult to define: what constitutes the fields of agriculture and food? What should be considered in the semantic perimeter? Such questions lead back to the definition of the scope of Agrisemantics as a community of interest. The work that has been done by group members to put words on general concepts in order to build the query contributed to clarify the vision. The corpus resulting from the constructed query tends to be as representative as possible of the field with yet inevitable biases related to the composition of the expert group who defined it and the choice of the source, i.e. the Web of Science.
The second challenge lies in how to represent the data. The graphs are issued from classical spatialisation algorithms for the detection of huge networks of actors and topics, based on their co-occurrences. Using them, we could produce the “big picture” of who works on what, but their presentation as pictures in a document limits the interpretation to tendencies and does not allow the reader to focus on parts of it and go deeper or back to the data.

This said, the analyses produced from the corpus of scientific publications lead to some general observations.
There is a growing interest for semantic approaches in the agriculture and food research community with a noticeable shift from information to data management. This confirms the observation mentioned in section 2 of the mutation in the use of thesauri and reference vocabularies not only for information retrieval purposes but now as collections of identities for the semantic web. This mutation, which is driven by sensor and high throughput data in particular, implies new research issues as well as new challenges in terms of efficiency and usability of semantic technologies and tools.

Actors are varied in terms of geographic location - semantics for agriculture and food data is a global concern; scientific disciplines and objects of study from protein to soil - with clear links and influence from the biomedical domain; application purposes from information discovery to decision support; types of data from scientific texts to data collected by farm equipment. While key players of the public research in agriculture and food research are shown to be active on various topics, it is not possible to identify from this study which are the main actors on the semantics side. The main reason is probably that for universities in particular, the identification of actors at the institutional level prevents from knowing if they work on agricultural or semantics issues. However, journals and conferences clearly show the integration of the agriculture and food community in the computer science and semantics research ones.

Finally, if many semantic resources are produced in the context of research activities, they are still too rarely considered as research products and promoted and published as such. Describing only the applications or research works they were developed for does not encourage their public release neither is sufficient to make them easily reusable by other actors. Too often, they remain private and low documented.

In a few words, the community of semantics for agriculture and food data is global, lively, and rich from many influences and collaborations. Future research works and projects could gain from more interactions between topic or local groups as well as from shared practices, standards and facilities.

4. Semantic resources in the agricultural domain

In this chapter we present an overview of the available semantic structures suitable for agricultural data. The first question to address then is - Where to find available semantic structures for agricultural data? There are a few existing services that can help outline the current landscape in this area. In this section, we focus on two initiatives, the "Map of data standards for food and agriculture" (for brevity, from now on “the Map”) and the AgroPortal, where the Map is a catalog only, and the AgroPortal is a repository supporting a number of functionalities for data maintainers and users. In Sec. 5.5, we give an extended overview of repositories, registries and catalogs of semantic resources relevant to agriculture.

The "Map of data standards for food and agriculture"[29] is a catalog of “data standards” created and maintained under the GODAN Action project[30], and builds on two existing efforts, the VEST Registry maintained by FAO and the AgroPortal maintained by the University of Montpellier. The Map is directly linked with the AgroPortal repository, which gives a more precise idea of the semantic interoperability of the featured standards. The AgroPortal[31] is a repository of ontologies and knowledge organization systems related to agriculture and neighbouring domains. It offers facilities to host, search, version, visualize, comment, and recommend and test semantic structures, as well as generate, store and

[29] http://vest.agrisemantics.org

[30]http://www.godan.info/news/godan-action-enabling-practical-engagement-open-data-agriculture-and-nutrition

[31] http://agroportal.lirmm.fr/

exploit ontology alignments. It also offer a text annotation service. These two resources are dynamic in nature - the figure reported here refer to data available as of June 2017, when this study was carried out.

4.1 A variety of “types” of semantic resources

As mentioned in Sec. 1, a large variety of resources are used for the purpose of describing data, from vocabularies to ontologies, through taxonomies, thesauri and glossaries, to mention only a few. Sometimes their name is meant to highlight their primary use and purpose, or structural features, or the formats used for encoding, or simply to reflect different conventions in different communities. Although we call them all “semantic resources” because they all serve the same purpose, classifying them by type can be useful because each type has different features that make it more or less suitable for certain uses.

To give an idea of the variety of different types of semantic resources around, we use a list compiled by the Dublin Core NKOS working group[32], reported in Table 3 below. The list focuses on value vocabularies (see Sec. 1). It should be clear from the list that many of listed “types” may be identical in abstract structure (say, lists) but different in terms of domain coverage (e.g., gazetteers and name authority lists), or specialization and amount of details included (e.g., dictionary and glossary). Sometimes, a mixture of historical flavor and format is key to understand their difference (e.g., semantic networks and ontologies). Also, resources that are identical in structure may be difficult to distinguish (e.g., categorization schemes, classification schemes and taxonomies) without considering their area of application or the the preferences in nomenclature of specific communities.

Table 3 : list of KOS vocabularies, as identified by the Dublin Core NKOS Working Group

Name	**Description**
categorization scheme	loosely formed grouping scheme
classification scheme	schedule of concepts and pre-coordinated combinations of concepts, arranged by classification
dictionary	a 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

[32] https://github.com/dcmi/archive/blob/master/mediawiki_wiki/NKOS_Vocabularies.md#kos-types-vocabulary

glossary	a collection of textual glosses or of specialized 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 kin
name authority list, **aka authority file**	controlled vocabulary for use in naming particular entities consistently
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)
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 organize items of knowledge or information
terminology	set of designations belonging to one special language
thesaurus	controlled and structured vocabulary in which concepts are represented by terms, organized so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms

Also for description vocabularies (i.e., lists of metadata elements) there is a certain variety of terminologies, as illustrated in Table 4 below.

Table 4 : list of vocabularies for metadata elements

Name	**Description**
metadata element set, aka schema	any set of metadata elements, like XML schemas, RDF schemas or less formalized 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	sometimes considered as sophisticated schema

The attentive reader will have noticed that “ontology” appears in both Table 3 and Table 4. This reflects a relatively common use of the term, not shared by the authors of this document. The confusion probably stems from the fact that an ontology typically consists of a set of axioms defining the classes of interest, and a (typically larger) set of instances of those classes. However, it is becoming more and more accepted and even encouraged to keep ontologies and instances apart.

A separate discussion deserves the formats in which semantic resources are made available (and other technical features that we will see in Sec. 4.2). Formats[33] themselves are not a simple and linear way of classifying semantic resources, as the “format” of a resource is in the end a combination of a file format (binary, text and all the sub-formats), a structural format (e.g., CSV, TSV, XML, Json, ttl, n3), a grammar framework (e.g., RDF, independently of its serialization), a grammar framework coupled with a vocabulary (e.g., RDF SKOS, Json LD SKOS, or OBO XML and OBO flat file). All the components of the format matter, because interoperability can be achieved at the level of the structural format (an application that reads Json will expect a Json format), at the level of the grammar (a semantic application that understands RDF will probably not care about the structural format) or at the level of the semantics (an application that understands RDF SKOS may not understand RDF OWL).

Some formats are tied to a specific type of semantic resource (e.g. RDFS is normally used for schemas, OWL for ontologies and sometimes for thesauri with added ontological features, SKOS for simple thesauri and classifications), but there is not an exact 1-1 relationship, so each semantic resource is best defined by a combination of type of resource plus its format.

[33] See also the list of URIs for file formats as given by the W3C: https://www.w3.org/ns/formats/

4.2 An investigation on the content of the Map and the AgroPortal

In order to give an idea of the distribution of existing semantic resources across these types and formats, we have looked at the content of the Map (excluding peripheral and generic vocabularies). In particular, we have considered ontologies, thesauri, glossaries, schemes and classification schemes. Labels are assigned by the authors registering data to the catalog. We leave to a further exercise to assess how consistently these labels are applied with respect to the structural features and format of implementation.

Ontologies = 81, of which 63 are maintained in AgroPortal
The 63 maintained in the AgroPortal are shared resources, in the sense that all classes and properties have URIs and many concepts have links to concepts in other vocabularies. Examples are the various crop-specific ontologies that are part of the CGIAR Crop Ontology (Chickpea Ontology, Banana Ontology, the GIAR IBP crop-specific Trait Ontologies and also the FAO/IPGRI Multi-Crop Passport Descriptor formalized as an ontology), other ontologies from the OBO family (like the Gene Ontology), the Planteome Plant Trait Ontology. A few ontologies are in the domain of animal health: the INRA Animal Disease Ontology and the INRA Animal Trait Ontology for Livestock. A few new ontologies hosted in the AgroPortal are in the domain of food: FOODON (also of the OBO family) and FOODIE (extending the INSPIRE data model for Agriculture and Aquaculture Facilities theme).

Thesauri = 27
Besides the major broadly-scoped thesauri (AGROVOC, the CABI Thesaurus, the US NAL Thesaurus, the Chinese Agricultural Thesaurus, all available as SKOS), there are several domain-specific thesauri. Only 5 of the thesauri present are expressed in a format recommended for the web (SKOS XML/RDF, Turtle, Json-LD): the US FDA Langual Thesaurus (International Framework for Food Description), the GSSoilSoilThes, the Water Quality Planning Bureau Library Thesaurus, the LandVoc (the Linked Land Governance Thesaurus managed by the Land Portal) and the INRA Thesaurus for Animal Physiology and Livestock Systems maintained in the AgroPortal. Others are more traditional thesauri available as PDFs and/or HTML pages, like the Finnish Agriforest Thesaurus, the FAO Aquatic Sciences and Fisheries Thesaurus (ASFA), the IRRI Rice Thesaurus etc.

Glossaries = 17
Most glossaries remain in traditional formats (text, PDF) and at best have been made available as HTML pages. Examples are the FAO glossaries part of FAOTERM (the Glossary of Aquaculture, the Fisheries Global Information System Glossary...), the FAO Glossary of Biotechnology for Food and Agriculture, the FAO Glossary of Phytosanitary Terms, the FishBase Glossary, the US EPA Glossary of Climate Change Terms, the IUFRO/FAO Multilingual Glossary of Forest Genetic Resources.

Schemas = 15
In some domains, schemas have been developed especially for representing and exchanging data about specimens or observations in natural science. Most of them are XML

schemas (like the GBIF Access to Biological Collection Data Schema, the Fisheries Metadata Element Set, the schemas of the INSPIRE data specifications for Soil and for Agricultural and aquaculture facilities, the OGC Geoscience Markup Language), a couple are RDF schemas (Darwin Core for germplasm, the agINFRA Soil Vocabulary) and a few are just a set of descriptors (like the FAO/Bioversity Multi-Crop Passport Descriptors - which have been translated to an ontology, see above -, the Bioversity Core descriptors for in situ conservation of crop wild relatives). Some schemas for farm management data have been developed by the AgGateway consortium, but most of them are only accessible to members.

Classification schemes = 13

Several classification schemes have been developed by experts in different domains, normally for classifying species and organisms or commodities and products. They also remain in traditional formats (text, PDF). Also in this case a good number has been produced by authoritative bodies like FAO and in some cases their use is prescribed in international information systems and official statistics.

Examples are the FAO Fisheries Commodities Classification, the WCO Harmonized Commodity Description and Coding System, the FAO International Standard Statistical Classification of Aquatic Animals and Plants, the UN Central Product Classification, the IUFRO Global Forest Decimal Classification, the Australian Soil Classification (ASC).

Semantic resources do not only differ in structure and format: in many cases they are specifically designed to model a certain type of entity/data. This is where semantic resources become relevant to specific domains. Some semantic resources are domain-agnostic and of general use. Consider for example the Dublin Core, defined to cover basic information valid to any type of information object; DCAT, to describe any type of datasets for data catalogs. Also some of these types, even though specific, may be not linked to a specific domain but relevant for many domains including most of the agricultural domains: a typical example are semantic structures designed for observations. Many of these structures cater for any type of observation and only in some cases more specific structures are developed for specific types of observation (e.g. weather observations, soil observations, crop growth observations...)

Here are some representative semantic structures for some very specific types of data from the Map of Standards:

- Out of 191 semantic resources, 111 are meant for types of data that belong to the general category of “Research and agronomic data”, of which the most populated sub-category is “**Plants / germplasm**” (66 resources), for which the following specific types of “things” or data(sets) can be found (as defined by the experts behind the AgroPortal): crops, general germplasm, germplasm accessions, location and environmental, phenotype and trait, plant anatomy and development, structural and functional genomic.

  There are 27 semantic resources suitable for the crops “things” and for the **phenotype and trait** datasets (besides the CGIAR crop ontologies under the Crop Ontology, there are specific ontologies like the Phenotypic Quality Ontology, the Plant Trait Ontology, the INRA Wheat Trait Ontology...) and 23 for **plant anatomy**

and development datasets (the Banana Anatomy ontology, the CGIAR IBP Wheat Anatomy and Development Ontology...).

- For **weather / meteorological data**, there are different types of datasets that in some cases may require or use different semantic structures: climate data, weather forecasts, weather monitoring infrastructure.
  An interesting case is the case of **weather observations**: first of all, it appears that most of the semantic structures used for weather observations are the same used in general for all types of observations and measurements; secondly, it appears that the most used of these structures have almost no semantics and are in the form of data specifications prescribing binary or textual array-based or tabular formats. An example of both tendencies is the NetCDF format, binary, array-based and for generic observations. More specialized for weather data but still binary are the WMO BUFR and GRIB formats and related code lists. A step ahead in the direction of more semantic structures are some XML schemas, still for generic observations, like the OGC "Observations and Measurements - XML Implementation" or the OGC "Timeseries Profile of Observations and Measurements". More specifically for the "climate data" type, an XML model based on the ISO and OGC feature type approach has been developed and is called "Climate Science Modelling Language", which is now the basis for the INSPIRE Data Specifications for Atmospheric Conditions/Meteorological Features and Oceanographic Geographical Features.

4.3 Laying the ground for an assessment of semantic resources

An overview of existing semantic resources should also aim at supporting potential adopters in evaluating the suitability of each standard for their needs as well as at providing a qualitative benchmark and gap analysis for providers and decision makers. Categorizing resources by type and domain helps, but more precise qualitative information should be provided about resources.

The above mentioned first "Gap exploration report" prepared in the GODAN Action project lays the ground for an assessment and analysis of gaps in the availability of data standards for food and agriculture. The evaluation framework of the map of standards was built on two existing frameworks: the assessment process used by the UK Government's Open Standards Board[34] and the ODI Open Data Certificates criteria[35] (applicable to standards published as open vocabularies). The criteria used in the gap analysis report were selected based on the assumption that users of standards need above all standards that are:

1. **fit for purpose =** compatible with other standards, scientifically sound, complete
2. **endorsed, adopted and authoritative** = therefore, a lack of standards in a domain is obviously a big gap, but also having a plethora of overlapping standards covering the same domain or even the same entities is a problem to be solved
3. **usable** = available in various forms, including widgets; integrated in tools; managed on a collaborative platform

[34] https://standards.data.gov.uk/assessing-standards-proposals

[35] https://certificates.theodi.org/en/

4. **open and interoperable** = especially for developers; therefore, standards that are only available as PDF or UML models represent a gap to be filled

All the criteria were translated into “questions” that were added as metadata to the Map.
The detailed assessment criteria used are published in the Map[36]. Some of the criteria used for assessing the openness and usability of standards can be of particular interest. In particular, they ascertain whether the standard is:

1. **Versatile**
   Is the standard available in different formats for different technologies? (e.g. XML, JSON, RDF)?
2. **Served by APIs**
   Are there APIs and web services that allow applications to:
   - Lookup terms / concepts using several parameters
   - Perform cross-walks between vocabularies
   - Extract / lookup subsets of vocabularies
   - Automatically “tag” with the vocabulary (term extraction plus advanced NLP in the case of text, other types of reasoning…)
   - Get more user- or web- friendly results (json, widgets…)
3. **Machine-readable**
   Is the standard available in machine-readable formats?
4. **Meaningful**
   Is it serialized in appropriate vocabulary format / semantics? (OWL, SKOS, OBO...)
5. **Referenceable**
   Does it use dereferenceable URIs (URLs) as identifiers of classes, properties and instances?
6. **Linked**
   Is it available as Linked Data? Does it actually link to other resources?
7. **Annotated**
   Is it accompanied by machine readable metadata?
8. **Clearly licensed / openly licensed**.

The report sketches some preliminary conclusions, drawn from the analysis of the metadata provided in the Map. Here below we report those that we consider relevant to the present work, and expand them further.

1. The number of resources in machine-readable format is low. Only 55% of the standards are presented in machine readable formats (40% with some semantics) - 29 out of 97 are from the plant sciences domain and 56 from the broader research and agronomy domain. Many are not even available on the web (16%).

2. **Licenses are often not stated**. Most resources fail to present a clear licence (only 21%) though where they do they are generally open (13%).

[36] http://vest.agrisemantics.org/content/assessment?qt-assessment_tabs=0#qt-assessment_tabs

3. **Resources are often not documented**. There is a gap between the information presented on the web and the documentation on the standard, with only 31% of the standards having documentation, only 5% having tests and only 40% being supported.

4. **Few APIs are available**. According to the assessment metadata in the map of data standards, excluding geospatial data standards that are at the same time peripheral and cross-cutting, around 41% of the featured resources are served by APIs: of these, 43 out of 79 are from the plant sciences domain. This is problematic as most of the uses described in Sec 2 require either the provision of an API (from a SPARQL endpoint to a REST API or any other type of web service) or at least an openly accessible machine-readable version of the semantic resource. In practice then, not many of the existing semantic resources offer the level of openness and usability required to implement the functionalities described there. Moreover, when this happens, they either cover specific domains, mostly plant sciences and geospatial information, or are broad general vocabularies that span across all agricultural domains. Moreover, even among resources that are served by APIs, the types of APIs are very different: on 191 resources in the narrower scope of agriculture, 64 have an API to automatically annotate text or data, 51 an API to lookup terms / concepts using several parameters, 47 to get web- or user- friendly results or perform cross-walks between vocabularies, 35 to perform cross-walks between vocabularies and only 1 to extract / lookup subsets of vocabularies.

5. **Big differences among domains.** It appears that certain domains are better covered than others (plant sciences above all, followed by natural resources).
More in details:

1. Most of the standards used in plant sciences are in the form of ontologies (therefore real semantic resources) and are highly open and usable.
2. Although at the level of domain “Natural resources” seems an area well covered by standards (also with a certain degree of openness), this is mostly due to the high number of geospatial semantic resources, while the level of openness and usability of standards in sub-domains of natural resources is lower and still varies. For example, good standards (widely adopted models, thesauri) exist in the soil domain, but in most cases they are not formalized as open standards. On the contrary, the land sector lacks widely accepted classifications and the existing ones are mostly on paper (with a subsequently low level of openness and usability) and only one open standard has been developed.
3. In the area of administration and legislation data, including official records and government finance data (which accounts for a huge amount of data potentially available and relevant for impact on farmers and industry), there seems to be a poor level of standardization. For some data, generic statistical standards apply, but there is little adherence to domain-specific semantics.
4. Value chain data is an area where standardization is picking up, with 11 standards of different types (from ISO standards to international product classifications to “messaging standards” to ontologies). However, most of these standards are either regulatory (mandated by governments or international bodies) or syntactic: there is very little reference to common semantics.

   Also in the area of food (interlinked with the supply chain area), standardization is clearly picking up, for instance in the area of food components / nutrients and ontologies to support diet and recipe applications.

These are only very broad initial insights. This work only laid the ground for further analyses that can be conducted based on the assessment metadata embedded in the map of standards. Better insight on the level of interoperability of a certain number of standards (of the ontology types) can probably be gained by extracting data from the AgroPortal repository of ontologies, which collects detailed information on formats, mappings, APIs and more.

4.4 Conclusions

In this section we have reported on our analysis of the current situation of semantic resources available in the area of agriculture. Our work was based on two initiatives in the area: the "Map of data standards for food and agriculture" (for brevity, “the Map”) and the AgroPortal, the former offering a comprehensive catalog of semantic resources in the area of agriculture, the later being a repository offering a wide range of facilities useful to both data maintainers and data users. To our knowledge, those are the only initiatives specifically addressing to the domain of agriculture. In Sec. 5 we provide a list of other similar initiatives, focussing on different areas.

We started off with a discussion of the various types of resources commonly falling under the label of semantic resource, in particular for value vocabularies (aka KOS). The rationale for such a discussion is to help non-experts in semantics find their way in the large variety in terminology used. We highlighted the fact that resources with the same structural features (say, lists) may be named differently depending on other non-structural characteristics such as the domain of coverage or the amount of details they include.

We also reported on our analysis of the Map and the AgroPortal. We found a large number of ontologies, followed by thesauri, glossaries, schemas and classification schemes, summing up to the majority of the types identified in the Map (the classification of a resource as one or another type depends on the authors submitting the data to the catalog). We notice that while ontologies are all expressed in the expected, W3C-promoted formats, namely either OWL or OBO (or maybe it is the other way around, that any resource in OWL or OBO is expected to be an ontology), only a few of the thesauri listed are encoded in formats suggested for the web. The situation is even worse for glossaries and classification schemes, while schemas are mostly in XML. The area of plant sciences seems is the one for which the most resources have been developed.

Based on the work done within the GODAN Action project, we listed a few criteria to assess the quality of semantic resources. Those criteria were used to extend and improve the metadata collected in the Map and the metadata analysed to get a general picture of the quality of the semantic resources available. We found that not only documentation is largely missing, and licenses of use often not stated (both somehow to be expected), but also that nearly half of the resources used to describe agricultural data are not even available online, many are only in pdf, and less than half have APIs. The field where most effort is made and visible is plant science, georeferenced data also are in general in quite good shape, while other areas still quite behind, e.g., classification schemes for statistics, and administration and legislation data. In the area of value chain and food, standardization is picking up.

We have not investigated the level of overlap among those resources, nor the possibility of aligning them so as to reduce duplication of work and enhance interoperability. This could be done in a further study. Further analysis could also try together evidence of the actual contribution to data interoperability provided by these resources, i.e., how much they are reused in different systems.

Anecdotal evidence suggests that the difficulty of finding about a semantic resource to reuse is one of the causes of the proliferation of resources discussed in Sec 2. In this respect, we look positively at initiatives such as the AgroPortal and the Map.

5. The Semantic Expert Toolkit

The goal of this chapter is to take stock of what is available in terms of online services or desktop tools for facilitating producers and users of semantic structures for accessing, building, managing, sharing and reusing semantic resources. More specifically, this chapter focuses on tools and services for the following activities: editing (Se. 5.1), visualize and produce documentation (Sec. 5.2), perform quality assessment (Sec. 5.3), establish alignment between semantic resources (Sec. 5.4). Finally, we survey existing repositories, registries and catalogues for sharing semantic structures (Sec. 5.5), and draw some conclusions (Sec. 5.6).

5.1 Editing Tools and Services

A rather large number of tools and services are available for editing semantic resources, and various attempts have been made to compile lists of those tools and services. Lists like those are important to facilitate the choice and adoptions of the right tool, and so spread the use of semantic resources. It is also important that these lists are regularly updated to keep track of the evolutions in the area. The RDA Vocabulary Services Interest Group has compiled a list of tools for editing thesauri[37]. Other lists of software for the management and editing of controlled vocabulary are also available online[38], with a certain number of features evaluated. Comparative analysis have also been published as scientific papers, as in (Stellato et al. 2015), where VocBench is compared with the currently most popular tools WebProtégé, Poolparty, TemaTres, and SKOSed.

The list of tools and services to edit and manage semantic resources is long, which poses problems to users having to choose the one best fitted to her needs, skills and technical

[37] “ANDS appraisal of thesaurus software tools”
https://www.rd-alliance.org/ands-appraisal-thesaurus-software-tools.html

[38] See for example
https://www.google.com/url?q=https://docs.google.com/spreadsheets/d/1AWKY86xLCBB0b50g2TXcG0iwIHcN6foSwrvW_LXPpvk/edit%23gid%3D0&sa=D&ust=1489657843094000&usg=AFQjCNGQy3o_BIgo2aiKK6feAsTEKrAeew

constraints. Having no clear guides often leads to investing energy and time successively in several tools with the risk of data loss or degradation and discouragement.

In this section, we provide a short list of key tools and services for the building and management of semantic structures. This tools in the list are recommended, based on the experience of the authors. For each of the presented tools and services, the name of the tool/service is provided as well as a general description, its business model, and the URL for accessing it.

Table 5: List of Editing Tools and Services

Title	**Description**	**Business model**	**URL**
For schemas (RDFS)			
Neologism	A web-based RDF Schema vocabulary editor and publishing system (not for ontologies nor SKOS vocabularies)	open-source	http://neologism.deri.ie/
For lists and hierarchies (SKOS, RDF)			
iQvoc	A SKOS(-XL) vocabulary management system for the Semantic Web	open-source	http://iqvoc.net/
SKOS Shuttle	A collaborative multi user / multi tenant thesaurus management system with an RDF store	commercial with limited free account	https://skosshuttle.ch/
SKOSjs	A JavaScript based editor for Simple Knowledge Organisation System (SKOS) data	open-source	https://github.com/tkurz/skosjs
ThManager	A tool for creating and visualizing SKOS RDF vocabularies.		http://thmanager.sourceforge.net/
For ontologies (OWL, OBO)			
Cognitum Fluent Editor	A collaborative ontology editor, interoperable with Protégé and R. It exports in RDF-S and OWL.	free for academics and researchers for non-commercial use	http://www.cognitum.eu/semantics/FluentEditor/
Obo-Edit	An open-source ontology editor using the OBO format	open-source	http://oboedit.org/

OWLGrEd	A free graphical ontology editor. It supports exports in OWL and is interoperable with Protégé.	free for use	http://owlgred.lumii.lv/
Semafora OntoStudio and OntoServer	A proprietary framework for building, managing, and mapping ontologies.	commercial	http://www.semafora-systems.com/en/products/ontostudio/
Protégé	A free, open-source ontology editor. It also offers a web environment, Webprotégé	open-source	http://protege.stanford.edu
For all kinds of semantic resources			
ITM (Intelligent Topic Manager)	A proprietary solution for creating, managing, and linking taxonomies, vocabularies and metadata.	commercial	http://www.mondeca.com/itm/
PoolParty	A proprietary Web based editor for taxonomies, thesauri, and ontologies utilizing Linked Data.	commercial	http://www.poolparty.biz/
TemaTres	An open-source vocabulary server, web application to manage and exploit vocabularies, thesauri, taxonomies and formal representations of knowledge.	open-source	http://www.vocabularyserver.com/
TopBraid EVN	A commercial web-based platform for creating and managing semantic structures	commercial	http://www.topquadrant.com/products/topbraid-enterprise-vocabulary-net/
VocBench	A web-based, multilingual, collaborative development platform for managing OWL ontologies (soon), SKOS(XL) thesauri and generic RDF datasets.	open-source	http://vocbench.uniroma2.it/

Even if the above mentioned tools are not all based on standards and thus interoperable, there is a real convergence trend towards implementing W3C recommendations and Semantic Web standards.

5.2 Visualization and Documentation Tools and Services

Semantic resources can be difficult to maintain and document if they are large. Besides, standard formats like OWL or SKOS are not dedicated to human reading. Editing tools can have limited visualization, or unsuitable for non specialist users, and poor or no (semi-)automatic documentation features making use of extra tools necessary. The section provides a summary of key tools and services for the visualization and for the documentation of semantic structures. For each of the presented tools and services, the name of the tool/service is provided as well as a general description and the URL for accessing it.

Table 6: List of Visualization Tools and Services

Title	Description	Business model	URL
Skos Play	A free application to render and visualise thesauri, taxonomies or controlled vocabularies expressed in SKOS.	free service	http://labs.sparna.fr/skos-play/
Skos Reader (Mondeca)	A web application providing HTML publication files for uploaded SKOS files.	free service	http://labs.mondeca.com/skosReader/
Skosmos	An open-source web-based SKOS browser and publishing tool. It is currently used to publish Agrovoc and GACS.	open-source	http://skosmos.org/
AgroPortal Browser	A web user interface to navigate and display ontology and vocabulary concepts in a hierarchy.	open-source	http://agroportal.lirmm.fr

Table 3: List of documentation tools and services

Title	Description	Business model	URL
LODE, the Live OWL Documentation Environment	An XSLT-powered on-line service that automatically generates a human-readable description of an OWL ontology (or, more generally, an RDF vocabulary).	free service based on an open-source development	http://www.essepuntato.it/lode

OWLDoc	A Protégé plug-in which generates JavaDoc-like HTML pages from OWL ontology.	open-source	https://protegewiki.stanford.edu/wiki/OWLDoc
Parrot	A RIF and RDF Ontologies documentation Tool. It provides users with useful reference documentation about rulesets and ontologies expressed in standard languages, such as OWL and RIF.	free service based on an open-source development	http://ontorule-project.eu/parrot/
Widoco	Wizard for documenting ontologies. WIDOCO is a step by step generator of HTML templates with the documentation of your ontology. It uses the LODE environment to create part of the template.	Open source	https://github.com/dgarijo/Widoco/

5.3 Quality Assessment Tools and Services

The quality of a semantic resource can be partly defined by the relevance of its content, namely its coverage, if it is scientifically sounded, etc. Some repositories (see subsection 5.5) like AgroPortal[39] offer recommandation facilities to address this facet of assessment. However, this is not enough to make the resource (re)usable as it is also crucial that its structure and the vocabularies used to implement it meet standards and practices shared within its author's community and among application developers. This is key to build trust and ensure reusability by other people and interoperability with other resources and systems. So assessing the quality is needed when deciding to adopt or align with an existing semantic resource as well as when creating a new one, being it published or not.

Some websites exist that provide pointers to tools and advocates for adopting semantic web best practices, see for example the PerfectO portal[40]. The W3C plays an active role in fostering the creation of major standards for semantics like OWL, and guides like the *Best Practice Recipes for Publishing RDF Vocabularies*[41] published in 2006. Published recently, the *Data on the Web Best Practices*[42] document offers advices on how data of all kinds can be shared on the Web, whether openly or not. It advocates the provision of a variety of metadata, the use of URIs as identifiers and multiple access options with the aim of maximizing data availability, the likelihood of its discovery and reuse. This fully applies to semantic resources whatever their nature and content.

[39] http://agroportal.lirmm.fr/recommender

[40] http://perfectsemanticweb.appspot.com/

[41] https://www.w3.org/TR/swbp-vocab-pub/

[42] https://www.w3.org/TR/dwbp/

Many of the editing tools cited above support standards like OWL or SKOS, but they do not necessarily allow either the producer or the user to check the conformance of a semantic resource with recommendations and best practices. Such checks are especially needed for SKOS as it is a low constrained vocabulary. 29 quality issues have been proposed and documented by (Suominen & Mader 2014). Three of the SKOS quality assessment tools in Table 7 are based on this reference work.

Table 7: List of Key Quality Assessment Systems

Title	Description	Business model	URL
For lists and hierarchies			
skosify	Skosify is a tool that can be used to convert vocabularies expressed as RDFS and OWL into SKOS. It can also be used to improve, enrich and validate existing SKOS vocabularies. An demo version of the tool is available[43], but it does not yet support all the options available in Skosify.	open-source	https://code.google.com/archive/p/skosify/
qSKOS	A tool for finding quality issues in SKOS vocabularies. It can be used as command line tool or API[44]. It is used in Poolparty	open-source	https://www.w3.org/2001/sw/wiki/QSKOS
SKOS testing tool	An online service base on qSKOS that provides a detailed and user-friendly quality report on uploaded or online SKOS and SKOS-XL vocabularies. Many options are offered.	free online service	http://labs.sparna.fr/skos-testing-tool/home?lang=en
PoolParty SKOS Quality Checker	An online service base on qSKOS that provides reports on uploaded SKOS vocabularies in n3, ntriples, rdf, rdfxml, trig, trix, and turtle formats. File size is limited to 100MB.	free online service	http://qskos.poolparty.biz/login
For ontologies			
OOPS! (OntOlogy Pitfall Scanner!)	A web application to identify errors and anomalies in ontologies during their development phase. A RESTFul Web Service is now offered that allows to integrate the pitfall detection in other applications.	free online service	http://oops.linkeddata.es/

[43] http://demo.seco.tkk.fi/skosify/skosify

[44] https://github.com/cmader/qSKOS/wiki/Quality-Issues

5.4 Alignment Tools and Services

Ontology alignment is the process of finding mappings between the elements of different representations. As ontologies constitute a fairly complex conceptualization mechanism, they entail different facets of information which can be used in the context of the alignment task. Consequently, certain alignment techniques work well under certain occasions, while failing over different alignment tasks where the compared ontologies do not carry adequately rich information on the dimension handled by these techniques. As a way to address these issues, most modern systems adopt a composite approach, incorporating different methodologies under an integrated framework. While such systems generally produce better results, their increased complexity impacts their use of resources (computational and temporal), a fact that – in combination with the continuously increase in the size of ontologies - poses one of the major and long-term challenges in the field. Another important facet of the problem is the efficient introduction of the human expert in the alignment process, i.e., how to maximize the usage of expert input while not making the process cumbersome or error-prone for the human agent.

Through the years, research activity on the ontology alignment problem has resulted to the establishment of an extended community which has reached meaningful results using a plethora of approaches on the problem. The multitude of researches and approaches and thus the need to coordinate their actions and progress on the relevant challenges of the field led to the creation and maintenance of the Ontology Matching website[45], which aims to act as a central information resource for the community. The website provides access to relevant publications, events and initiatives around ontology alignment.

The most prominent initiative related to the community is the Ontology Alignment Evaluation Initiative (OAEI)[46]. Its goal is to formalise the development and evaluation of alignment techniques and systems, offering a common base for testing and assessing the performance of matching approaches under a broad range of tasks. Activities of the initiative include the organization of the Workshop on Ontology Matching, usually in conjunction with the important ISWC conference, as well as, the coordination and execution of a yearly evaluation event. Several test sets in the domain of biomedicine are proposed, e.g. "anatomy", "Large Biomedical Ontologies", and more recently "Disease and Phenotype" (human and animals).

The evaluation event provides a set of matching tasks organized into tracks, where the participating systems are evaluated in terms of their precision and recall against reference alignments provided by the initiative. An important outcome of the movement is the production and establishment of standards for representing alignments, namely the Alignment[47] and the more expressive EDOAL[48] formats. Furthermore, the initiative encourages the usage of a specific set of Java libraries, the Alignment API[49], for developing matching algorithms and systems, and their submission to the SEALS platform[50]. The latter

[45] http://ontologymatching.org/

[46] http://oaei.ontologymatching.org/

[47] http://alignapi.gforge.inria.fr/format.html

[48] http://alignapi.gforge.inria.fr/edoal.html

[49] http://alignapi.gforge.inria.fr/

[50] http://www.seals-project.eu/

aims to act as a repository for alignment systems, while also providing a standardized way for deploying and testing the available platforms.

5.4.1 Initiatives / Resources / Standards

International Workshop on Ontology Matching[51] (Shvaiko and Euzenat 2013) provides a benchmark about ontology alignment Interesting research tools with advanced features but not packaged, not documented, lack of sustainability, too difficult to use.
OAEI food track

Table 8 below summarizes key artefacts and information sources commonly used by the Ontology Alignment community to build or extend software, communicate information and reporting on results and systems.

Table 8: List of Key Initiatives / Resources / Standards

Title	Description	URL
Alignment API	An API for expressing ontology alignments. It uses the Alignment Format (http://alignapi.gforge.inria.fr/format.html) and EDOAL ((http://alignapi.gforge.inria.fr/edoal.html) formalizations for representing the produced mappings.	http://alignapi.gforge.inria.fr/
Ontology Alignment Evaluation Initiative	An initiative to provide a standard for the evaluation of alignment techniques. It incorporates multiple evaluation tracks, targeting generic matcher capabilities, as well as, specific challenges in the field (large-scale matching, instance matching, user involvement, etc.).	http://oaei.ontologymatching.org/
Ontology Matching	A web portal for observing developments in the field, relevant events, etc.	http://ontologymatching.org
SEALS	A repository and testbed for ontology alignment systems. The SEALS platform allows the obtaining and execution of multiple alignment frameworks and systems, using predefined benchmarks through the SEALS client, or their application over user-defined alignment tasks.	http://www.seals-project.eu

[51] http://oaei.ontologymatching.org/

5.4.2 Matching Systems

Table 5 presents available systems some of which have participated in the aforementioned Ontology Alignment Evaluation Initiative (OAEI) and achieved notable results. In general, the systems are built to handle the matching of ontologies expressed in OWL, however some systems also handle other specifications and formalisms.

Table 5: List of Key Matching Systems

Title	**Description**	**Business model**	**URL**
AgreementMakerLight	An automated ontology matching system able to tackle large ontology matching problems. It is primarily based on the use of element-level matching techniques supported by background knowledge. It has been used by the GACS initiative and AnaEE research infrastructure..	open-source	http://somer.fc.ul.pt/aml.php
Falcon-AO	An automatic ontology matching tool that has become a very practical and popular choice for matching Web ontologies expressed by RDF(S) and OWL.	open-source	http://ws.nju.edu.cn/falcon-ao/index.jsp
LogMap and LogMap 2.0	A highly scalable ontology matching system with 'built-in' reasoning and diagnosis capabilities. A web front-end is also provided.	open-source	http://www.cs.ox.ac.uk/isg/tools/LogMap/
Onagui (Ontology Alignment Graphical User Interface)	An alignment helper and viewer. User can edit alignments between concepts of two semantic resources either in SKOS or OWL. Several algorithms are provided for discovering candidate alignments.	open-source	https://sourceforge.net/projects/onagui

PARIS	A probabilistic matching system, based on the analysis of common relations between the compared ontologies and the calculation of the probability that the connected entities are related, propagating these probabilities to cover the entire ontologies.	open-source	http://webdam.inria.fr/paris/
Yam++	An ontology matching tool, which supports discovering alignment of ontologies by either machine learning approaches or generic methods. It supports multilingual alignments. A web tool is provided as Yam++ online as well as an API.	free service based on an open-source library	http://yamplusplus.lirmm.fr/
LOOM	Syntactic mapping component used by NCBO BioPortal and AgroPortal to generate mappings automatically between ontologies uploaded in the repositories.	Open source	Include within NCBO technology (https://www.bioontology.org/wiki/index.php/Category:NCBO_Virtual_Appliance)

Only a few tools deal with the evaluation of alignment strategies, and this is the case of KitAMO and SAMBO[52]. Their evaluation is based on performances on different test cases. Its architecture is based on the SAMBO framework, a tool for aligning biomedical ontologies using external information like thesauri, dictionaries, and instance collections.

5.5 Repositories, Registries and Catalogues for Sharing Semantic Structures

In this section, we list a number of initiatives, be those repositories, registries or catalogs, all aimed at providing a single access point to semantic resources. To our knowledge, the Map and the AgroPortal are the only one specifically related to agriculture.

Name	**Description**	**URL**
Agriculture-centered		
Map of agri-food Data Standards	A **catalog** of metadata about all types of data standards for food and agriculture in all core and neighbouring disciplines.	vest.agrisemantics.org

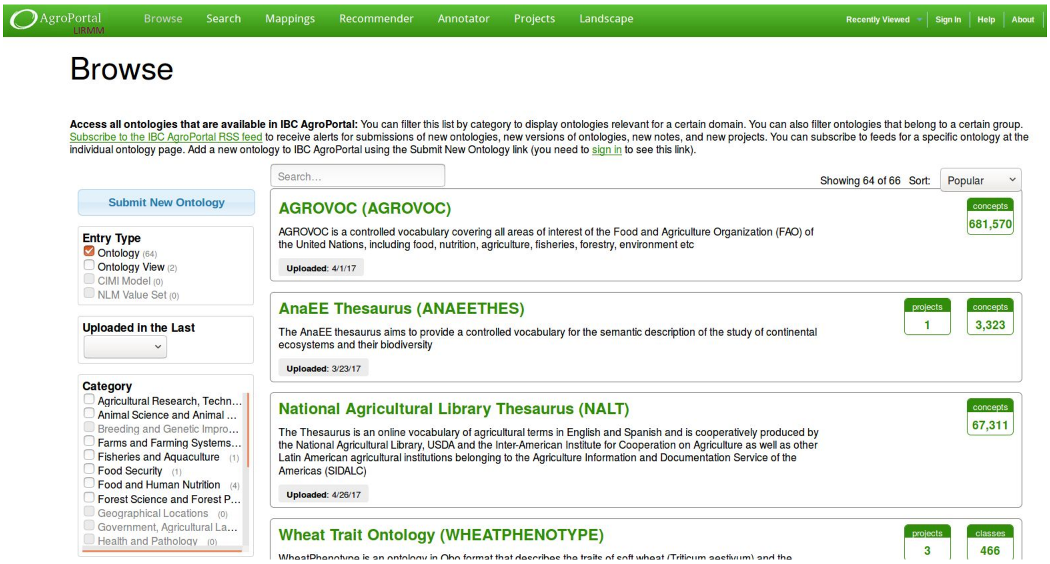
[52] http://www.ida.liu.se/~patla00/research/KitAMO/index.html

AgroPortal	A **repository** of ontologies and value vocabularies in the field of food and agriculture.	http://agroportal.lirmm.fr/
Neighbouring domains		
BioSharing	A **registry** of standards in life sciences, environment and biomedicine	https://biosharing.org/standards/
NCBO BioPortal	A **repository** of biomedical ontologies	https://bioportal.bioontology.org/
EBI OLS	A **repository** for biomedical ontologies	http://www.ebi.ac.uk/ols/index
OBO Foundry	A **catalogue** for biomedical ontologies	http://www.obofoundry.org/
General or multi-domain		
Protégé Ontology library	A **catalog** of ontologies developed using Protégé	https://protegewiki.stanford.edu/wiki/Protege_Ontology_Library
Ontology Design Patterns	A **catalog** of ontology design patterns (ODPs)	http://ontologydesignpatterns.org
LOV	A **registry** of linked open vocabularies with a SPARQL endpoint	http://lov.okfn.org/dataset/lov/
BARTOC	A **registry** of any kind of knowledge organization system from any subject area, in any language, any publication format, and any form of accessibility	https://www.bartoc.org/
TaxoBank	A **registry** of controlled vocabularies of all types	http://www.taxobank.org/
OntoHub	A **repository** of ontologies	https://ontohub.org/
FINTO	A **repository** of vocabularies, ontologies and classifications with browsing features and API access	https://finto.fi/en/
Colore	An open **repository** of first-order ontologies that can support the design, evaluation, and application of ontologies in first-order logic	https://github.com/gruninger/colore/tree/master/ontologies

In our view, initiatives like these are important, in that may help users find relevant, adequate, trustworthy semantic resources. However, the relative large number of them poses problems as well, to both data users and data producers - How to eventually select the most appropriate? And, What is the best place to have my resource adopted? How to make my community benefit from my work? Our contribution to solve these questions is to develop a ndfacilitate a community dialogue around all phases of the production and use of semantic resources, and so promote best practices and the tools and services that best fit our purposes.

AgroPortal

The AgroPortal project (http://agroportal.lirmm.fr) aims at offering a vocabulary & ontology repository for agronomy and related domains such as biodiversity, plant sciences and nutrition. AgroPortal specifically pays attention to respect the requirements of the agronomy community in terms of ontology formats (e.g., SKOS, trait dictionaries) or supported features (metadata, annotation).



Browse

Access all ontologies that are available in IBC AgroPortal: You can filter this list by category to display ontologies relevant for a certain domain. You can also filter ontologies that belong to a certain group. Subscribe to the IBC AgroPortal RSS feed to receive alerts for submissions of new ontologies, new versions of ontologies, new notes, and new projects. You can subscribe to feeds for a specific ontology at the individual ontology page. Add a new ontology to IBC AgroPortal using the Submit New Ontology link (you need to sign in to see this link).

Showing 64 of 66

Entry Type
☑ Ontology (64)
□ Ontology View (2)
□ CIMI Model (0)
□ NLM Value Set (0)

Uploaded in the Last

Category
□ Agricultural Research, Techn...
□ Animal Science and Animal ...
□ Breeding and Genetic Impro...
□ Farms and Farming Systems...
□ Fisheries and Aquaculture (1)
□ Food Security (1)
□ Food and Human Nutrition (4)
□ Forest Science and Forest P...
□ Geographical Locations (0)
□ Government, Agricultural La...
□ Health and Pathology (0)

AGROVOC (AGROVOC)
AGROVOC is a controlled vocabulary covering all areas of interest of the Food and Agriculture Organization (FAO) of the United Nations, including food, nutrition, agriculture, fisheries, forestry, environment etc
Uploaded: 4/1/17
concepts 681,570

AnaEE Thesaurus (ANAEETHES)
The AnaEE thesaurus aims to provide a controlled vocabulary for the semantic description of the study of continental ecosystems and their biodiversity
Uploaded: 3/23/17
projects 1
concepts 3,323

National Agricultural Library Thesaurus (NALT)
The Thesaurus is an online vocabulary of agricultural terms in English and Spanish and is cooperatively produced by the National Agricultural Library, USDA and the Inter-American Institute for Cooperation on Agriculture as well as other Latin American agricultural institutions belonging to the Agriculture Information and Documentation Service of the Americas (SIDALC)
Uploaded: 4/26/17
concepts 67,311

Wheat Trait Ontology (WHEATPHENOTYPE)
WheatPhenotype is an ontology in Obo format that describes the traits of soft wheat (Triticum aestivum) and the
projects 3
classes 466

AgroPortal features ontology hosting, search, versioning, visualization, comment, ontology recommendation, enables semantic annotation, as well as storing and exploiting ontology alignments. In addition, all the previous features are available through two endpoints allowing automatic querying of the content of the portal: (i) a REST web service API (http://data.agroportal.lirmm.fr/documentation); and (ii) a SPARQL endpoint (http://sparql.agroportal.lirmm.fr/test). Indeed, a new metadata model has been implemented to support better descriptions of ontologies and their relations with respect of the standards metadata vocabularies used in the semantic web community. This has resulted in the capability of automatically aggregating information about ontologies to facilitate the comprehension of the whole agronomical ontology landscape by displaying diagrams, charts and networks about all the ontologies on the portal (grouping, types of ontologies, average metrics, most frequent licenses, languages or formats, leading contributors & organization, etc.). A specific page dedicated to visualizing this landscape is now available in AgroPortal: http://agroportal.lirmm.fr/landscape. The latest version (v1.4) was released in July 2017 and currently hosts 64 public ontologies (and about ten of private ones); 95 other ontologies are candidate resources for AgroPortal.

5.6 Conclusions

In this section, we have proposed lists of tools and services for a few main activities that the maintainers of semantic resources regularly perform - editing, visualize and produce documentation and produce alignment. We also provide a list of initiatives, repositories, registries and catalogues, that aim at providing single access points to the many semantic resources available. To our knowledge only two focus specifically on agriculture, but we also list other with different scope.

Given the abundance of tools and services for each of those activities, especially editing tools, we have preferred to mention the tools that are best known and appreciated by the contributors to this document and provide pointers to surveys providing a broader lists wherever possible. Given the observations reported above, on one hand on a proliferation of semantic resources (Sec. 2), and on the other hand, on the little number of resources in machine-readable formats, the scarcity of APIs, not to mention the lack of documentation and of clearly stated licenses (Sec. 4), we suggest that more attention should be devoted to the promotion of coordinating efforts to address these gaps.

6. Conclusions and Next Steps

This document represents the first activity of the RDA Agrisemantics Working Group. The overarching goal of the group is to make agricultural data more interoperable by an effective use of semantics, that is why the group engaged in this landscaping exercise, to investigate what the state of semantics in our domain is like. This study was compiled on the basis of group discussions held both online and face-to-face, and integrates activities and results produced by our group members within other projects. The intended readers of this document are end users, above all managers, project coordinators, data scientists, and researchers interested in getting the big picture of semantics in agriculture (especially for Sec. 2 and Sec. 3), but it also addresses those more interested in the practical side of using semantic resources, by providing a discussion of the semantic resources currently (Sec. 4) and the tools and services (Sec. 5) available for editing, visualize, produce documentation, as well as mapping and listing/searching semantic resources.

In Sec. 2 we presented five major areas of use of semantics in agricultural data management. Those high-level areas are meant to provide an exhaustive grouping, and were elaborated and agreed during the RDA Agrisemantics Working Group discussions - elaborated during the regular group calls and several private discussions and approved and consolidated during RDA Plenary 9[53]. The underlying idea was to show that the use of semantics, or rather, the use of semantic resources, is ubiquitous in data management and use, although with different degrees of expressivity of the semantic resources used, and different support to the involved users.

We have provided a rather large definition of semantics, to reflect what nowadays seems a common practice - use “semantics” to refer to any information that allow a system to (semi-) automatically identify the “meaning” of data. This includes the use of “traditional” metadata

[53] https://www.rd-alliance.org/plenaries/rda-ninth-plenary-meeting-barcelona

describing entire datasets or information items such as publications, but also more fine-grained, shared and machine readable description of individual pieces of data - not only serving the purpose of making datasets findable on the web, but connect their contents in meaningful ways (this latter vision is graphically sketched in Figure 1). Most of our work here focuses on resources the original purpose of which was to to serve as value vocabularies (Sec. 1 and Sec. 2), but we also touched on metadata schema and ontologies. Unfortunately, an in-depth review of the use of ontologies for agricultural data was out of the scope of this landscaping exercise, but we believe it will be part of subsequent work of the working group. Thesauri and controlled vocabularies have been originally devised for the purposes of indexing and retrieval of information resources, while ontologies are needed every time an advanced knowledge application is needed (e.g. reasoning). In either case, we have collected evidence of a tendency to proliferation of resources, both for vocabularies and for ontologies (Sec. 2.1). What has been indicated as reasons for such proliferation is the fact that resources are often not easily findable, or rather that the knowledge about them tends to be limited to specific communities, despite their coverage could possibly span more topics and communities. Moreover, resources are often not available online and therefore scarcely reusable. Finally, it was highlighted that governance of the editorial control is often a problem. In other words, people report the fact that often it is easier to create a new resource from scratch, or starting from reusable fragments, instead of requesting that an existing resource is improved and expanded.

Some of the observations reported above have been confirmed by the analysis of the Map in Sec. 4, from which we learned that almost half of the semantic resources listed in the catalogue are not in machine-readable format, and that only a fraction has APIs (which makes their reuse more difficult). In our view, the way out to these problems is to promote more awareness in the community, support the selection and adoption of both specific semantic resources (Sec. 4) and services and tools for working with them (see Sec. 5). It is also important that more awareness is promoted in the area of licensing, with the idea of promoting open data, and address the issue of governance of such semantic, shared, machine-readable and open semantic resources. Under another angle, some of these problems may be addressed by adopting Linked Open Data (LOD) approaches, as the online publishing of resources makes them more easily findable, while the adoption of open formats and the creation of public APIs makes them accessible.

We also noted that thesauri and controlled vocabularies keep playing their main role of resources for indexing and retrieval, but there is a tendency to use them also to “tag” the identity of the subject measured in datasets of entities - observations, as per the scientific jargon. In other words, thesauri and controlled vocabularies are extending out of their traditional area of use, i.e., metadata schemas and information retrieval, and are more and more used as repositories of identities for the semantic web. This shades a different light on the point of the proliferation of resources. Consider for example all those applications where it is important to distinguish and unambiguously refer to different species of plants and animals, or chemical compounds, or even agricultural practices or land uses. The suggestion we gathered from this landscaping exercise is that it is time to start carefully considering the difference between *keywords* and *codes* on one side, as normally dealt with in metadata schemes, indexing and information retrieval applications, and entities and URIs on the other side, as needed by reasoning and information integration in a world of abundant, distributed and related data. It is dubious that the Linked Data approach alone may solve the two issues here highlighted, namely the proliferation of semantic resources and the practical confusion between tagging and reasoning. Linking data and vocabularies helps, but ontologies are still needed, especially when the need is to integrate data on different object, at different scale and produced by different users (Sec. 2.4, Sec. 2.5).

The community should take seriously the task of lifting the existing semantic resources to the web and spread knowledge on benefits and possible areas of applications (Sec. 2), appropriate tools (Sec. 5), existing resources (Sec. 4), and research trends (Sec. 3). As per discussions held internally in the group, one way to get out of the proliferation problem would be to rely less on extensive linking by identit exploit the advantages of the web, namely in the possibility of publishing and linking distributed data.

References

(Caracciolo et al. 2012) Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Keizer, J. (2012) The AGROVOC Linked Dataset. Semantic Web Journal 4(3). 2013 pp. 341-348. IOS Press, Amsterdam. ISSN: 1570-0844

(Celli et al. 2015) Celli F, Malapela T, Wegner K *et al.* AGRIS: providing access to agricultural research data exploiting open data on the web. *F1000Research* 2015, 4:110 (doi: 10.12688/f1000research.6354.1)

(Çağdaş & Stubkjær 2015) Çağdaş, V., Stubkjær, E., (2015) A SKOS vocabulary for Linked Land Administration: Cadastre and Land Administration Thesaurus, Land Use Policy, Volume 49, 2015, Pages 668-679, ISSN 0264-8377, http://dx.doi.org/10.1016/j.landusepol.2014.12.017.
(http://www.sciencedirect.com/science/article/pii/S0264837715000654)

(Ilic et al. 2007) Ilic, K., Kellogg, E. A., Jaiswal, P., Zapata, F., Stevens, P. F., Vincent, L. P., Avraham, S., Reiser, L., Pujar, A., Sachs, M. M., Whitman, N. T., McCouch, S. R., Schaeffer, M. L., Ware, D. H., Stein, L. D., Rhee, S. Y. (2007). The Plant Structure Ontology, a Unified Vocabulary of Anatomy and Morphology of a Flowering Plant. Plant Physiology Feb 2007, 143 (2) 587-599; **DOI:** 10.1104/pp.106.092825

(Laporte et al. 2016) Laporte M-A, Valette, L., Cooper, L., Mungall, C., Meier, A., Jaiswal, P., & Arnaud, E. (2016). Comparison of ontology mapping techniques to map plant trait ontologies. Unpublished. https://doi.org/10.13140/rg.2.2.29413.40166

(Lindblom et al. 2017). Lindblom, J., Ljung M., Jonsson, A. (2017) Promoting sustainable intensification in precision agriculture: review of decision support systems development and strategies. Precision Agriculture 18(3), pp. 309-331. DOI: 10.1007/s11119-016-9491-4

(Mir & Quadri 1970). Mir, S., M. K. Quadri, S. (1970). Decision Support Systems: Concepts, Progress and Issues - A review. In: Climate Change, Intercropping, Pest Control and Beneficial Microorganisms, pp. 379-399. DOI: 10.1007/978-90-481-2716-0_13

(Wegner 1996) Wegner, P. (1996). Interoperability. ACM Computing Surveys. 28(1), pg. 285-287. March.

(Rodríguez-Iglesias et al. 2013) , A., Egaña-Aranguren, M., Rodríguez-González, A., Wilkinson, M.D.
International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2013)

(Shvaiko & Euzenat 2013) Shvaiko, P., and Euzenat, J. Ontology Matching: State of the Art and Future Challenges. IEEE Transactions on Knowledge and Data Engineering (2013) 25:1. Doi: 10.1109/TKDE.2011.253

(Stellato et al. 2015) Stellato, A., Rajbhandari, S., Turbati, A., Fiorelli, M., Caracciolo, C., Lorenzetti, T., Pazienza, M. T. (2015). VocBench: A Web Application for Collaborative Development of Multilingual Thesauri. The Semantic Web. Latest Advances and New Domains. Springer International Publishing. https://doi.org/10.1007/978-3-319-18818-8_3

(Suominen & Mader 2014) Suominen, O., Mader, C.J. (2014) Assessing and Improving the Quality of SKOS Vocabularies. Data Semantics (2014) 3:47. Springer Berlin Heidelber. DOI: https://doi.org/10.1007/s13740-013-0026-0

(Villa et al. 2017) Villa F., Balbi S., Athanasiadis I.N. and Caracciolo C. Semantics for interoperability of distributed data and models: Foundations for better-connected information [version 1; referees: 1 approved with reservations]. *F1000Research* 2017, 6:686 (doi: 10.12688/f1000research.11638.1)

(West & Fowler 1999) Matthew West and Julian Fowler (1999). Developing High Quality Data Models. The European Process Industries STEP Technical Liaison Executive (EPISTLE).

(Wilkinson et al. 2016) Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*3:160018 doi: 10.1038/sdata.2016.18 (2016)

Annex 1: bibliometric study details

The following query was applied (12th of april 2017) on the Web of Science database, using SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH indexes, with a time restriction to the 2006-2016 period, and a document type restriction to Article, Book, Book chapter, Proceedings paper and Review.

The resulting corpus made of **2,800 publication records** is the intersection between the “semantic query” and the “agriculture query” detailed hereafter. Both lists have been discussed with the group, in particular during the IGAD meeting in Barcelona.

Step 1 : semantic query

TS=("linked data" OR "linked open data" OR "web of data" OR "web data" OR Semantic* OR "semantic information*" OR "semantic annotation*" OR "semantic relation*" OR “semantic representation*” OR "conceptual structure*" OR "description logic*" OR "semantic resource*" OR "knowledge organi?ation*" OR “knowledge engineering” OR “knowledge map*” OR “inference engine*” Or “reasoning engine" Or "knowledge retrieval" OR "information retrieval” OR "term alignment" OR "term extraction" OR “term recognition” OR "entity recognition" OR “concept alignment” OR “concept extraction” OR NLP OR “Natural Language Processing” OR “information extraction" OR "relation extraction" OR "context aware system" OR "key discovery" OR "semantic sensor based" OR DAML OR RDF OR RDFS OR "RDF(S)" OR "RDF/S" OR SKOS OR "SKOS-XL" OR "SKOS XL" OR SWRL OR OWL OR SPARQL OR "controlled vocabular*" OR KOS OR "common logic")

OR

TS="concept mapping" OR "concept alignment" OR "vocabulary alignment" OR "vocabulary mapping"

OR

TS=(AGROVOC OR (Biorefinery AND Semantic*) OR CAB Thesaurus OR "Cell Ontology" OR "Chemical Entities of Biological Interest" OR ChEBI OR "Crop ontology" OR "Crop Research Ontology" Or "plant trait ontology" OR "Crop Research Ontology" OR "CO_715" OR "Environment Ontology" OR ENVO OR "Experimental Factor Ontology" OR "Feature Annotation Location Description Ontology" OR FALDO OR "NAL Thesaurus" OR NALT OR "Phenotype And Trait Ontology" OR "Plant Experimental Conditions Ontology" OR "Plant Environment Ontology" OR "Plant Ontology" OR "Plant Trait Ontology" OR "Population and Community Ontology" OR "Protein Ontology" OR "Sequence Ontology" OR "Variation Ontology" OR "Agronomy Ontology" OR "OBO Foundry" OR OBOE)

Step 2 agricultural query

TS=(agricult* OR agronom* OR agrifood OR food OR “food transformation” OR wine* OR “oenolo*” OR “climate change” OR “agro*environmental” OR “cultural system” OR “crop system*” OR "agro*ecology" OR “crop management” OR “fruit*” OR cereal* OR “pest attack” OR “Plant science*” OR “plant development” Or rice OR farming OR farm* OR “agricultural intensification” OR bioeconomy OR “food pack*” OR biotechnolog* OR biorefiner*)

This set of 2,800 records was checked manually and cleaned from irrelevant content, e.g. texts about “owl” as a animal or “RDF” as “recommended dose of fertilization”. Publications about medicine with no mention of agriculture were also discarded.

The analyses presented in the report were built from a corpus made of 1441 publications.