Working with Semantic Resources in Agriculture. Requirements and Recommendations from the RDA Agrisemantics WG

Caterina Caracciolo

Food and Agriculture Organization of the UN

Abstract. Semantics, including both metadata, vocabularies and ontologies, is an important component to achieve data interoperability. In this paper we report on a set of requirements for tools and services, and high level recommendations to software developers ... and

1 Introduction

This document reports on the second step taken by the Agrisemantics Working Group towards the definition of recommendations for future e-infrastructures to support semantic resources (SR) in agriculture. "Semantic resources" in this context refers to "... structures of varying nature, complexity and formats used for the purpose of expressing the "meaning" of data" [REF], be those textual or numeric. Controlled vocabularies, value lists, classification systems, glossaries, thesauri, and ontologies are all example of semantic structures. They may be expressed in a variety of formats, open or proprietary, machine-readable or not. This broad definition then includes both the "vocabularies" as defined by W3C (i.e., including metadata elements and value vocabularies, aka knowledge organization systems), and ontologies, be those lightweight or with richer descriptions and logical axioms. We prefer to distinguish the content and use of semantic resources (e.g., thesauri for indexing or classification systems) from their formats (e.g., relational format, RDF or OWL), to avoid the sometimes misleading equivalence between the formats used to express and make resources available, and the semantic content (and purpose) of the resource itself.

Our first activity focussed on delineating the applications of SR in agriculture. Now, we report on our second activity, aimed at surveying the real-life problems and bottleneck that researchers and practitioners encounter when using semantic resources, together with their wishes and/or proposed solutions. We digested the input gathered from the community into requirements. The next step will be to distill our findings into a set of recommendations for e-infrastructures that aim at supporting researchers and practitioners in their work with agricultural data. Adopters of these recommendations include policy makers, funders, software developers, research scientists, data managers and the wider community that provided us with the input to define them.

We were particularly interested in identifying needs concerning:

1. Access to useful semantic resources
2. Reusability of semantic resources either by human or machines
3. Tools and services to create, manage, improve, interlink, publish semantic resources
4. Use of semantic resources or services in applications
5. Standards and best practices to represent and exchange semantic resource

To this end, we defined a template to facilitate contributing to the answers. The template was designed to be rather essential and suitable to accommodate to different cases. We received open problems, ideas for solutions at different stage of development, including ongoing or future projects to address those problems.

In the following we describe the process followed to collect and analyse the use cases (Sec. 2) and the requirements we drew from them, as resulting from the Workshop (Sec. 3). In Sec. 4 we discuss our findings.

??. vagaa.

2 Methodology

Input was collected using a template, defined by the group chairs with feedback from the Agrisemantics WG members. Other sources for the template include documents produced within RDA, use cases provided with no specified template during RDA P10, Barcelona.

Respondents were invited to answer 4 core questions (describe the limitations or difficulties they face) and 2 additional questions concerning the context of their work, plus two more questions about the respondent. All questions were open-ended and with some explanations added in the form of questions to help respondents to articulate their answer. The template is attached in Annex I.

The survey was made available as a Google Doc. Other, more formal, survey tools were too restrictive in the types of responses the user is meant to provide. The survey was distributed by means of the mailing lists of: Agrisemantics and IGAD (and all working groups in the area of IGAD), Agroportal, AGINFRA+, and personal communication.

Answers were collected from mid November 2017 through end of January 2018, with three general reminders. As a result, we received 20 use cases, most of which (13) were written directly by their providers while the remaining seven were written collaboratively as a result of an interview conducted by one of the chairs and the provider. All use cases are available from the RDA Agrisemantics Working Group web space. The list of use cases (title, author, and institution) are provided in Annex II.

Analysis was done by the group chairs, then submitted to the working group for review and comments. Each use case was analyzed by two chairs and the result summarized in a spreadsheet in order to provide an unified view on all

pieces of information collected. The requirements drawn from this step where then visually organized using an online mind map software, that was also the basis for the discussion within the working group. A static version of the map is in Annex III. The original use cases were linked from the map and their entire original text and intermediate summary always available to the group. The set of requirements resulting from this process were further discussed and finalized in the course of a workshop during the RDA P11 in Berlin (March 2018), with the participation of about 30 people. In the following, the requirements gathered are synthesized and presented.

3

Questo e'il riferimento.
2

4 Results of Use Case Collection

We collected 20 use cases, from institutions based in 10 distinct countries from 4 continents (15 from Europe, 2 from North and 2 from South America, 1 from Asia (China)), mostly from research organizations (15), 3 international organizations, 1 professional and 1 governmental organization.

From the use cases, it emerges that a number of different roles and backgrounds are involved in different tasks dealing with SR. They include:

- computer scientists, application developers, and data managers are largely represented both as producers and users;
- information technology professionals and librarians also;
- knowledge engineers and linguists are present but to a lesser extent;
- domain experts and researchers participate in the production but are also important users of semantic resources.

It is clear then that the process of producing semantic resources is highly collaborative and requires various competencies.

Also, virtually all tasks are mentioned in the use cases, from when SR are first created to their retrieval and use in applications.

The evidence we collected shows that there are as many toolkits as projects, covering all steps in the data life cycle and project workflow, from editing a semantic resource to its use in a given application. The great majority of use cases combine open source and ad-hoc tools, often developed in-house, while the commercial solutions adopted tend to be integrated platforms covering various phases of the semantic resources life cycle, for which no equivalent product is available for free and/or open source. Almost half of the use cases mention of RDF technologies, in particular triple stores.

5 Requirements

The high level message that we gathered from the use cases and the discussion that followed (RDA P11) is semantic technologies/methodologies need to be more accessible in terms of both skills and resources required for their development and use. In particular:

Rq1. Tools designed for use with semantic resources should also be accessible to non-ontologists. In particular, more attention should be payed to graphical interfaces, support for validation, and for application of methodologies appropriate for each task.

Rq2. Online platforms are needed to lift the burden of local (or ad-hoc) installations and maintenance from users or individuals.

Rq3. Common tasks involving semantic resources (e.g. editing, format conversion, etc.) should be integrated (or integratable), interoperable and flexible workflows, to minimize the breadth of skills required to work with semantic resources.

We further analyzed the last requirement above identifying four tasks:

1. Creation and maintenance
2. Mapping
3. Use in applications
4. Discoverability & Availability

Figure 1 below maps the four groups of requirements against a generic data lifecycle.

Figure 1. Semantic resource life cycle: green boxes represent production tasks while orange ones are for consumption. Smaller boxes are subtypes (plain arrows) of tasks in the larger boxes. Double arrows represent the life cycle and clouds are issues of concern for each task.

After presenting the requirements corresponding to those four groups, we discuss some issues related to availability and formats of actual SR, as gathered from the use cases and the face-to-face discussion.

5.1

5.2 Tasks involving semantic resources

5.3 Creation and Maintenance

This phase includes all tasks involved in the creation and evolution of a SR.

1. Editing tools should be designed having in mind that different users, and therefore competencies, are involved in various (sub)phases of the editing tasks. For example, often editing involves domain experts providing domain knowledge to the modeller, and then validating the resulting semantic model.

Therefore it is important that domain experts be able to understand and provide feedback on the semantic resources implemented by the knowledge manager.
2. Tools used in different phases of the editing process should be integrated. Editing a semantic resource is often articulated in subtasks, including eliciting knowledge from domain experts, formalizing that knowledge into a specific semantic structure, validating the resulting structure with domain experts, searching and reusing fragments from other resources or creating alignments with other sources. It should be possible to move from one activity to the other in an unfragmented way..
3. Tools should integrate methodologies for modelling, quality checking, and validation. Tools should support users in applying existing methodologies and practice while performing editing tasks, including modelling, quality check of the formalized and populated resource, validation. For example, in order to avoid overloading ontologies with an excess of classes that should better be organized in different ontologies, e.g., foundational or domain specific, or different domain-specific ones, tools may implement heuristics to warn risk of overloading and possibly suggest alternative modelling decisions.
4. Ontology editing tools should encourage to separate the definition of low-level resources from that of ontologies. Given the distinction between high- and low-level resources made elsewhere (ADD REF), editing tools should support users to implement it, by incorporating methodological and design principles and also by recommending specific resources to reuse.
5. Online platform(s) should be available to those who cannot afford hosting and maintaining platform in-house. Creating and maintaining SRs (either created from scratch or converted from existing resources) may involve more resources than actually available, for example in terms of skills, dedicated personne or IT infrastructure. Online platforms are also important to enable collaborative work.

Mapping This phase focuses on the alignment of SR, consisting in the creation of mappings between them. Here we refer to the mapping activity in general, independently of the type of mapping to establish, or of the reason for engaging in the task. This task could be discussed as part of the editing phase, but we present it in isolation because it does not require having editing rights on the resources to map.

1. Tools should make available state-of-art algorithms for the automatic extraction of candidate mappings. Competitive algorithms too often remain as research products that require advanced computing skills to reuse in another context and, as such, are difficult to install and configure, have poor or no interface at all, and offer no support to users.
2. Tools should integrate methodologies for mapping. A number of issues are critical to the production of good mappings. Methodologies and best practice

should support users during the various steps involved in the process of mapping creation, including searching for existing mappings to reuse, supporting the actual mapping creation (in case of manual creation) or validating those automatically generated.
3. Develop and make easy to use a specific semantic resource that would function as a hub to interconnect resources instead of creating many-to-many mappings between semantic resources.
4. Promote a standard to represent mapping between pre-semantic resources like spreadsheets and SR. In many cases, widely used SRs are not available in open, machine-actionable formats (See Chapter 4 in Landscaping document). This implies that ad-hoc solutions are devised in order to create mappings to them, with consequences on interoperability and possibility of reuse.
5. Promote a standard way to annotate spreadsheets with semantic resources. Spreadsheets are the principle way to manipulate or exchange data in many environments and for many purposes. In that context, column headers typically (could) belong to some types of semantic resources but the reference is easily lost and commonly established in ad-hoc manners. Guidelines and tools should be available to users to exploit those references within applications.
6. Appropriate graphical interface should be available to allow users validate mappings. Similarly to the requirements described in the section on editing, above, different users, hence different competencies may be involved in this task. Appropriate graphical interface and interaction mechanisms should be available to support the involved competencies and roles. This requirement is especially important considering the critical role that human validation plays in making mappings useful.

Use of SR in applications Under this heading we group together tasks related to the actual use of SR in applications. Some overlaps exist with the previous two groups (for example, SR need to evolve in order to be used, and mappings may be needed for the same reason), others span into the actual availability of resources or even their modelling. We discuss this group in isolation to emphasize the variety of factors essential to make SR used and usable.

1. Services should be available that notify updates of a SR to the applications using it. This is to avoid that changes in a SR are not reflected in the applications, causing delays in updates and possible breaks in the services provided by the application.
2. The use of SR should be facilitated to extend their adoption by a wide range of user profiles. The use of semantic resources is often perceived as something that requires very specialized knowledge, and a steep learning curve to achieve it. This may be related to the formal languages used (e.g., RDF, OWL) or to the logical modelling of some resources (e.g., symmetric properties, use of reasonings), or both. User interfaces oriented to the needs of application developers should be available, together with dedicated tutorials should be made available.

3. "Low-level resources" should be made created and made available, or well maintained when already existing, for use in applications. Ontologies, the resources with the highest level of formalized semantics (e.g., presence of axioms and possibility of applying rich inference) typically define classes, or generic categories (e.g., "crop" or "species") instantiated by individuals (e.g., the specific crop or species, defined in "low-level" resources). It is good practice to keep instances separated from ontologies, as they usually are in larger number than classes, are often maintained by different curators and tend to evolve much faster than ontologies. Such "low-level" resources are of fundamental importance in real-life applications.
4. Services and metrics to assess resources usage should be developed. Being in a world of limited resources, it is advisable to have a measure for the use of resources. This could help maintainers prioritize their resources and effort, and funders to get a grasp of the use of their fundings.

Accessibility and Discoverability This section focuses on all elements considered relevant to find and access SR online.

1. The use of global identifiers should be encouraged and supported. Global identifiers are the basis of accessibility over the web. Services should be made available that provide global identifiers, e.g., URIs or DOIs, to semantic structures so as to enable referencing, citation, mapping, and in general, reuse in information systems.
2. Metadata creation should be supported by tools to the greatest extent possible. Currently, much of the metadata generation task is on the data curator, with relatively little support by tools. This leads to little availability of metadata, often of poor quality (e.g., not up-to-date, sketchy or in inconsistent formats), with a consequent untapped potential for the programmatic access of data.
3. The description of the SR used in datasets should always be part of a dataset metadata. The vocabularies, classifications or ontologies used to collect and distribute data are a fundamental component of a dataset, but often "hidden" in the data. Despite major metadata schemes, e.g., DCAT, do include properties for that purpose, these properties are often not supported by data and content management systems (i.e., services like CKAN, Dataverse, DataCite, and CrossRef) or not enforced. This limits the possibilities of automatic search and integration of datasets.

5.4 Semantic resources in agriculture and nutrition

While most of the input we gathered from our correspondents focused on tools and services, it also touched on the availability of SR on specific topics. This section reports on needs related to the accessibility or usability of semantic resources on specific topics that are particularly strategic or intensively used in the domain. The main claims for such reference resources are 1) to avoid duplicated

efforts, and 2) to augment interoperability among datasets, information systems, and semantic resources themselves. Efforts should be made to:

1. Have machine-actionable lists of "entities" important to agriculture provided with global identifiers for use in applications, such as pests, diseases, livestock, agricultural activities (i.e., the "low-level resources" mentioned above). The point here is that such reference lists exist and are commonly used, but they are scarcely available in machine-oriented formats, such as spreadsheets, when not available as PDF. These converted resources should also be maintained over time. More generally, a significant number of global identifiers (e.g.,URIs) need to be dereferenceable with long-term sustainability, to instill trust and support use and alignment/mapping.
2. Support the use of SRs in conjunction with quantitative data, eg. involving measurements units or processes. Many semantic resources are traditionally and successfully used to tag/index textual information data. However, their use in the context of scientific datasets (e.g., observations) presents additional changes, such as expressing the unit of measurement (eg cubic tonnes or cubic meters) or the measuring method (e.g., ph in water or non-aqueous solutions). These fundamental pieces of information are usually treated in an ad-hoc manner, often by reusing SR originally designed for different purposes. As a consequence the interoperability of datasets is limited. Tools should support users in correctly handling quantitative data and the full set of attributes that define them.
3.
4.

Develop semantically enabled data types for commonly used objects in agriculture and nutrition. The values associated to a given type, e.g. "soil quality", would be declared and maintained by the community in an appropriate semantic resource which would provide global unique identifiers, and, ideally, labels in many languages.

```
<Insert Code Here>
```

1.

6 Discussion

The ultimate goal of the Agrisemantics Working Group is to serve as a community-based space to discuss and share experience on the use of semantics to enhance the interoperability of data in agriculture. The work presented in this document aimed at gaining

evidence on the most urgent needs felt by researchers and practitioners when dealing with semantic resources, and focus on requirements of broad scope, useful to the entire community, including funding agencies and research coordinators.

We tried not abstract away from the fine-grained details of current research or implementation problems, such as specific algorithms, optimization problems, or referring to specific methodologies when alternative ones are available. However, our work necessarily reflects the status of current research and practice in the area, and does hint some methodologies over others. For example, the issue of strategies for mapping creation and reuse (e.g., the pros and cons of 1-1 mappings compared to the mapping to a central hub) is currently receiving much attention, with different views regarding its goals and how to address it. Although not new, the distinction between low- and high-level resources is increasingly accepted, together with an emphasis on their separate but coordinated management and access. However, the actual implementation of this distinction varies and is still subject of research.

The requirements presented in this report are based on input provided by members of the RDA Agrisemantics working group and individuals, groups of institutions reached by them. The use cases collected mostly came from Europe and Research organizations. We have no use case from Africa, one from Asia (China), 2 replies each from South and North America. Most of the respondents and on site participants were both producers and users of semantic resources, while relatively few are "pure users". We received no use cases from the private sector, although the private sector is represented in the Agrisemantics group and in the face-to-face workshop.

Many of the requirements hint at need to publish existing semantic resources according to Semantic Web standards, to make them openly accessible, machine-readable, and exposed in triple stores with the twofold goal of increasing data interoperability and avoiding duplication. We appreciate that some initiatives are already being carried on in this sense (e.g. within GODAN and by individuals and organizations gathering around the RDA and GODAN communities) but, as also reported as a finding of our landscaping activity, this effort certainly needs to be further promoted.

We notice that many of the requirements presented are not specific to agriculture. This matches our understanding of semantics as something general, cross-domain. Instead, what we found very domain specific is the community environment, characterized by the resources used, and the social side of the work, i,e., the terminology adopted, the place where people gather or publish, the type of training they have access to, and the expectations about interfaces and functionalities. Similar evidence resulted from the bibliographic study included in the landscape report where publications were almost equally distributed in journals and conferences of the Agriculture and Information Management sectors.

As a next step, the group will distill the requirements presented in this document in the form of recommendations to project funders, research and data managers, as well as fellow researchers, in order to broaden up the use of semantic resources to improve the interoperability of data in the ag sector. We plan on phrasing these recommendations in different ways and formats, and possibly with different levels of details, in order to address the great variety of skills and profiles involved in the production and use of agricultural data.

[**?**]