Semantics for Data in Agriculture: A Community-Based Wish List

Caterina Caracciolo[1], Sophie Aubin[2], Brandon Whitehead[3] and Panagiotis Zervas[4]

[1] Food and Agriculture Organization of the United Nations, Rome, Italy
caterina.caracciolo@fao.org
[2] INRA, UAR 1266 DIST Délégation Information Scientifique et Technique, Versailles, France
sophie.aubin@inra.fr
[3] CABI, Wallingford, UK
b.whitehead@cabi.org
[4] Agroknow, Athens, Greece
pzervas@agroknow.com

Abstract. The paper reports on activities carried within the Agrisemantics Working Group of the Research Data Alliance (RDA). The group investigated on what are the current problems research and practitioners experience in their work with semantic resources for agricultural data and elaborated the list of requirements that are the object of this paper. The main findings include the need to broaden the usability of tools so as to make them useful and available to the variety of profiles usually involved in working with semantics resources; the need to online platform to lift users from the burden of local installation; and the need for services that can be integrated in workflows. We further analyze requirements concerning the tools and services and provide details about the process followed to gather evidence from the community.

Keywords: Semantics, agricultural data, vocabularies, ontologies.

1 Introduction

Increasing attention is being devoted to the use of semantics to achieve data interoperability [1], [2]. However, challenges still remain in making the technology of broader use. The goal of the Agrisemantics Working Group (WG) within the Research Data Alliance (RDA) is to gather researchers and practitioners interested in the use of semantics in conjunction with agricultural data. In this paper we report on one activity of the group, aim at finding out what the main issues and bottlenecks the community experience when working with semantic resources, and what are the requirements to overcome them. “Semantic resources” in this context refers to “...structures of varying nature, complexity and formats used for the purpose of expressing the “meaning” of data" [3], 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[1] (i.e., including metadata elements and value vocabularies, aka knowledge

1 https://www.w3.org/standards/semanticweb/ontology

organization systems), and ontologies, be those lightweight or with richer descriptions and logical axioms.
Our first activity focused on delineating the applications of semantic resources in agriculture [3]. Now, we report on our second activity, aimed at surveying the real-life problems and bottlenecks 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. We were particularly interested in identifying needs concerning: (a) access to useful semantic resources, (b) reusability of semantic resources either by human or machines, (c) tools and services to create, manage, improve, interlink, publish semantic resources, (d) use of semantic resources or services in applications and (e) standards and best practices to represent and exchange semantic resources.

2 Use Case Collection: Methodology and Results

Input was collected using a template, defined by the group chairs with feedback from the Agrisemantics WG members. Respondents were invited to answer 4 core questions (to describe the limitations or difficulties they face) and 4 additional questions concerning their role and the context of their work. All questions were open-ended, provided with some explanations expressed in the form of questions to guide respondents in articulating their answer.
As a result, we received 20 use cases. All use cases were summarized in a spreadsheet, then the requirements drawn from each use case were organized using an online mind map software. The graphical mind map was also used as a basis for discussion within the working group. The map together with all use cases are available from the RDA Agrisemantics Working Group web space[2]. 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.
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), 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 semantic resources, showing 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 semantic resources 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

2 https://www.rd-alliance.org/deliverable-2-use-cases-and-requirements

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 as open source. Almost half of the use cases mention of RDF technologies, in particular triple stores.

3 Requirements

The high level message collected is that semantic technologies/methodologies need to be made more accessible both in terms of 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. More specifically, more attention should be paid to graphical interfaces, terms used, support for validation, and for methodological support in 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 to form flexible and interoperable 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 and 4) Discoverability and availability.

3.1 Creation and Maintenance

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

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, it is important that domain experts are enabled 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 like eliciting and formalizing the knowledge, 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 modeling, quality checking, and validation. They may implement heuristics to warn risks and possibly suggest alternative modeling decisions or specific resources to reuse.
4. Online platform(s) should be available to those who cannot afford hosting and maintaining platform in-house. They are also important to enable collaborative work.

In the following, we provide specific requirements for each of the main task above. Then we discuss some issues related to availability and formats of semantic resources, as emerged from the use cases and the face-to-face discussion.

3.2 Mapping

This phase focuses on the alignment of semantic resources, consisting in the creation of mappings between them [4]. 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.

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.
5. Tools should integrate methodologies and best practice to support users during the various steps involved in the process, including searching for existing mappings to reuse, supporting the actual mapping creation (in case of manual creation) or validating those automatically generated.
6. Promote a standard to represent mapping involving semantic resources in not or little machine-actionable formats, e.g., spreadsheets.
7. Promote a standard way to annotate spreadsheets with semantic resources, in particular column heading referring to common concepts of the domain.
8. Appropriate graphical interface should be available to allow users validate mappings independently from their skill level regarding semantics. This requirement is especially important considering the critical role that human validation plays in making mappings useful.

3.3 Use

Under this heading we group together tasks related to the actual use of semantic resources in applications. We discuss this group in isolation to emphasize the variety of factors essential to make semantic resources used and usable.

1. Services should be available that notify updates of a semantic resource to the application using it. This is to avoid that changes in a semantic resource are not reflected in the applications, causing delays in updates and possible breaks in the services provided by the application.
9. Appropriate interfaces, formats, training, and documentation should be made available to tool developers to encourage the introduction of semantics in end user applications. The use of semantic resources is too often perceived as something that requires very specialized knowledge, and a steep learning curve to achieve it.
10. “Low-level resources” should be created and made available by and to the community, and well maintained when already existing. Such “low-level” resources are of

fundamental importance in real-life applications as they represent the actual subjects of observation, measurement and research - e.g., crop varieties, livestock, pests.

11. Services and metrics to assess resources usage should be developed. Ways to quantify and evaluate their use could help maintainers prioritize their resources and effort, and funders get a grasp of the use of their funding.

3.4 Discoverability and Accessibility

This section focuses on all elements considered relevant to find and access semantic resources online. In this area, we support the recommendations made through the FAIR principles [5].

1. The use of global identifiers should be encouraged and supported. Global identifiers, e.g., URIs or DOIs, are the basis of accessibility over the web.
12. Automatic creation of metadata should be supported by tools to the greatest extent possible, leading to increased availability of metadata and better quality (e.g., up-to-date, rich or in consistent formats).
13. Datasets' metadata should always specify the semantic resources in them. Despite major metadata schemes, e.g., DCAT[3], do include properties for that purpose, these properties are often not supported by data and content management systems (i.e., services like CKAN[4], Dataverse[5], DataCite[6], and CrossRef[7]) or not enforced. This limits the possibilities of automatic search and integration of datasets.

3.5 Semantic Resources in Agriculture and Nutrition

While most of the input collected focused on tools and services, it also touched on the availability of semantic resources on specific topics. 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 reference 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).
14. Support the use of semantic resources in conjunction with quantitative data as the usefulness of many semantic resources developed to tag or index textual information data is limited when applied to numeric data qualified by measurements (e.g., different units, such as cubic tons or cubic meters, or different measuring methods, such as pH in water or in non-aqueous solutions).

---

3 https://www.w3.org/TR/vocab-dcat/
4 https://ckan.org/
5 https://dataverse.org/
6 https://www.datacite.org/
7 https://www.crossref.org/

4 Conclusions

Many of the requirements hint a 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 noticed 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 type of training they have access to, and the expectations about interfaces and functionalities.
Considering that semantics is key to both efficient data discoverability and integrability to serve better research in agriculture, we call on the community of engineers and researchers who develop methods and tools to manipulate and use semantic resources to consider the requirements expressed in the use cases we collected and synthetized in this paper.

Acknowledgments

The work of Caterina Caracciolo contributing to this paper was supported by the Cross-Cutting project, funded by the Bill and Melinda Gates Foundation. The views expressed in this information product are those of the author and do not necessarily reflect the views or policy of FAO.
The work of Panagiotis Zervas presented in this paper has been partly supported by the AGINFRA PLUS Project that is funded by the European Commission's Horizon 2020 research and innovation program under grant agreement No 731001.

References

1. Haav, H., Kungas, P.: Semantic data interoperability: the key problem of big data. In: Big Data Computing. CRC PRESS (2014).
2. Villa, F., Balbi, S., Athanasiadis, I., Caracciolo, C.: Semantics for interoperability of distributed data and models: Foundations for better-connected information [version 1; referees: 1 approved with reservations]. F1000Research. 6, (2017).
3. Aubin, S., Caracciolo, C., Zervas, P.: Landscaping the Use of Semantics to Enhance the Interoperability of Agricultural Data. Agrisemantics Working Group. https://www.rd-alliance.org/deliverable-1-landscaping, last accessed 2018/08/22.
4. Shvaiko, P., and Euzenat, J. Ontology Matching: State of the Art and Future Challenges. In IEEE Transactions on Knowledge and Data Engineering 25:1 (2013). Doi: 10.1109/TKDE.2011.253.

5. Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg et al. The FAIR Guiding Principles for scientific data management and stewardship. In Scientific Data 3 (1): 160018 (2016). doi:10.1038/sdata.2016.18.