Uses and Needs of Semantic Resources in the Area of Agriculture

A Report from the RDA Agrisemantics WG

12th December, 2018

Table of contents

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",[1] be those textual or numeric. Controlled vocabularies, value lists, classification systems, glossaries, thesauri, and ontologies are all example of semantic resources. 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[2] (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 SRs (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.

The first activity of the Agrisemantics Working Group focused on delineating the applications of SR in agriculture. Now, we report on our second activity, aimed at surveying the real-life problems and bottlenecks that researchers and practitioners encounter when using SRs, 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 targeting a number of profiles, including 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:
- **Access** to semantic resources
- **Reusability** of semantic resources
- **Tools and services** to create, manage, improve, link, and publish semantic resources
- **Usage** of semantic resources or services in applications
- **Standards and best practices** to represent and exchange semantic resources

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

---

[1] RDA Agrisemantics Working Group (2017) Landscaping the Use of Semantics to Enhance the Interoperability of Agricultural Data
[2] https://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/

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.

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[3], and group discussions held during RDA P10 in Barcelona, including test use cases provided with no specified template[4].

Respondents were invited to answer 4 core questions, to describe the limitations or difficulties they face; 2 additional questions concerning the context of their work; ands two more questions about the respondent. All questions were open-ended, provided with some explanations expressed in the form of questions to guide respondents in articulating their answer. The template is attached in Annex I. The survey was made available as a Google Doc. Dedicated online survey tools were also considered, but discarded to allow respondents maximum flexibility in providing their contribution.

The survey was distributed by means of the mailing lists of the RDA Agrisemantics WG and all the groups affiliated with the RDA Interest Group IGAD. The the mailing list of Agroportal[5] and the EC-funded AGINFRA+ project[6], plus personal communications were also used. Answers were collected from mid November 2017 through the 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[7]. The list of use cases collected (title, author, and institution) is provided in Annex II.

Use cases were analyzed by the WG chairs, the resulting analysis regularly submitted to the working group for discussion and comments. All use cases were summarized in a spreadsheet in order to provide an unified view on all pieces of information collected, then the requirements drawn from each use case were organized using a collaborative, online mind mapping tool[8]. In each step, at least two chairs per use case were involved, in order to facilitate harmonization of the process and results. The graphical mind map was also used as a basis for discussion within the working group. A static version of the map is in Annex III

---

[3] Wu, Mingfang; Psomopoulos, Fotis; Khalsa, Siri Jodha; Larkin, Jennie; de Waard, Anita (2018) Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories
[4] In particular, thanks to Ferdinando Villa, from BC3, Spain,for the use-case: 'Agrisemantics use case: Semantically-driven assessment of economic returns from biodiversity protection'
[5] http://agroportal.lirmm.fr/
[6] http://plus.aginfra.eu/
[7] https://www.rd-alliance#.org/system/files/documents/AgrisemanticsCollectionOfUseCAses.pdf
[8] https://www.mindmup.com/

and can be downloaded from the WG website[9]. The map includes links to the original use cases for reference, and an 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[10] in Berlin (March 2018), with the participation of about 30 people. In the following, the requirements gathered are synthesized and presented.

Results of Use Case Collection

We collected 20 use cases[11], from institutions based in 10 distinct countries on 4 continents (15 from Europe, 2 from North and 2 from South America, 1 from Asia (China)). They were primarily from research organizations (15), with a handful from international (3) professional (1) and governmental (1) organizations.

From the use cases, it emerges that a number of roles and backgrounds are involved in different tasks dealing with SR, showing that the process of producing SRs is highly collaborative and requires various competencies. 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 SRs.

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 tools and toolkits as projects,** covering all steps in the data life cycle and project workflow, from editing a SR 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 SRs 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.

Requirements

The high level messages that we gathered from the use cases and the discussion that followed (RDA P11) is that semantic technologies/methodologies need to be made more

---

[9]
https://www.rd-alliance.org/system/files/documents/requirements_from_use_cases_mindmap_v2_04-05-2018.pdf

[10] https://www.rd-alliance.org/plenaries/rda-eleventh-plenary-meeting-berlin-germany

[11] Use cases were provided by GFAR, U of Tor Vergata, INRA, AgroParistech, Embrapa, Ikerbasque Center for Climate Change, U of Montpellier, AgMIP, Irstea, CREA, CAAS, Solidaridad Network, Poznan Supercomputing and Networking Center, the German Federal Institute for Risk Assessment, U of British Columbia, ISKO, FSU Jena.

accessible both in terms of skills and resources required for their development and use, as indicated by the three requirements below, Rq1-Rq3:

General requirements to facilitate the creation and use of semantic resources	
Rq1	Tools designed for use with SRs should be accessible to non-ontologists. In particular, more attention should be payed to graphical interfaces, 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 SRs (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 SRs.

We further analyzed the last requirement above, Rq3, according to the following four macro-tasks:
   a. Creation and maintenance
   b. Mapping
   c. Use in applications
   d. Discoverability & Availability

Figure 1 below maps the four groups of requirements against a generic data lifecycle.
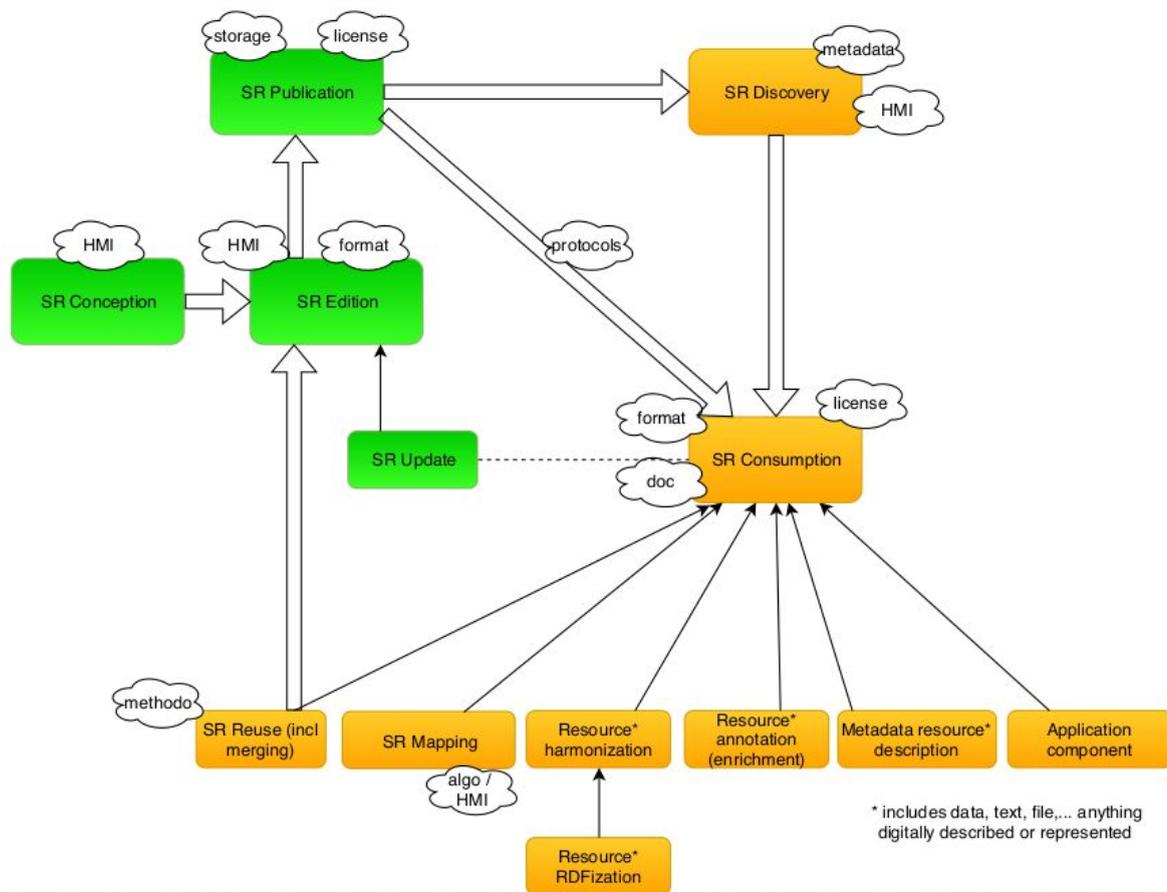
Figure 1. Semantic resource life cycle: green boxes represent production tasks while orange boxes are for consumption. Smaller boxes are subtypes (plain arrows) of tasks in the larger boxes. Double arrows represent the life cycle and clouds are issues of concern for each task.

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

Tasks involving semantic resources

Creation and Maintenance

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

Requirements for improved SR creation tools and methodologies	
ReqC1	Editing tools should be designed having in mind that different users, and therefore competencies, are involved in various 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 are enabled to understand and provide feedback on the SRs implemented by the modeller/knowledge engineer.

ReqC2	Tools used in different phases of the editing process should be integrated. Editing a SR 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.
ReqC3	Tools should integrate methodologies for modelling, quality checking, and validation. Tools should support users in applying existing methodologies and practices while performing editing tasks, when such methodologies and practices exist. For example, in order to avoid overloading ontologies with an excess of classes which are better organized in a separate ontology/resource (e.g., foundational or domain specific ontology, or a taxonomy), tools may implement heuristics to warn modellers and possibly suggest alternative modelling approaches.
ReqC4	Ontology editing tools should support the separation and coordination of low-level resources (or value vocabularies). Ontologies, the SRs 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, for ease of maintenance (they usually are in larger number than classes, are often maintained by different curators, are language-dependant, and tend to evolve much faster than ontologies) and use. Editing tools should support user implementation by incorporating methodological and design principles, and potentially recommending specific resources to reuse.
ReqC5	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 personnel or IT infrastructure. Online platforms are also important to enable collaborative work.

Mapping

This phase focuses on the alignment of SRs, 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.

Requirements for supporting mapping activities	
ReqM1	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.
ReqM2	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.
ReqM3	Develop and make easy to use a specific SR that would function as a hub to interconnect resources instead of creating many-to-many mappings between SRs.
ReqM4	Promote a standard to represent mapping involving SRs expressed in little or no machine-actionable formats, e.g., spreadsheets (this is a frequent case, as per findings reported in the Landscaping document (Chapter 4). This implies that ad-hoc solutions are regularly devised in order to create and store mappings to them, with consequences on interoperability and possibility of reuse.
ReqM5	Promote a standard way to annotate spreadsheets with SRs. Spreadsheets are the principle way to manipulate or exchange data in many environments and for many purposes. In that context, column headers typically belong to some types of SRs but the reference is commonly established in ad-hoc manners. In some cases, this is also true for values of spreadsheets. Guidelines and tools should be available to users to exploit those references within applications.
ReqM6	Appropriate graphical interface should be available to allow users validate mappings. Given that different users, hence different competencies may be involved in validating mappings, appropriate graphical interfaces and interaction mechanisms should be available to support the various competencies and roles involved. This requirement is especially important considering the critical role that human validation plays in making mappings useful.

Accessibility and Discoverability

This section focuses on all elements considered relevant to find and access SRs online.

Requirements for enhanced access and discoverability of semantic resources	
ReqA1	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. Services should be made available that provide global identifiers, to semantic structures so as to enable referencing, citation, mapping, and in general, reuse in information systems.

ReqA2	Automatic creation of metadata 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.
ReqA3	Datasets' metadata should always specify the SRs in them. 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,[12] do include properties for that purpose, these properties are often not supported by data and content management systems (i.e., services like CKAN[13], Dataverse[14], DataCite[15], and CrossRef[16]) or not enforced. This limits the possibilities of automatic search and integration of datasets.

Use of SR in applications

Under this heading we group together tasks related to the actual use of SRs in applications. Some of the requirements presented here overlap with the two groups discussed above, editing and mapping (SRs need to be maintained in order to be used, and mappings may be needed for the same reason). Other requirements deal with the actual availability of resources or their modelling. We discuss this group in isolation to emphasize the variety of factors essential to make SRs used and usable.

Requirements towards increased and easier (re-)use of semantic resources	
ReqU1	Services should be available that notify updates of a SR to the application 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.
ReqU2	Appropriate interfaces, formats and documentation should be made available to tool and service developers. The use of SRs 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. A wide range of user profiles should be taken into account when making SRs available, so as to facilitate their adoption and reuse.
ReqU3	"Low-level resources" should be created and made available, and well maintained when already existing, for use in applications. Such "low-level"

[12] https://www.w3.org/TR/vocab-dcat/
[13] https://ckan.org/
[14] https://dataverse.org/
[15] https://www.datacite.org/
[16] https://www.crossref.org/

	resources are of fundamental importance in real-life applications as they represent the actual subject of observation, measurement and research - e.g., crop varieties, livestocks, pests.
ReqU4	Services and metrics to assess resources usage should be developed. Most SRs are made public and reused, but actually lack ways to quantify and evaluate their use. If such ways were in place, they could help maintainers prioritize their resources and effort, and funders get a grasp of the use of their fundings.

Semantic resources in agriculture and nutrition

While most of the input we gathered from our correspondents focused on tools and services, some also touched on the availability of SRs on specific topics. The main requests for such reference resources are that 1) effort is not duplicated, and 2) interoperability among datasets, information systems, and SRs themselves is improved. In particular, efforts should be made to:

Requirements for domain-specific semantic resources	
ReqSR1	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). Many such reference lists exist and are commonly used, but they are scarcely available in machine-oriented formats and should then be lifted up to formats more suitable for use in applications, and reliably maintained. Special attention should be given to the long-term sustainability of global identifiers (e.g.,URIs) , to instill trust in users.
ReqSR2	Support the use of SRs in conjunction with quantitative data, i.e., data involving measurement units or processes. Many semantic resources, such as thesauri and controlled vocabularies, are traditionally and successfully used to tag/index textual information data. However, their usefulness is limited when dealing with 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). These fundamental pieces of information are often 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.

	Develop semantically enabled data types[17] 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.

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 SRs, and focus on requirements of broad scope, useful to the entire community, including funding agencies and research coordinators. We tried to abstract away from the fine-grained details of individual's research or implementation problems, and avoid 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 at 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 a need to publish existing SRs 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.

---

[17] The data type of a value is an attribute that tells what kind of data that value can have or determines how tools should handle the given values. Common simple data types are integer or string. They can be more specific like date, or boolean values.

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 conferences or journals they chose to 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[18] 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 SRs 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.

---

[18] https://www.rd-alliance.org/system/files/documents/Deliverable1%20-%20Landscaping.pdf

ANNEX I - What are your Needs when working with Semantic Resources?

If you are a software developer, a knowledge engineer, or a project manager, and if you are currently using or planning on using semantic resources on agriculture or food data, this question is for you!

Semantic resources include controlled vocabularies, classifications, ontologies, or any similar structure you use to describe/define data in your work. The goal of this survey is to identify what needs should be addressed in order to improve work with semantic resources. Your answers will contribute to shaping recommendations for e-infrastructures to work with semantic resources, final output of the RDA Agrisemantics Working Group.

We are then interested in hearing the problems you face, especially in terms of tools and functionalities to work with semantic resources, but not only. Areas of interest include but are not limited to:

- **Accessibility of** useful semantic resources
- **Reusability** of semantic resources either by human or machines
- **Tools and services** for managing the lifecycle of semantic resources (e.g., creation, maintenance, linking, publication)
- **Standards** and best practices to represent and exchange semantic resources
- **Connection to end-user applications** that use semantic resources or services.

Please, make sure you answer at least the questions marked with * !

Thank you!
The Agrisemantics WG

Use case title*
Can you synthesize your use case with a title? You may want to start with a tentative title and revise it after the you have answered all questions.

...

1. Problem statement*
What does the use case aim to address? How? Who are the beneficiaries? What are the impacts? etc.

...

2. Ontologies and vocabulary requirements

A. Can you say something about the semantic resource(s) you use? Do they include hierarchical/specific relationships, logical axioms, synonymy, textual definitions, several languages, references to external concepts, etc.

…

B. How are they used? What for?

….

C. Are they local or shared resources? Are they publicly accessible? If so, can you share an URL?

…..

D. Do they come with a clear license? Which license?

…..

3. Semantic toolkit

What are the tools and services used to manipulate/use those semantic resources? Is it standalone software or integrated in a complex workflow?

….

4. Limits and expectations*

Can you summarize the difficulties/shortcomings of your work with semantic resources? Where are the bottlenecks? What would you like to improve? What could be the solutions?

….

Tell us a bit more about…

5. Data requirements

What type of data, format, storage, size, workflow, etc apply to this use case?

….

6. Manpower

What is the profile of persons interacting with the semantic resource or their application?

….

7. Focal point*

Can you leave us your name and email address? This is to contact you in case of doubts

…

8. Organization and role*

What is your organization and your role in it?

….

THANK YOU!!

ANNEX II - Collected use cases

The list is formatted as follows: Identifier: Title, Provider, Organisation, type of requirements
The use cases are presented in the online document[19]:

UC01: *Search data set by the semantic resource it uses* **by Valeria Pesce (GFAR)**
    mapping, use, documentation, resources
UC02: *Metadata Exploitation* **by Armando Stellato (Tor Vergata University)**
    mapping, documentation,resources
UC03: *VocInra: turning an institutional keyword list into a linked open thesaurus* **by Sophie Aubin (Inra)**
    editing, mapping, resources
UC04: *Linking Wheat Data With Literature* **by Robert Bossy and Thomas Letellier (Inra)**
    editing, mapping, use, format
UC05: *Data integration for sensory and environmental quality in food* **by Liliana Ibanescu (AgroParistech and Inra)**
    editing, resources
UC06: *Improve management, application, validation of terminologies at Embrapa, and training on using them* **by Ivo Pierozzi Jr (Embrapa)**
    use, editing, training
UC07: *Link and search ontologies and vocabularies to achieve semantically-driven assessment of economic returns from biodiversity protection* **by Ferdinando Villa (Ikerbasque Center for Climate Change)**
    use, documentation
UC08: *Lack of support for managing/finding/validating/reconciling/accessing alignments between ontologies* **by Clement Jonquet (LIRMM, University of Montpellier)**
    mapping
UC09: *AgMIP Data Interoperability* **by Cheryl Porter (AgMIP)**
    use, resources
UC10: *Farm Data Storage and Access, and Field Data Observation* **by Catherine Roussey (Irstea)**
    use, documentation, resources
UC11: *Soil Data Interoperability* **by Giovanni L'Abate (CREA-AA)**
    editing
UC12: *Make your soil research data available, accessible, discoverable and usable* **by Giovanni L'Abate (CREA-AA)**
    use, format
UC13: *Agricultural Science and Technology thesaurus* **by Xuefu Zhang (CAAS)**
    editing
UC14: *Farmer and farming data for sustainability* **by Amanda Moura (Solidaridad Network)**
    documentation, training,

---

[19] https://www.rd-alliance.org/system/files/documents/AgrisemanticsCollectionOfUseCAses.pdf

UC15: *Publication of Inspire-based agricultural Linked Data* **by Raul Palma (PSNC)**
    mapping
UC16: *Food safety model repositories* **by Matthias Filter (BfR)**
    editing, use, training
UC17: *High-throughput phenotyping* **by Alice Boizet (Inra)**
    discovering, resources
UC18: *Food Traceability with respect to foodborne pathogen outbreak investigations* **by Damion Dooley (University of British Columbia)**
    mapping, training, resources
UC19: *Professional Society member needs* **by Stella Dextre Clarke (ISKO)**
    training
UC20: *AquaDiva* **by Alsayed Algergawy (FSU Jena)**
    use

Annex III - Map of use cases