

# 39 Hints to Facilitate the Use of Semantics for Data on Agriculture and Nutrition

## Recommendations from the RDA Agrisemantics Working Group

This document presents the recommendations of the RDA Agrisemantics Working Group (WG) to promote the use of semantics for agricultural data for the purpose of enhancing data interoperability in agriculture. These recommendations are high-level, to encourage researchers and practitioners to extend them according to their area of expertise.

The activities of the Agrisemantics WG started off with an analysis of the current landscape of use of semantic resources with agricultural data, based on first-hand experience of our group members, as well as on bibliometric investigation, and an analysis of existing repositories<sup>1</sup>. From that first activity, we found that despite the many possible applications of semantics and the interest shown by research and industry alike, their actual implementation and use is lagging behind. A large number of resources are not in machine-readable formats or do not have public APIs, while those available are often used beyond their intended area of use, possibly leading to problems. Our conclusion is that there is a strong need for “lifting” semantic resources to the web, and create new ones appropriate to their intended use.

Based on the insights gained from that analysis, we moved on to look in more detail at the actual reality of working with semantic resources, we opened a call for use cases, to collect examples of actual work and limitations, together with evidence of work already ongoing to address them. From that, we found that a large variety of profiles (and so skills and roles) are involved in working with semantic resources, and that there seem to be as many tools and toolkits as projects. Following the explicit suggestions provided by our respondents, we identified three types of implications of this fact, namely:

1. **Implications on software usability.** Given the variety of profiles involved, tools designed for use with SRs should be accessible to non-ontologists. In particular, more attention should be paid to graphical interfaces, support for validation, and for methodological support in each task.
2. **Implications on tools availability.** Reportedly, not all institutions or organizations can afford the whole set of activities associated to them, online platforms are needed to lift the burden of local (or ad-hoc) installations and maintenance from users or individuals.
3. **Implications on tools interoperability.** From the use cases presented, it emerges that the different tasks commonly required to work with SR are commonly achieved by pipelines of different tools. Given this fragmented landscape, we recommend that tools to perform common tasks involving SRs (e.g. editing, format conversion, etc.)

---

<sup>1</sup> The report is available from the Agrisemantics page of the RDA website. See <https://www.rd-alliance.org/system/files/documents/Deliverable1%20-%20Landscaping.pdf>

should be integrated, or integratable, to form flexible and interoperable workflows. This approach would contribute to minimizing the breadth of skills required to work with SRs.

Now, we distilled our findings into a set of high level recommendations (HLR) for future activities and implementations.

**HLR1. Foster the development of a generic framework to work with Semantic Resources (SR). This framework should be available online and be easily adaptable** according to:

- **the tasks to perform:** Tasks along the entire lifecycle of SRs and data should be supported. Examples for SR producers include: conception, editing, update, quality check, validation, metadata creation (preferably automatic), and publication. Examples for SR consumers include: discovery, mapping, merging and annotation;
- **the domains** to cover: Users working with different domains should be able to connect to relevant SR repositories and data type registries;
- **the user competencies** required. By providing a large catalog of tools with different levels of complexity, appropriate documentation, methodological tips, and training.

Such a framework should support **collaborative work**, and take into particular account the needs of non-specialists of semantics, like domain experts, data managers and application developers (consuming SRs). It should encourage the **reuse** of existing resources whenever possible to reduce workload and increase interoperability. In order for such tools to effectively extend the framework, the appropriate licencing policy needs to be in place, to ensure the legal interoperability of the framework and the candidate tools for inclusion. The legal interoperability aspect needs to be considered since the beginning of any project or initiative.

**HLR2. Support initiatives aimed at developing tools and resources for alignment** (semantic mapping). Activities of interest include, but are not limited to:

- The development of **innovative tools for creating, validating, and maintaining alignments** between two semantic resources or between a dataset and a semantic resource. Such tools should integrate state of art algorithms, have user-friendly GUIs, and be as integrated as possible in the working environment of the user.
- The promotion of a **standard for expressing alignment between SR expressed in little or no machine-actionable formats.**
- The development and long term maintenance of **low-level resources<sup>2</sup> for common objects of interest in agriculture** (e.g., species and varieties, diseases, food products) offering persistent identifiers (PIDs). Supported resources should conform to W3C recommendations and be published according to the FAIR principles.

**HLR3. Promote the adoption of common metadata models for the description of semantic resources and alignments, and the use of global identifiers to facilitate**

---

<sup>2</sup> Reference lists of entities to be used in conjunction with ontologies (i.e., corresponding to instances of ontologies). To this end, it is essential that those lists be provided with global identifiers, to enable programmatic reuse in applications. Examples of low-level resources in agriculture include lists of crops, livestock, pests, diseases, or agricultural activities

**reuse, ensure usage tracking, and citation.** This is to promote the access and reuse of semantic resources. To this end, improving information on provenance is a priority. A short set of metadata models fitted to describe semantic resources and alignments needs to be gathered and promoted. Whenever needed, existing models should be completed with adequate metadata items to allow, for instance, the production of usage metrics (if not available, they should be created). This should come with a more generalized adoption of persistent global identifiers (PIDs) at various level (e.g., the resource itself, the concepts, mappings, its creators). We do not recommend a specific PID system (as its choice also depends on the needs of the specific application and domains), but do recommend that the system chosen is widely adopted, and that the possibility of creating correspondences with other systems is taken into consideration since the beginning (especially to increase the possibility of data integration and interoperability). Shared and adapted metadata as well as persistent, global identifiers will contribute to develop better practices in reusing and citing resources in the agri-food domain.

**HLR4. Create courses on semantics and integrate them into curricula** in computer sciences, data & information management and analytics, agronomy, bioinformatics, etc. This is the condition for a broader adoption of semantics to handle agrifood data. As semantics is needed in all steps of data life cycle, relevant knowledge and skills should be shared by all roles involved in the workflow - domain experts and software managers alike. People developing software to store, share, integrate and analyze the data must be considered as prior targets if we want semantics to be integrated in mainstream platforms and tools.

## Recommendation according to profiles and roles:

The recommendations given above are organized according to their area of coverage. In the following, we arrange them according the skill sets and roles that in the best position to implement them. At the same time, we also provide more details for as many recommendations as possible.

### 1. Developers of tools for manipulating semantic structures

We wish to address software engineers, from both academia and industry, who produce tools and platforms to conceive, edit, merge, map, and share semantic resources. With these recommendations, we intend to guide them towards easier-to-use, more interoperable, smarter technologies for the benefit of non experts in particular. A table of correspondences between the more specific recommendations given below and the Requirements listed in Deliverable 2 can be found in Annex I.

#### General recommendations

1. Build **tools that may be integrated into suits**, to unburden users from tasks such as format transformation and transfer of metadata. Integratable tools should be available at all phases of the editing process, e.g., eliciting knowledge, formalization, reuse and alignment. This implies **choosing standard technologies and shared**

**input/output formats** for both content representation (e.g. SKOS<sup>3</sup> and TEI<sup>4</sup>) and metadata encoding.

2. Use **open source** licenses that will facilitate , at a minimum, allow for the integration of your software into pipelines and public e-infrastructures.
3. Build **tools that support known best practices for the task at hand**, including the implementation of FAIR principles, and for reusing, citing and providing SR with metadata and documentation, distinguishing conceptual from linguistic content. Implement FAIR metrics to assign level of FAIRness to SRs
4. **Provide** (if possible, multilingual) **documentation on your tools and toolkits**. Make it possible for third parties (e.g., e-infrastructure developers) to provide user support when needed.

## Usability

5. **Consider non-semantic experts** (e.g. biologist, agronomist) when **developing tools**. Tools should be user-friendly both at the level of software interaction and through graphical interfaces that use understandable terminology. Tools should smoothly support those users in a number of activities, such as bulk edits, or editing by drag & drop. A cloud-based platform (with no local installation required) also contributes to the usability of tools, especially if also collaborative features are implemented.
6. Implement **(semi)automatic generation of metadata** and transfer protocols to pass metadata on from one tool to another. Duplicated effort in recording metadata at different stages of SR creation and reuse must be avoided.

## Features

7. **Guide users in choosing appropriate format and modelling approach** to represent a SR. For example, OWL may be too expressive for a vocabulary that could be effectively modelled as a SKOS resource, and conversely, an OWL *file*, does not necessarily imply an ontology.
8. Allow for the **customization** of editing and alignment tools, such as the possibility of (de)activate features according to the nature of the SR to be edited (e.g., simple lists, thesauri, formal ontologies).
9. Include automatic **quality check** on both structure and content of SR, by allowing for reuse of existing tools and algorithms.<sup>5</sup>
10. Connect tools or services to SR repositories in the user's domain(s) of interest in order to facilitate the reuse of resources (e.g., terms, concepts, low-resources) whenever possible, instead of creating new content.
11. Implement notification mechanisms to inform users, both human and applications, when a new version of a SR is available.

---

<sup>3</sup> <https://www.w3.org/2004/02/skos/>

<sup>4</sup> <http://www.tei-c.org/>

<sup>5</sup> For example, for SKOS, qSKOS and Skosify.

12. Include concept/class **alignment** advanced features in SR editing tools, integrate state-of-art alignment algorithms. In particular, consider the alignment of resources of different levels of expressiveness, e.g. thesauri with ontologies. Allow for **collaborative evaluation** of alignments. Implement the **use of metadata and recording of provenance** of concepts in alignment processes.
13. Add **alignment discovery services** that deal with SPARQL endpoints and develop algorithms and tools to **mine large sets of alignments**.

## 2. Semantic professionals

14. Make your SRs easily findable, accessible, interoperable and reusable, by applying W3C recommendations and the FAIR principles<sup>6</sup>. In particular:
  - **Share your resources** in various repositories and catalogues, be they generic or domain specific;
  - **Reuse common metadata schemes for SRs, and provide adequate documentation;**
  - **Provide persistent identifiers (PIDs)** to your resources, e.g. DOIs, and/or **dereferenceable URIs**
15. Publish new versions in repositories that handle versions and devise mechanisms to inform users about changes.
16. **Allow for the possibility of exposing/publishing SR at different levels of maturity** (e.g., alpha, beta, with draft/proposed concepts) by providing adequate (human and machine readable) documentation.
17. **Use best practices for modelling, implementing and using SR whenever possible.** For example, avoid ontology overload by keeping ontologies (knowledge models) separate from the instances (and provide mechanisms for keeping their alignment(s) in sync); preference given to smaller, modular, interoperable ontologies over monolithic structures.
18. **Reuse existing resources as much as possible (e.g., concepts, vocabularies, metadata scheme). Create new resources only if nothing reusable exist (then, try to align with related/more generic resources if possible).**
19. Consider joining/forming a community of practice for the creation and long-term maintenance of SRs, especially for those widely used in the community.
20. When alignments are needed, check the possibility of reusing existing “hubs” to interconnect SRs so as to avoid the proliferation of 1-1 alignments.
21. **Promote (or, if none available, create) standards for expressing alignments between resources expressed in different formats.**
22. **Develop metrics to assess resources usage.**
23. Produce recommendations and training on key issues, such as:
  - collaborative work, e.g., intellectual properties and licenses, modeling and use of provenance information;
  - **reusing (parts of) resources to create new ones.**

---

<sup>6</sup> <https://www.force11.org/group/fairgroup/fairprinciples>

24. **Communicate** the benefits of semantics through case studies, best practices, lessons learned, shared experiences, success stories.

### 3. Developers of tools/platforms consuming SRs

25. **Stay abreast of development in semantic technologies.**
26. Include functionalities for searching, discovering, documenting, visualizing, etc., data according to the SR used.
27. **Integrate semantics (features and resources) into mainstream tools and services** (data repositories, analysis platforms, extract-transform-load tools...) to describe, annotate and/or structure data and services. It is particularly important that data collection tools utilize shared, machine actionable semantics for data structuring and documentation.
28. **When possible, implement semantically enabled data types into data, tools and services** to describe common data features used in agriculture and nutrition, e.g. measure units, parameters for experimental or observational data, soil properties, etc.
29. **Support conversion of data to and from RDF (import/export).** In particular:
  - covering the most widely used formats,
  - according to appropriate, normative, semantic models (eg INSPIRE)
  - in a reproducible manner.

### 4. Data managers, data producers

30. **Stay abreast of development in semantic technologies.**
31. **Make explicit (searchable and findable) what SR(s) are used within your dataset.** To this end, pay special attention to the associated metadata scheme, and chose those that also allow for the description of the SR used.
32. When not available, **develop semantically enabled data types** for common features used to describe data in agriculture and nutrition, e.g. measure units, parameters for experimental or observational data, soil properties, etc.
33. **Prefer data collection tools which support semantically enabled data types** to structure and document the data.
34. If alignment/mapping between datasets are created, **provide documentation for the processes as well as for the result.**

### 5. Policy makers & Funders

35. **Foster the development of a generic, extensible and web-based, framework to work with Semantic Resources.**

36. Encourage and support the creation and long term maintenance of **open tools to manipulate semantic resources**.
37. Support initiatives to increase the **discoverability of resources, including both datasets and semantic** resources and services (e.g., concepts, mappings, SPARQL endpoints, ..).
38. Support the creation and long-term maintenance of resources of strategic importance,<sup>7</sup> such as low-level resources relevant to the domain (e.g., crops, plant and animal diseases, livestock, grazing activities). These resources should be:
  - a. available online and in various formats with PID like URIs, DOIs,
  - b. with rich metadata
39. With the aim to make the technology more available, promote courses and trainings on semantics (e.g., on theoretical foundations, modelling, use and implementation of tools), to be proposed either as stand-alone courses, or included in discipline-specific education programs.

## Annex 1 : References to Requirements

Recomm endation #	Statement	Distilled from requirements #
<b>High Level Recommendations</b>		
HLR1	Foster the development of a generic framework to work with SRs.	Rq1 Rq2 Rq3 ReqC1 ReqC2 ReqC4 ReqC5
HLR2	Support initiatives aimed at developing tools and resources for alignment	ReqM3 ReqM4

<sup>7</sup> In particular, the Global Agricultural Concept Scheme ( GACS, see <http://agrisemantics.org> ) which is intended to serve as a hub of concepts to interconnect SRs, data, services, etc. This work has been jeopardized by the lack of political support, despite its recognized interest by the agri-food community.

HLR3	Promote the adoption of common metadata models for the description of semantic resources and alignments and the use of global identifiers	ReqU4
HLR4	Integrate courses on semantics in training programs	
<b>Developers of tools for manipulating semantic structures</b>		
1	Build tools that may be integrated in suits... by choosing standard technologies and shared input/output formats	ReqC2
2	Use open source licenses	ReqC2
3	Build tools that support known best practices	ReqC4 ReqM2 ReqA1 ReqA2
4	Provide user documentation on your tools	ReqC3 ReqU2
5	Take into account the profile of non-semantic experts	Rq1 Rq2 ReqC1 ReqC5 ReqM6
6	Implement (semi)automatic generation of metadata	ReqA2
7	Guide users in choosing appropriate format and modelling approach	ReqC3
8	Allow for the customization of editing and alignment tools	Rq3 ReqC1 ReqM6
9	Include automatic quality check	ReqC3
10	Connect tools or services to SR repositories in the user's domain(s) of interest in order to facilitate the reuse of resources	ReqC3 ReqC4
11	Implement notification mechanisms to inform users on new version of SRs	ReqU1
12	Include concept/class alignment advanced features in SR editing tools, integrate state-of-art alignment algorithms	ReqC2 ReqM1 ReqM6
13	Add alignment discovery services	ReqM2

		ReqM6
<b>Semantic professionals</b>		
14	Make your SR easily findable, accessible, interoperable and reusable	ReqA1
15	Publish new versions in repositories that handle versions	ReqU1
16	Allow for the possibility of exposing/publishing SR at different levels of maturity	
17	Use best practices (e.g., avoid ontology overload)	ReqSR2
18	Reuse existing resources as much as possible	ReqU3 ReqC3 ReqSR1
19	Consider joining/forming a community of practice for the creation and long-term maintenance of SR	ReqM3 ReqSR1 ReqSR2
20	When alignments are needed, check the possibility of reusing existing “hubs” to interconnect SRs	ReqM3
21	Promote (or, if none available, create) standards for concept alignment	ReqM4
22	Develop metrics to assess resources usage	ReqU4
23	Produce recommendations and training on key issues	ReqM5
24	Communicate on benefits of semantics	ReqU2
<b>Developers of other tools/platforms consuming SRs</b>		
25	Stay abreast of development in semantic technologies	ReqU2
26	Include functionalities for searching, discovering, documenting, visualizing, etc., data according to the SR used	
27	Integrate semantics into mainstream tools and services	ReqM5
28	When possible, implement semantically enabled data types into data, tools and services	ReqSR2
29	Support conversion of data to and from RDF	ReqM5
<b>Data managers, data producers</b>		

30	Stay abreast of development in semantic technologies	ReqU2
31	Make explicit (searchable and findable) what SR(s) are used within your dataset	ReqA3
32	When not available, develop semantically enabled data types for common features used to describe data	ReqSR2
33	Prefer data collection tools which support semantically enabled data types	ReqA2
34	If alignment/mapping between datasets are created, provide documentation for the processes as well as for the result.	ReqA3
<b>Policy makers &amp; Funders</b>		
35	Foster the development of a generic, extensible and web-based, framework to work with Semantic Resources	Rq1 Rq2 Rq3 ReqC1 ReqC2 ReqC4 ReqC5
36	Encourage and support the creation and long term maintenance of open tools to manipulate semantic resources	ReqC2 ReqM1
37	Support initiatives to increase the discoverability of semantic resources and services	ReqA1 ReqA2 ReqA3
38	Support the creation and long-term maintenance of resources of strategic importance	ReqM3 ReqSR1 ReqSR2
39	Promote courses and trainings on semantics	