RDA Agrisemantics WG Case Statement

September 1st 2016

Rationale

Agricultural information encompasses **diverse types of data**, from policy, technical and scientific textual documents, through visualizations of temporally and spatially explicit information, to scientific data and models to allow nowcasting and forecasting of agricultural outcomes. Many, if not all, of these types of data are regularly sought after and used by experts in agriculture and poverty reduction, in order to understand the current situation in a given geographical context, identify current and predicted trends, and define actions and policies to implement. However, much of that information is described and catalogued using natural language and **ad-hoc semantics**, which limits the potential for unified search across different information systems, and makes the use and reuse of data for tasks such as impact assessment, extraction of trends or projections into the future extremely laborious and hardly replicable.

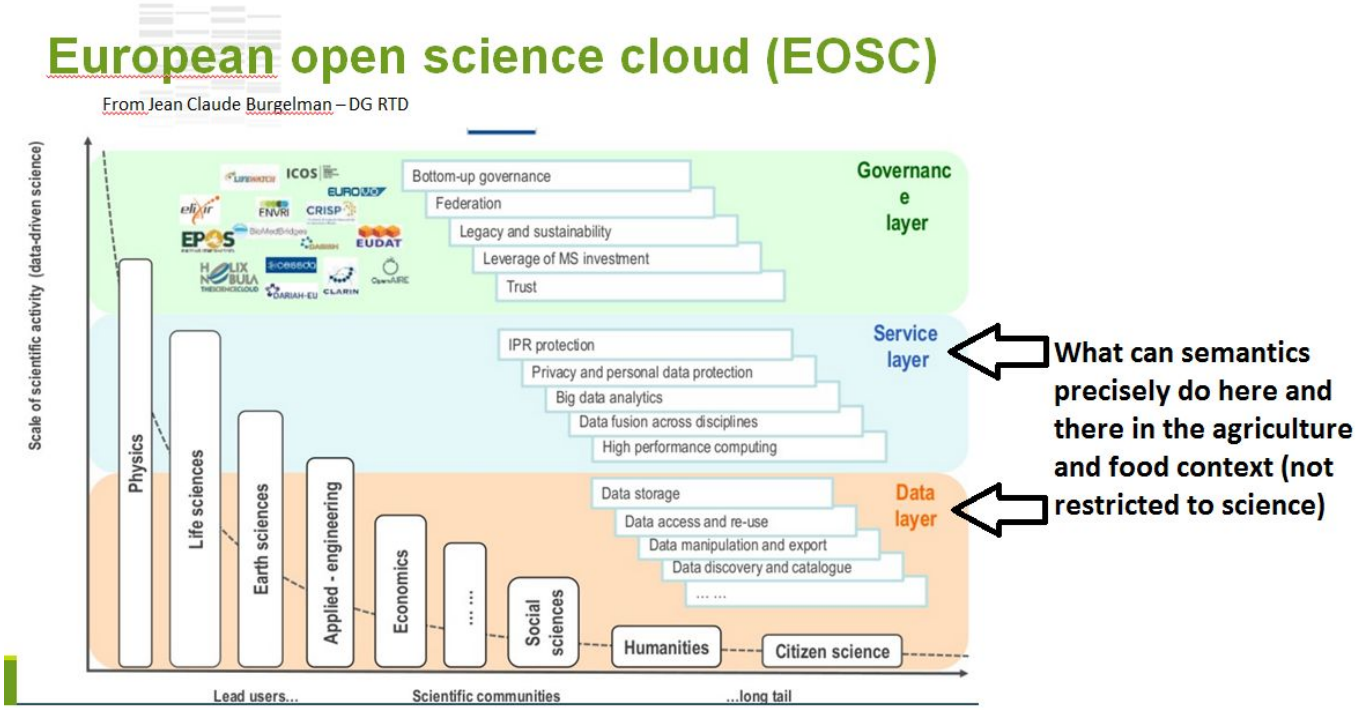
"Semantic Interoperability is usually defined as the ability of services and systems to exchange data in a meaningful/useful way."[1] In practice, achieving semantic interoperability is a hard task, in part because the description of data (their meanings, methodologies of creation, relations with other data etc.) is difficult to separate from the contexts in which the data are produced. This problem is evident even when trying to use or compare data sets about seemingly unambiguous observations, such as the height of a given crop (depending on how height was measured, at which growth phase, under what cultural conditions, ...). Another difficulty with achieving semantic interoperability is the lack of the appropriate set of tools and methodologies that allow people to produce and reuse semantically-rich data, while staying within the paradigm of open, distributed and linked data.

The use and reuse of accurate semantics for the description of data, datasets and services, and to provide interoperable content (e.g., column headings, and data values) should be supported as community resources at an infrastructural level. Such an infrastructure should enable data producers to find, access and reuse the appropriate semantic resources for their data, and produce new ones when no reusable resource is available. The Agrisemantics working group aims at being a community hub for the diffusion of knowledge and practices related to semantic interoperability in agriculture, and to serve a common place where the future of data interoperability through semantics will be envisaged.

International context & related actions

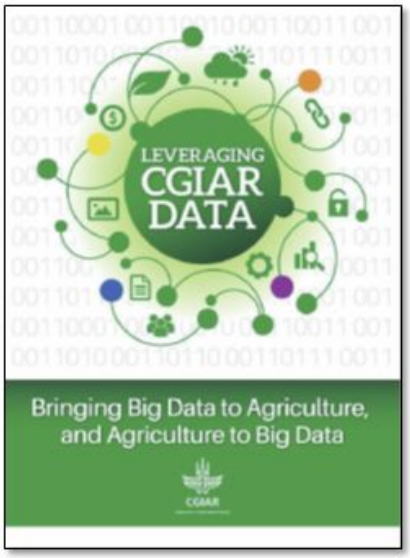
Many communities and institutions are organizing themselves to develop open data policies, virtual environment for research and distributed infrastructure to enable them. The diagram below (J-C Burgelman keynote http://bit.ly/1C03XWf) represents the importance and use of data in a few broad domain of knowledge.

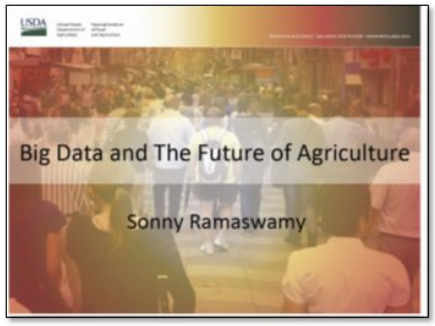
[1] RDA DFT group definition: http://smw-rda.esc.rzg.mpg.de/index.php/Semantic_Interoperability



One can notice that physics and citizen science are mentioned as the branches of science producing the most and least of data respectively. This is especially understandable in the European context where CERN alone is regularly producing impressive amount of data, and where citizen science is still predominantly a niche. One can also notice that agriculture is not mentioned anywhere in the graphic, leaving one to wonder what the reason for this is. We believe that there are two causes. On the one hand, the domain is simply behind on this path, as evidenced by the small number of open data sets in agriculture. On the other hand, agriculture is a broad domain, and any assessment of agricultural systems cannot do without knowledge about species and varieties, climate, and cultural and economic issues, to mention only a few. This latter consideration further emphasizes the importance of data sharing and semantic interoperability, and pushes us to work on a new version of the graphics, where agriculture appears as a topic, and as a community able to reuse and process data produced within different contexts.

There have been several reports established by important agriculture organizations (see below) about the importance of data in agriculture. A number of projects and initiatives are currently ongoing that relate to one or more of semantics, agricultural data and e-infrastructure. Initial participants in the WG are associated to one or more of those listed below, which ensures that relevant work produced in those projects is reused within the WG and diffused to the community of peers. Such close ties with relevant ongoing initiatives will also allow the Working Group to reach a wide audience and so effectively propagate its results.





e-ROSA (Towards an e-infrastructure Roadmap for Open Science in Agriculture, 2017-2019) is an 18-month H2020 coordination and support action[2] financed by the European Commission, co-led by the French National Institute for Agricultural Research (INRA), Alterra - Stichting Dienst Landbouwkundig Onderzoek and Agro-Know, in association with the Food and Agriculture Organization (FAO) of the United Nations. Through a foresight approach, the project will build a shared vision of a future sustainable e-infrastructure for research and education in agriculture and make it operable through pragmatic recommendations that will be reflected in a common roadmap.

2 https://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/23332-infrasupp-03-2016.html

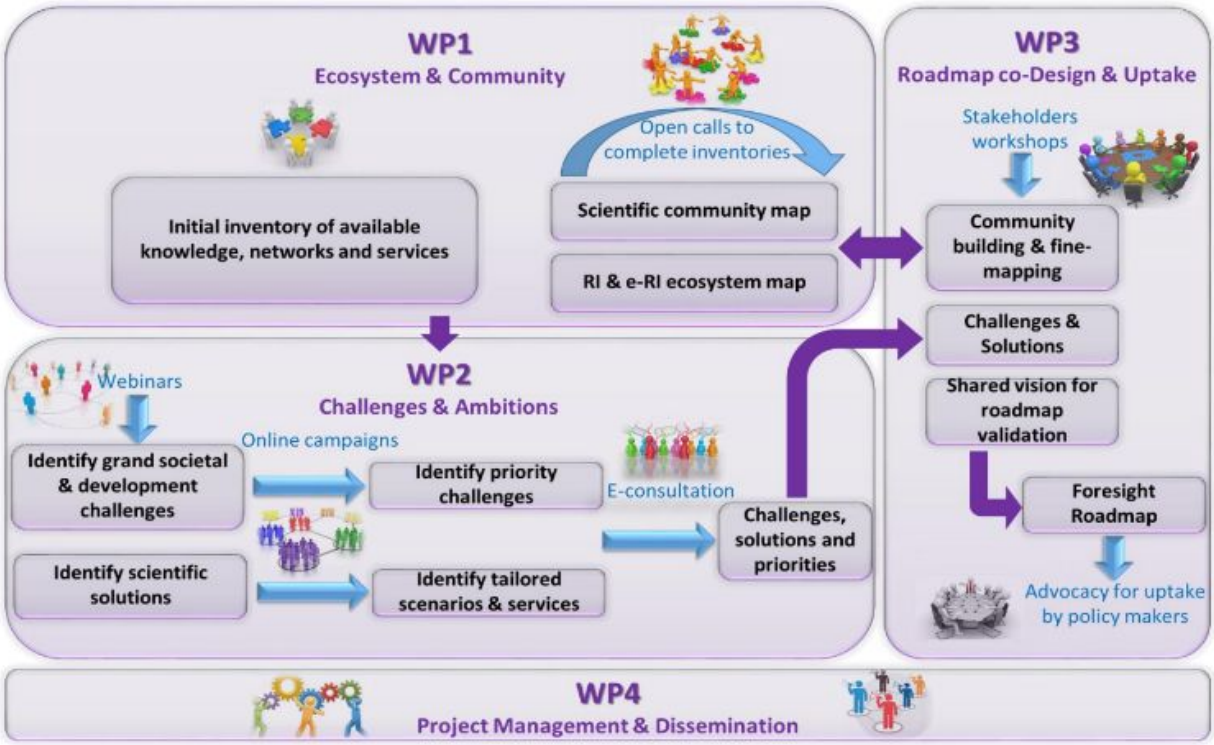


Figure 2. Workflow of the e-ROSA project

GODAN Action (2016-2019) is a 3,5 year project funded by the Department for International Development UK, and led by Wageningen UR, The aim of the project is to produce more evidence-based decision making, to improve service delivery in agriculture and nutrition, databased business creation, actor empowerment and increased transparency of decision making. GODAN Action has identified three focal areas where research and capacity building investment is urgently required. The area relevant to the WG is interoperability of agriculture and nutritional open data (including metadata schemes, semantic standards and classification systems), with the goal to increase use of standards in data management and information systems, and increase collaboration between agricultural and nutritional open data standards initiatives.

GACS (Global Agricultural Concept Scheme, 2014-2016) was a working group formed by FAO, CABI and NAL to explore the possibility of a common repository of conceptual and terminological information related to agriculture. Shaped as a project funded by in-kind contribution, GACS brought together the thesauri of the three organizations - AGROVOC, CAB Thesaurus, and NAL Thesaurus, respectively and produced a “core” of about 15,000 multilingual concepts shared by the three thesauri. The first result of the GACS working group was presented in May 2016, and after a phase of community feedback, the workplan of GACS is currently being updated.

AgroPortal (http://agroportal.lirmm.fr/) is an inter-institutional, LIRMM-supported initiative, a web ontology portal which features ontology hosting, search, versioning, visualization, comment, recommendation, enables semantic annotation, as well as storing and exploiting ontology alignments. It reuses the technology developed by the National Center for Biomedical Ontology within the BioPortal project. AgroPortal gives access to a set of

semantic assets relevant to agriculture and nutrition. It embodies a distributed approach to ontology management but central registry and storage, and unified querying functionalities. As of September 2016, AgroPortal offers 51 ontologies dedicated to agronomy, food, and plant and the repository is intended to grow in the next months.

agINFRA+ (2017-2020) is a EC-funded project embracing the vision of an open and participatory data-intensive science. agINFRA+ continues the work carried out within the agINFRA project (2011-2014) on the design and implementation of a data infrastructure. It makes use of core e-infrastructures such as EGI, OpenAIRE, EUDAT and D4Science, in order to provide a sustainable channel addressing adjacent but not fully connected user communities around agriculture and food.

AGRIS (http://agris.fao.org/) (International System for Agricultural Science and Technology) is a global public database providing access to bibliographic information on agricultural science and technology. The database is maintained by CIARD, and its content is provided by participating institutions from all around the globe that form the network of AGRIS centers. One of the main objectives of AGRIS is to improve the access and exchange of information serving the information-related needs of developed and developing countries on a partnership basis. At the same time, AGRIS is a collaborative network of more than 150 institutions from 65 countries, maintained by FAO of the UN, promoting free access to agricultural information and covering the wide range of subjects related to agriculture, including forestry, animal husbandry, aquatic sciences and fisheries, human nutrition, and extension.

NOAW (No Agro-Waste) is a H2020 project (2016-2020) involving 32 partners from 10 european countries, China and Taiwan. Driven by a “near zero-waste” society requirement, the goal of NoAW project is to generate innovative efficient approaches to convert growing agricultural waste issues into eco-efficient bio-based products opportunities with direct benefits for both environment, economy and EU consumer. INRA will test and extend the semantic web platform @Web (http://www6.inra.fr/cati-icat-atweb) to annotate a selection of experimental data produced during the project. Those annotated data will be reused in specific decision supports systems (DSS) (by example food packaging selection DSS). Ontologies used to annotate data are automatically replicated on the AgroPortal. NOAW is a candidate use-case for Deliverable 2 and could contribute to develop recommendations in Deliverable 3.

The **Integrated Modelling partnership** (IM) is a global consortium dedicated to supporting the production of semantically rich, consistent and interoperable scientific data and models. IM uses high level domain ontologies, a semantic annotation language (k.IM), and an open source software stack (k.LAB) to achieve reuse and interoperability of data and model artifacts in the ecological, social, agricultural and economic domains. Particular emphasis is put on to the use of open standards and controlled vocabularies maintained by authoritative bodies, that the k.LAB software links with its formal ontology framework. The flagship project of the IM partnership is ARIES (http://aries.integratedmodelling.org). ARIES is dedicated to the assessment and valuation of ecosystem services, and used worldwide by policy makers

and scientists for environmental assessment and decision in coupled human-natural systems.

The **Wheat Information System** (www.WheatIS.org) is an international project supported by the Wheat Initiative, a G20 initiative. It promotes and improves standards for data sharing in a common web portal, backed by a distributed search engine, therefore supporting the wheat research community. The IGAD Wheat Data Interest Group has built for the WheatIS a cookbook to disseminate good practices for semantically rich data sharing in the WheatIS community, including key research institute like INRA, BBSRC, CSIRO, TGAC, EMBL-EBI, Rothamsted and USDA.

The **Crop Ontology** (www.cropontology.org) initiated by the Generation Challenge Programme (GCP, http://www.generationcp.org/) provides a framework and a portal for trait and agronomy controlled vocabularies building and sharing. Those are used in the CGIAR research centers and their usefulness has been demonstrated in the field for years. The crop ontology good practices are used in other systems and are disseminated through the WheatIS cookbook. They are used in key european projects like Elixir or EPPN.

AgGateway (www.aggateway.org) is a non-profit consortium of over 200 organizations, primarily in the private sector, dedicated to the implementation of data standards for interoperability in agriculture. AgGateway has operated for over 10 years in supply-chain processes. For the last 5 it has also been active in field operations, where it has multiple active projects: SPADE (interoperability in field operations, grain handling and asset management), PAIL (interoperability in irrigation and observations & measurements), and ADAPT (implementation of a common object model and format conversion framework). All of these have strong semantic interoperability components, including the creation of variable registries, reference data APIs, and leveraging of existing controlled vocabularies..

Charter

The RDA Agrisemantics WG aims at serving as a community-driven initiative to advance on a shared view on the role of semantics for data sharing and interoperability, especially for data related to agriculture. The Working Group will define a high level view on the use of semantics assets and on the needed infrastructure to achieve better data sharing and interoperability in agriculture.

The WG intends to organize its work in phases. First, the group will set a common ground for the working group and the community (e.g., shared definition of infrastructure, semantic interoperability, and the general view of an agrisemantics landscape). Then, the group will proceed and identify a few use cases to facilitate the sharing of ideas community-wide, and will then move and collect functional, non-functional and social requirements for the effective use of semantics to improve data interoperability. The **final output of the working group will be a set of recommendations for designing the future of semantics to enable interoperability of agricultural data. The document is expect to serve as a roadmap for future design of semantically-oriented software and data infrastructural components.**

During its activities, the group will consider the various aspects involved in an infrastructure, especially software and services for the various stages of a data lifecycle. The data publication on the Web using linked open data principles will be particularly encouraged. Moreover, semantic assets (ontologies, thesauri, code lists, ..), data and metadata schemes for various areas related to agriculture will be considered.

The WG intends to:

1. Liaise with the GODAN Action on a survey the project is going to distribute in late 2016, to gather information on the use of semantic assets to describe datasets. The WG will build its activities also on the information collected through the survey.
2. Liaise with INRA on the bibliometric study on academic publications that it is going to carry on to identify key players and most studied semantics related issues, better qualify which agricultural subdomains already benefit from semantics and which do not, and to what purposes.
3. Liaise with GODAN Action to provide input to the project deliverable on mapping of standards, by reporting on the technical solutions and practices adopted by WG members for editing, mapping, publishing and using semantic assets; and finally identify possible difficulties faced by the community.
4. Promote the reuse and sharing of semantic assets, following the W3C semantic web format, to describe agricultural data.
5. Identify a set of Use Cases to represent various domains and situations in which the correct semantic interpretation of agricultural data is crucial for building systems or support analysis. Subsequent work on requirements and a roadmap will build on the Use Cases.
6. Study e-infrastructures involving semantics in other domains (biomedicine, environment, pharmaceutics, etc.)
7. Liaise with the e-ROSA project to select infrastructural component related to the management, use or reuse of semantic assets for which recommendations should be developed.
8. Liaise with the GACS Working Group as a key stakeholder in the area of maintenance and use of semantic assets.
9. Liaise with other RDA groups (see next section for details) to reuse their outputs and coordinate efforts and vision.
10. Liaise with private sector initiatives such as those of AgGateway, to identify common semantic interoperability problems and work toward common solutions.

Value proposition

Individuals, communities and initiatives that will benefit from the WG recommendations.

The WG recommendations will be a major contribution to the e-ROSA roadmap. Specifically, the WG Recommendations will contribute to the part of the roadmap that concerns infrastructure to support editing, access, publication and use of semantic assets.

As a consequence, the WG recommendations are expected to benefit the private sector by highlighting services and requirements to be taken into considerations when developing new products.

The recommendations developed by the WG are intended to benefit data and information managers dealing with semantic assets, by pointing at future directions in the field.

Key impacts of WG recommendations

The tangible impacts expected from the WG output:

1. Spread knowledge within (and outside) RDA on the role of semantics for addressing interoperability issues.
2. Guide agricultural stakeholders into the development of semantic solutions and products using standards and shared tools, services and practices for data maintenance, publication and use.
3. Produce a community-built proposal for the role and requirements for the development of an infrastructure that supports semantics for data interoperability in agriculture.
4. Move semantics into the agenda of EC funding frameworks.
5. Present to funding bodies the “why” and “what” should be promoted for the advances of knowledge and policy-making in agriculture.

Engagement with existing work in the area

The Agrisemantics WG is a working group within the RDA Interest Group on Agricultural Data (IGAD). The working group will closely liaise with the other working group within IGAD, primarily the now concluded Wheat Data Interoperability to reuse the part of their output on vocabularies, and the soon to be formed group on Rice Data Interoperability.

Other Working and Interest groups in RDA are also relevant to the Agrisemantics WG. In particular, we have identified the following groups for participation and exchange of information and feedback:

- Vocabulary Services Interest Group
- BoF on Domain Vocabulary Development, Standardization, Registration, Harmonization and Support
- Repository Core Description WG
- Metadata IG

New relevant development within RDA will be closely monitored, to ensure that no effort is duplicated and that the impact of RDA outputs is maximized.

Moreover, the group will leverage the active involvement of its members in many related and complementary projects already in place (see Section “International context” in this document) to better achieve its goals. The WG will also organise and encourage the

community to participate in education and dissemination events such as workshops, datathons or hackathons (e.g., AgroHackathon Meetup Series).

During the course of its activities, the WG will also actively promote its activities so as to involve more experts in the group, to enhance the possibilities of dialogue in the community.

Work Plan

Form and description of final deliverables

1. A report on “Semantics Landscape for Agricultural Data”, presenting the current situation on semantics for data interoperability in agriculture. This document will serve as a basis for discussion among practitioners and data producers dealing with different types of agricultural data.

2. A report consisting of a set of use cases and requirements drawn from the discussions engaged within the Agrisemantics community.

3. A document on “Recommendations” for the future of semantics for agricultural data and supporting infrastructure. The document will touch upon software, functionalities, and semantic assets to enhance data interoperability in agriculture. This document will build upon the knowledge and experience gained during the first phase of the WG activities, and the results presented in the abovementioned reports. The recommendations will provide a roadmap for future work on infrastructures to support semantic assets.

Milestones

Month of the WG	Timeline (indicative)	Milestone
-	Nov 2016	Agrisemantics panel
M1	Dec 2016	WG starts
M5	Apr 2016	Landscape report out
M8	Jul 2017	Use case and req out
M12	Nov 2017	First version of recommendation report
M14	Jan 2018	Feedback from the community
M16	Mar 2018	Final version of recommendation report

M17-18	Apr-May 2018	Dissemination

WG's mode and frequency of operation

Participation in the WG is voluntary and not every WG member will be able to contribute equally. Therefore, we aim to maximise members' contributions by focussing on members' specific interests, and also ensure that all members can contribute to internal reviews.

Communication will be based on regular online meetings and group writing on shared documents. Extra meetings and events will be organized to carry out specific activities. A few face-to-face meetings and workshops will also be organized, as much as possible co-located with other events so as to maximise participation and feedback. The events currently foreseen are the following:

1) Panel session at MTSR, Nov 2016
2) Kick off meeting of e-Rosa and agINFRA+ projects (co-located), January, 2016 in Paris
3) RDA P9 in Barcelona, Spain (5-7 April 2017)
4) Co-located Workshop with e-ROSA (around June 2017, organized by FAO)
    a) During e-ROSA: WG landscaping results handover to e-ROSA WP1 and community feed for requirements gathering activities of the WG
    b) After e-ROSA: WG workshop (allowing participation of external experts)
5) AgroHackathon 2017 (June, possibly in Montpellier)
6) RDA P10 in TBD (around September 2017)
7) Co-located Workshop with e-ROSA (around December 2017, organised by INRA)
    a) During e-ROSA: get inspired by presentations of grand challenges, innovative ideas/solutions and selected visionary speakers
    b) After e-ROSA: WG workshop to initiate or continue writing down our recommendations
8) AgroHackathon 2018 (June, possibly in Montpellier)

Achieving consensus, addressing conflicts, and staying on task and within scope

- Consensus will be reached via open discussion, voting, and majority considerations informed by evidence where possible.
- Conflict will first be addressed by WG leaders. An escalation procedure will be drafted, for example the RDA Council will be consulted, and an independent person not in the WG will be brought in to mediate the conflict.
- Staying on task and within scope: the WG leaders have good experience in projects management and standards development. The key mechanism for reaching consensus will be through examining evidence and identifying limitations of applicability of competing ideas. In addition, of course, we will agree on a detailed schedule and track action items.

Planned approach to broader community engagement and participation

The WG creation will be announced widely using RDA communication means, including participation to P8 IGAD meetings, mailing lists, social networks. The WG launch is planned to take place (or at least announced) at the MTSR conference (22-25 November 2016, in

Göttingen, Germany). The participation of WG members in the AgroSEM special track on Metadata and Semantics for Agriculture, Food & Environment should attract additional domain experts and allow the identification of new use cases.
The activities of the WG will be regularly presented by its members in events (conferences, project meetings, etc) in particular those organized by the agriculture and nutrition communities, in order to get more people involved and have some feedback on ongoing activities. Social networks and RDA communication facilities will be used to inform on the WG activities and outputs.

Adoption Plan

e-ROSA

The H2020 Coordination and support action **e-ROSA** will take into account the “State of the Art” WG output in its planned activity of landscaping the digital agriculture science community and practices (WP1). Our results may enrich the agINFRA portal. The strengths, weaknesses, requirements and gaps in the specific field of semantics that will have been identified by the WG will help e-ROSA describing what is available in terms of data and technical solutions to address grand challenges (WP2). We aim at the recommendations delivered by the WG to be included in the Foresight Roadmap Paper that will be the final output of the H2020 Coordination and support action e-ROSA (WP3).

AgroPortal

Within the WG span, we plan to submit a H2020 e-infra project mainly dedicated to the use of ontologies and the development of AgroPortal. We have already designed and implemented an advanced prototype. With the Agrisemantics partners, we plan to turn that prototype into a real service to the community. We think that this platform can offer a robust and stable reference repository that will become highly valuable for the agronomic domain.

NOAW

INRA will inspire from the recommendations and other outputs of the WG to design and implement the H2020 NOAW project Data Management plan of which INRA is in charge. INRA will also communicate WG recommendations to partners of the project.

GACS

The Global Agricultural Concept Scheme (GACS) Working Group concluded its first phase of activities in May 2016. The product of that first phase was a set of some 15,000 concepts in up to 25 languages, originated from three major thesauri in the area of agriculture, namely AGROVOC, CAB Thesaurus and NALT. Work is planned to continue from late 2016 through 2017, placing a stronger emphasis on topic-driven sets of concepts, to be used as reference identities for the expression of the semantics of data. The future work of GACS is expected to be an important source of requirements and first-hand experience in the use of semantic assets in the context of textual documents as well as numeric data.

FAO

As an active member of the GACS working group, FAO is going to actively participate also in the RDA Agrisemantics WG, to contribute to the community-wide discussion promoted by RDA with its experience in terms of requirements and recommendations. After validation by the community, FAO will promote internally to the organization the adoption of the WG recommendations for what concerns the lifecycle of metadata production.

k.LAB

As k.LAB uses formal semantics to provide access to distributed datasets and allow their use in dataflows, recommendations and vocabularies endorsed by the WG will be adopted as authoritative in all semantic annotations related to the themes of interest of the WG. This will affect all users and developers of projects for which k.LAB provides the enabling infrastructure, such as ARIES.

Use Cases
The providers of the collected use cases are expected adopters of the WG recommendations as they will be developed partially upon the requirements formulated from the use cases.

INRA
Recommendations and other outputs of the WG will be adopted by the Department of Scientific and Technical Information to improve the services and tools offered to INRA agents who actually or want to maintain, publish, and use vocabularies and semantized data.

Initial Membership

Brandon Whitehead, CABI
Caterina Caracciolo, FAO
Catherine Roussey, Irstea
Cyril Pommier, INRA
Clement Jonquet, LIRMM
Devika Madalli, ISI, India
Ferdinando Villa, Ikerbasque, Basque Centre for Climate Change (BC3)
François Pinet, Irstea
Gary Berg-Cross, USA
Ivo Pierozzi Junior, Embrapa, Brasil
Panagiotis Zervas, Agroknow
Pascal Aventurier, INRA
Patrice Buche, INRA
Pierre Larmande, IRD
R. Andres Ferreyra, Ag Connections LLC and AgGateway, USA
Richard Finkers, WUR

Simon Cox, CSIRO, Australia
Sophie Aubin, INRA
Tom Baker, FAO consultant

List of acronyms

AGRIS	International Information System for the Agricultural Science and Technology
ARIES	ARtificial Intelligence for Ecosystem Services
BBSRC	Biotechnology and Biological Sciences Research Council
CABI	Centre for Agricultural Bioscience International
CERN	European Organization for Nuclear Research
CGIAR	Consultative Group on International Agricultural Research
CIARD	Coherence in Information for Agricultural Research for Development
CSIRO	Commonwealth Scientific and Industrial Research Organisation
EC	European Commission
EGI	European Grid Infrastructure
EMBL-EBI	European Molecular Biology Laboratory - European Bioinformatics Institute
Embrapa	Brazilian Agricultural Research Corporation
EOSC	European Open Science Cloud
EPPN	European Plant Phenotyping Network
FAO	Food and Agriculture Organization
GCP	Generation Challenge Programme
GACS	Global Agricultural Concept Scheme
GODAN	Global Open Data for Agriculture and Nutrition
H2020	Horizon 2020
IG	Interest Group
IGAD	RDA Agricultural data interest group

IM	Integrated Modelling
INRA	Institut National de la Recherche Agronomique
IRD	Institut de Recherche pour le Développement
Irstea	Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture
ISI	Indian Statistical Institute
LIRMM	Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier
MTSR	Metadata and Semantics Research Conference
NAL	National Agricultural Library
TGAC	The Genome Analysis Centre, now Earlham Institute
UN	United Nations
USDA	United States Department of Agriculture
WG	Working Group