GACS: a FAIR investment in open agricultural knowledge

Sustainable agricultural value chains and global food security cannot be addressed without intelligent use of data. When data is also 'FAIR' - Findable, Accessible, Interoperable, and Reusable, open data on weather, soil, production, market prices, pests, and diseases can lead to faster and effective decision-making while increasing the reproducibility, transfer and impact of scientific discoveries.

The impact of FAIR data increases by an order of magnitude when the information is mapped to a common descriptive framework -- semantics -- which allows both humans and machines to understand and make use of the data and relationships within, and between, them, making them more interoperable and reusable.

The Global Agricultural Concept Scheme (GACS) is the first step toward a mapping of the data and relationships through a community of interconnected, interoperable, semantic assets and services relevant to agriculture and food security. GACS will transform massive *data silos* to a more reusable *web of data and services*.

GACS was derived from three well known and trusted sources within the community, AGROVOC, NALT and CABT. GACS is currently situated as a synthesised set of 15,000 terms that map across all three thesauri. This is a key contribution. If any data, or web, is mapped to one GACS term, it is mapped to all three -- an exemplar of data collaboration and interoperability.

What problems can GACS help us solve?

Making public data transparent: Donors of agricultural research spend billions of dollars on hundreds of projects. Due to the structure of many donor organizations, it's almost impossible to know and understand every project that is currently underway, let alone the data and results that emerge. Although several projects may be operating in the same region or on the same crop, the results and publication of the project are restricted to the specific grant. With GACS, the research, statistical and programmatic data that emerges each project would be mapped to each other, thus increasing the amount and quality of information to make better decisions. This would allow the donors to understand the impact of their funding, leading to higher return on investment, higher quality research and science, and innovation that everyone can benefit from.

Other uses include: Integrating corporate research with public data for more efficient R&D; Linking commercially published research with related resources and datasets; Compiling statistical indicators on Sustainable Development Goals (SDGs); Improving visibility of research in non-English languages; Naming common animals, crops, and diseases unambiguously; Making sense of research archives.

CABI as a global provider of information, skills and tools to help solve problems in agriculture and the environment promotes GACS as a key element in building global open knowledge infrastructure for agricultural discourse that allows us to solve wicked problems more intelligently.

GACS already exists as a strong proof of concept. We now seek support and donor funding of €1 million to pilot an operational service using real world cases. Using the knowledge of operational challenges and with feedback from users of the pilot service, a fully-production level service will be specified involving SLAs, support, and data governance team. Key to this 'production' specification is the establishment of a fully self-sustaining business model. This model, along with commitments to long-term availability of the data, information models and services, will be essential in bringing government, academic and commercial organisations to the platform.

For more information see our full vision for GACS that follows.

GACS vision within Agrisemantics for a food secure world

By GACS WG & Agrisemantics WG

Executive Summary

Challenges in agricultural production and food value chains can be addressed in part by improving access to research results, advice for farmers, and data analyses. Intelligent use of data can make agriculture and its value chains more efficient. Open data on weather, soil, production, market prices, pests, and diseases can level the playing field for all actors in the value chain, not just in farming but in husbandry, horticulture, forestry, fisheries, and aquaculture. Open data can increase the reproducibility, transfer and impact of scientific discoveries especially when data is also 'FAIR' - Findable, Accessible, Interoperable, and Reusable. One way to achieve the FAIR data vision is by a wider adoption of Semantic Web principles, semantically describing and interconnecting data to achieve the vision of a Web of Data (or Linked Open Data). Health sciences and biomedicine have made considerable steps in implementing semantic web principles, but agriculture is lagging behind. In recent years various components are starting to come together in a framework termed 'Agrisemantics'.

The Global Agricultural Concept Scheme (GACS) is one of the key components of the Agrisemantics framework. FAO with the USDA National Agricultural Library (NAL) and the Centre for Agriculture and Biosciences International (CABI) have collaborated to create GACS, a concept scheme mapped to the most frequently-used concepts in AGROVOC, NAL Thesaurus, and CAB Thesaurus. With support from the Global Open Data for Agriculture and Nutrition initiative (GODAN) we have built a broader partnership for nurturing GACS and for positioning GACS as a major component of Agrisemantics, a future community network of interconnected semantic assets and services relevant to agriculture and food security including Agroportal and the GODAN Map of Standards. A number of tangible outcomes that can be readily enabled, or achieved more cost-efficiently using Agrisemantics have already been identified including: Making public data transparent; Integrating corporate research with public data for more efficient R&D; Linking commercially published research with related resources and datasets; Improving the reusability of donor-funded grant results; Compiling statistical indicators on Sustainable Development Goals (SDGs); Improving the visibility of research in non-English languages; Naming common animals, crops, and diseases with global URIs; and Providing global identity to entities matched to text patterns.

We describe a straightforward and pragmatic route towards a self-sustaining operating model for GACS within the wider Agrisemantics landscape based on phased development

from proof of concept, to pilot, to production.

Building a semantic web of food and agriculture

Securing food for a projected nine billion people will require significant increases in agricultural production, and value chains will need to become more efficient to reduce waste. These challenges can be addressed in part by improving access to research results, advice for farmers, and data analyses. Intelligent use of data can make agriculture and its value chains more efficient. Open data on weather, soil, production, market prices, pests, and diseases can level the playing field for all actors in the value chain, from smallholder farmers to governments, universities, and businesses; not just in farming but in husbandry, horticulture, forestry, fisheries, and aquaculture. Scientific communities have realized the importance of open data in increasing the reproducibility, transfer and impact of scientific discoveries. We have recently seen the emergence of the FAIR data principles - Findable, Accessible, Interoperable, and Reusable[1] - within a supportive data ecosystem[2]. One way to achieve the FAIR data vision is by a wider adoption of Semantic Web principles. The Semantic Web offers a panel of standards and technologies, endorsed by the W3C, to semantically describe and interconnect data to achieve the vision of a Web of Data (or Linked Open Data). Although the semantic web principles have been largely adopted by some important scientific communities, for example health sciences and biomedicine with measurable impact, the vision still needs to be realized in food and agriculture.

Data that is interoperable. A dataset is useful on its own, but its value increases if it can be integrated with other datasets, perhaps in unforeseen ways, to yield new insights. Data that plays well with other data is said to be 'interoperable'. Potentially interoperable datasets can be discovered through their metadata - structured descriptions of their sources and contexts, tagged according to their topics, such as 'rice' or 'rainfall'. Such topical tags may be selected from hierarchically organized vocabularies, thesauri, or concept schemes which capture a certain level of 'semantics'. Datasets are themselves structured by schema elements, field names, spreadsheet column headings, or properties. The semantics of data and metadata may be hardwired into specific applications or described in globally available ontologies, which define the meaning of these elements as well as how they are interrelated.

Data that is open and reusable. Data can be used to drive change in agriculture and food value chains. The more data can be integrated, the more insights can result. Data has the biggest impact when it is openly reusable for multiple uses.[3] If data providers invest in the interoperability of open data sources, data can be easily discovered and assessed for reuse

[1] https://www.force11.org/group/fairgroup/fairprinciples
[2] http://www.godan.info/documents/data-ecosystem-agriculture-and-food
[3] http://www.godan.info/documents/open-data-farming-101

enabling service providers to create more value added products to meet global food challenges. The GODAN initiative supports the proactive sharing of open data to make information about agriculture and nutrition available, accessible and reusable in support of urgent challenges in ensuring world food security and therefore has taken on the task to promote GACS development.[4]

Data that is based on global semantics. Open data is optimally reusable when it is based on global semantics. The semantics in thesauri, concept schemes, classifications, vocabularies, and ontologies can be declared global by leveraging the infrastructure of the World Wide Web to assign globally unique identifiers (URIs[5]) to their concepts and properties[6]. When data is published on the Web, using URIs as identifiers for their semantics, it is easier to interlink related elements among multiple online datasets ("Linked Open Data"). The techniques of Linked Open Data overcome many of the limitations of traditional information technology by expressing mappings and data semantics with global identifiers that can be looked up in schemas published on the Web, turning the Web into a vast, distributed dictionary for knowledge organization.

What does global semantics enable ?

The creation of innovative tools, techniques, services, and applications in agriculture is made possible when data that is comparable, well-linked, and interoperable. Tangible outcomes that can be achieved more cost-efficiently, through the adoption of global semantics include:

- **Making public data transparent.** Government initiatives such as the Open and Transparent Water Data Act in the US are tasked with making all of the data they publish interoperable, reusable, and thus more valuable. This requires publishing data available in ad hoc local formats in widely available standardized formats, models and vocabularies.
- **Integrating corporate research with public data for more efficient R&D.** Syngenta must frequently clean and normalize its internal data for compatibility with public sources to enable new analyses. This can be done more efficiently by using global semantics curated by domain experts.
- **Linking commercially published research with related resources and datasets.** Commercial STM publishers can add value for their readers by pointing to and recommending related content.
- **Improving the reusability of donor-funded grant results.** Donors such as BMGF, DFID and USAID find that requiring grantees to publish data openly is not enough.

[4] http://www.godan.info/about
[5] https://en.wikipedia.org/wiki/Uniform_Resource_Identifier
[6] By analogy, the binomial Latinate names of Linnaean classification have historically functioned as globally unique names for species.

Annotating the data with terms from common vocabularies would dramatically improve the value of that data and thus the return on investment of the donor.

- **Compiling statistical indicators on Sustainable Development Goals (SDGs).** Reporting on SDGs requires standard approaches and terminology. Agencies such as the Food and Agriculture Organization of the UN (FAO) find themselves in real need of a platform where statistical classifications in agriculture can be shared for reuse.
- **Improving the visibility of research in non-English languages.** The French National Institute for Agricultural Research (INRA), which indexes publications in French, can map to global indexing terms identified with URIs that have multilingual labels, making their work instantly findable by speakers of other languages.
- **Naming common animals, crops, and diseases with global URIs.** Researchers constantly re-compile lists of common entities for local use. It is easier simply to use lists that are already available, and if the entities are identified with URIs, resources described with those URIs become easier to reuse.
- **Providing global identity to entities matched to text patterns.** Text mining techniques are used to identify salient text patterns within a given context in scientific literature. Reference entities commonly detected can be given stable, global identities in the form of URIs from common vocabularies such as GACS.

Progress to date

Early Web semantics for agriculture. The web of global semantics ("Semantic Web") has been forming since the late 1990s - that is, for almost as long as the Web itself. Global semantics for agriculture began with initiatives such as AGROVOC, a decades-old, print-based thesaurus of agricultural terminology maintained at the Food and Agriculture Organization of the UN (FAO) that was transformed into a Semantic Web vocabulary and published on the Web in the early 2000s.

Plant sciences, a domain close to agriculture has already developed advanced ontologies such as the Plant Ontology[7], the Crop Ontology[8], the Environment Ontology[9], and more recent initiatives - for example: Food Ontology, TOP Thesaurus, etc. As more vocabularies and ontologies are produced in the domain, the greater the need to discover them, evaluate them, and manage their alignment and interfaces. This was the driver to create reference vocabulary and ontology repositories such as AgroPortal[10], a platform for ontology-based services for agronomy, plant sciences, food and biodiversity.

[7] http://planteome.org/
[8] http://www.cropontology.org/
[9] http://purl.bioontology.org/ontology/ENVO
[10] http://agroportal.lirmm.fr/

Global Agricultural Concept Scheme (GACS). Over the years, FAO coordinated with the USDA National Agricultural Library (NAL) and the Centre for Agriculture and Biosciences International (CABI) on the improvement of their respective thesauri. Since 2014, in response to feedback from users of their bibliographic databases, the three partners have created the Global Agricultural Concept Scheme (GACS), a smaller concept scheme mapped to the 15,000 most frequently-used concepts in AGROVOC, NAL Thesaurus, and CAB Thesaurus[11]. GACS Beta was formally launched at the GODAN Global Summit in September 2016 by US Secretary of State for Agriculture, Tom Vilsack.[12] The GACS partners have committed to the long-term persistence of its URIs.

Uses for GACS. The 15,000 high-level, lightly defined concepts of GACS constitute a pool of URIs that are already usable:

- For tagging information and datasets for discovery (semantic annotation).
- As building blocks for constructing other, more detailed knowledge organization systems such as ontologies.

All this initiatives are now working together within the 'Agrisemantics' framework.

Next steps in the development of GACS

Extending GACS support for structured data. A living language, GACS can be extended beyond the core requirement; namely the bibliographic description to support concepts frequently used in datasets, by providing key terminology in many languages for crops, soil types, diseases, and other biological entities .

Mapping GACS to domain ontologies. By mapping the generic concepts of GACS to more granular, domain-specific concepts in ontologies, taxonomies, and specialized vocabularies, GACS can function as a switching language, glueing together a diverse collection of loosely compatible domain languages (see Figure 1). Users can focus searches on increasingly specific resources by following links from GACS to more specialized ontologies.

Mapping domain ontologies to data in local formats. Domain ontologies can provide global identity to elements of structured data embedded in a diversity of local formats. For example, the Agrisemantics initiative will explore methods for associating a column in a spreadsheet of field data with a well-defined and well-contextualized concept in a domain ontology, itself linked, in turn, to a generic, high-level concept in GACS (see Figure 2). Giving

[11] http://browser.agrisemantics.org/gacs
[12] See the press release (http://bit.ly/gacs_launch_press_release) and video (http://bit.ly/gacs_launch_video) of the GACS launch.

global identity to local data elements can provide application-specific data formats with a form of semantic authority control, facilitating the normalization of diverse data sources for merging into an application or analysis.

Shifting maintenance responsibility to expert communities. Traditionally, thesaurus editors were responsible for maintaining concepts in all fields within scope of a thesaurus. Ideally, responsibility for maintaining specific types of concept, such as countries, viruses, and organisms, would be shifted to communities of specialists. The Web infrastructure for global semantics could thus be leveraged to create a new division of labour for maintaining shared semantics.

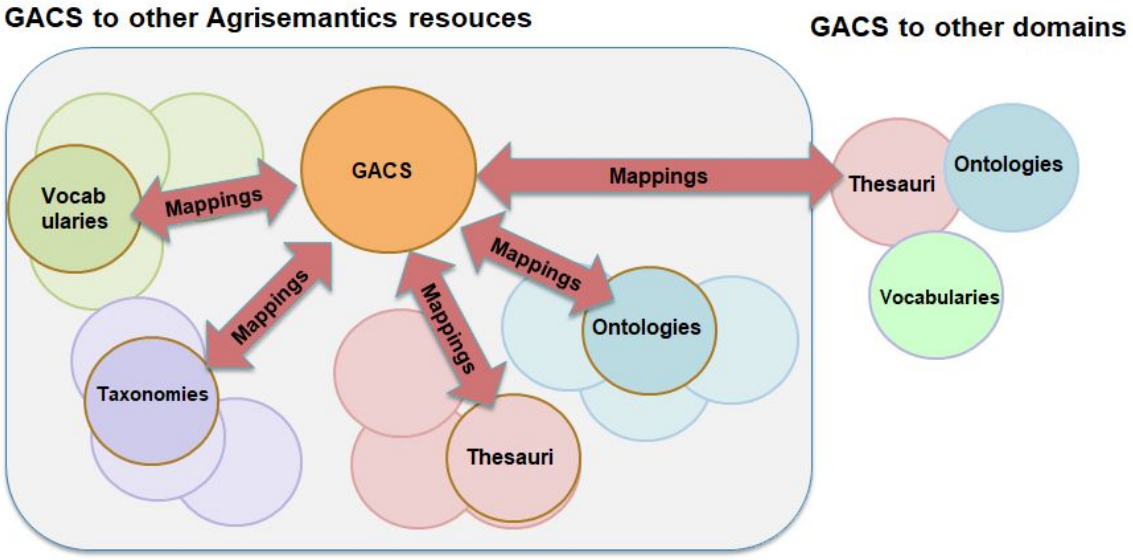

Figure 1. The generic concepts of GACS are mapped to concepts in more granular, domain-specific concepts in ontologies, taxonomies, and specialized vocabularies, thus avoiding laborious many to many mappings.

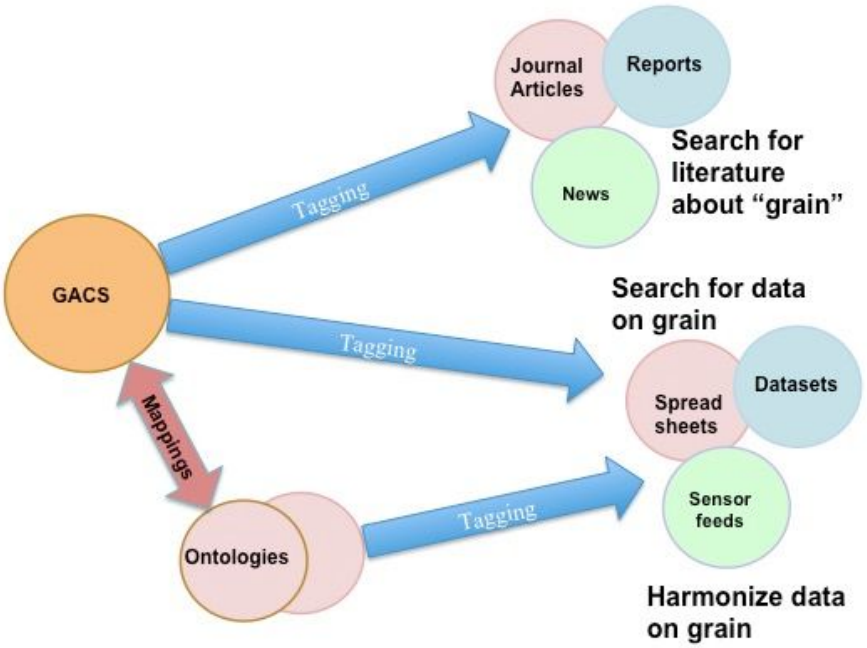

Figure 2. The high-level concepts of GACS can be used to tag, and thus find, information in the form of research reports, news items, and videos, or structured datasets in the form of spreadsheets, statistical tables, and sensor feeds. Domain-specific ontologies mapped to GACS can be used to normalize a diversity of datasets on the basis of global semantics.

GACS in the Agrisemantics framework

With support from the Global Open Data for Agriculture and Nutrition initiative (GODAN) we have built a broader partnership (see Appendix A) for nurturing GACS and for positioning GACS within Agrisemantics, a future community network of interconnected semantic assets and services relevant to agriculture and food security. In order to provide "a well-integrated clearinghouse of machine-readable semantic assets in agriculture and nutrition, such as vocabularies, code lists, ontologies, taxonomies, and statistical indicators", Agrisemantics aims at integrating existing assets such as the following:

- The Global Agricultural Concept Scheme (GACS).
- The Agroportal[13] platform, with its interlinked ontologies from INRA, CGIAR, and other agricultural research institutions and projects.
- The DFID-funded GODAN Map of Standards[14].
- The Crop Ontology[15].

[13] http://agroportal.lirmm.fr/
[14] http://vest.agrisemantics.org/
[15] http://www.cropontology.org/

- Interfaces with industry standards such as those maintained by AgGateway[16] and the Open Ag Data Alliance (OADA)[17].
- Specialist vocabularies on for example land use (LandVoc) or project funding (Open Ag Funding of IATI) are looking for places to publish their own vocabularies and lists.
- VocBench and other tools and services useful to create, map, and maintain semantic resources.

GACS is an essential and pivotal first step towards ontology-based data access and knowledge discovery in agriculture. GACS can be reused by in a number of important ways inside information systems, namely (i) as a query model over data resources permitting end users to search for data, (ii) as a resource for authoring semantically linked data, (iii) a reference vocabulary / gazetteer for text analytics tasks, (iv) as knowledge representation for inference and reasoning about agricultural concepts and relations, (v) as common terminologies for describing the inputs and outputs of web services (making them discoverable and interoperable), (vi) a repository of domain specific rules and definitions for automatically classifying data to semantic categories.

Agrisemantics community and capacity building. Aside from providing a platform where vocabularies can be published in open, machine-readable formats and linked among themselves, the Agrisemantics initiative aspires to build a community of users. This effort will require capacity building in order to help agricultural data and service providers engage with global semantics and use them effectively in their own systems and products.

Achieving and maintaining the Agrisemantics vision

We will take a "disciplined and pragmatic path to progress around the principles of 'thinking big but starting small'..."[18] We have already developed GACS based on resources mobilised by FAO, CABI, and USDA. New partners will provide other in-kind contributions to promoting its organic growth and help plan what it might become.

The Agrisemantics initiative has been inspired by developments in the Life Sciences, where common semantics have been created through systems such as NCBO BioPortal[19] and the UMLS[20]. BioPortal is a one-stop access point for biomedical ontologies; the Unified Medical Language System created for the healthcare sector UMLS "integrates and distributes key terminology, classification and coding standards, and associated resources"[21],[22]. Through integrating Agrisemantics services such as AgroPortal, the Map of Standards, and GACS,

[16] http://www.aggateway.org/
[17] http://openag.io/
[18] Stanley Wood - on the need to develop a data ecosystem for agriculture in http://www.godan.info/documents/data-ecosystem-agriculture-and-food
[19] https://bioportal.bioontology.org/
[20] https://www.nlm.nih.gov/research/umls/
[21] https://www.nlm.nih.gov/research/umls/
[22] Other sources of inspiration include the Financial industry's FIBO system https://www.edmcouncil.org/financialbusiness and OpenPhacts https://www.openphacts.org/ .

Agrisemantics aspires to create efficient and shared semantics in Agriculture and Nutrition, exploiting all possible synergies with the Life Science Community.

The Research Data Alliance (RDA) has set up an Agrisemantics Working Group to define a broad vision and to position the various assets in the framework. The GACS initiative is fully part of this endeavour.

A sustainable business plan for Agrisemantics. Achieving the bolder vision for Agrisemantics, including GACS, AgroPortal, and the Map of Standards, will require one or more step change investments based on a sustainability plan with a well-defined business model. Sustainability is a key issue for any business plan.

A pragmatic implementation proposal

Following the outline vision above, it is important to describe a straightforward and pragmatic route towards a self-sustaining (or largely self-sustaining) operating model for GACS and the wider Agrisemantics landscape. Key to this is building confidence. We suggest using 3 horizons or stages for implementation[23]. Taking GACS as an implementation case, we can recognise the foci required in the different phases.

PoC – proof of concept

If we take the current GACS *Beta* as a starting point, then we can treat this as our proof of concept for the proposed future service. It has clarified what is needed to make a service work technically and it has helped those involved agree on possible scope for a future implementation. A set of future services can be defined around it, and it has pushed the need for sustainability to the fore.

Pilot

Building on this, and using a simple "PoC – Pilot – Production" progression, we propose gaining support and donor funding to set up a working service as a true pilot. This will allow the broader team to gain experience of operational service needs as well as the technical aspects of combining data and ontologies. Some real world use cases for interoperability should be included into the pilot. A robust pilot stage will promote confidence that the service is 'safe' to consume.

Production

Using the knowledge of operational challenges and with feedback from users of the pilot service, a fully-production level service can be specified in conjunction with the appropriate SLAs, support team, and data governance team. Key to this 'production' specification is the

[23] The implementation proposal largely refers to the GACS element of Agrisemantics, but it needs to be kept in mind that the role of GACS is embedded in the overall framework, with the AgroPortal and the Map of Standards as core services.

establishment of a fully self-sustaining business model. This model, along with commitments to long-term availability of the data, information models and ontologies, and services, will be essential in bringing academic and commercial organisations to the platform.

Likely funding needed to implement

Clearly details of full costs will need to be worked out but from the outset it is important to be clear about the desire to make Agrisemantics services such as GACS self-sustaining service offerings. To reach that state from the current position will need funding. Doing this in stages should help to build confidence, minimise risks, and give the chance to bring a wider set of potential funders on board, beyond the obvious larger governmental or charitable donors. By targeting the creation of a service of value to the commercial sector it is possible to bring smaller and larger private organisations in the Ag sector into the discussion early, and convert them into paying customers who can help to support the production service.

Options for self-funding could include:

- Consortium model - Members pay to take part in or influence governance (this could include in kind commitments by people's time). Those sectors who could benefit from the GACS service are unpacked in market analysis in a Business Model Canvas;
- Wikipedia-like crowd funding model;
- Payment for quality certification of semantic assets;
- Support for the creation of, or access to, customized vocabularies and relationships. Clients could also pay to be able to add new items to an existing controlled vocabulary list, concept set, or ontology. We believe that farm management information systems (FMIS) and equipment companies could be willing to pay for such functionality.
- Payment for a high-service/private offering APIs (noting that by definition low service means low SLAs, low guarantee of availability, low bandwidth or number of calls possible in a unit of time; high might be 24/7, high throughput, queries/calls kept private, etc.)
- Training and integration of services on top of the raw concept and API offering – training, KT/upskilling, and integration services for addition of other ontologies/concept sets/data sets, whether external (grow the public set) or internal (private extensions).

The final business model for a sustainable running of common semantic services will most likely draw on more than one of these elements. More generally it will be important to secure a role for GACS and other semantic services for agriculture and nutrition in the EOSC concept that foresees community services for the more efficient 'FAIR'ization of data; there has been some consideration as to how that might best be achieved. If the EU follows the same trajectory as that taken by US NSF, by which it mandates specific expenditure in

trusted services for data management in all EU-funded projects, then this could deliver sustainable income to Agrisemantics.

As of now we estimate that with an investment of €1 million, the three main Agrisemantics services (GACS, Agroportal and Map of Standards) could be brought from PoC phase to Pilot, while resourcing thorough investigation of options for a sustainable business model.

Appendix A: Agrisemantics partners in the GACS working group

- GODAN – Global Open Data for Agriculture and Nutrition
- CABI – Centre for Agriculture and Biosciences International
- FAO – Food and Agriculture Organization of the United Nations
- NAL – USDA National Agricultural Library
- INRA – Institut national de la recherche agronomique
- Syngenta AG
- CGIAR – Consultative Group on International Agricultural Research
- AgroPortal - University of Montpellier (LIRMM)

Appendix B: Links

- Agrisemantics - http://agrisemantics.org/
- GACS - http://agrisemantics.org/gacs/
- GACS Browser - http://browser.agrisemantics.org/gacs/en/
- AgroPortal - http://agroportal.lirmm.fr
- GODAN Map of Standards - http://vest.agrisemantics.org/