**ICT Seventh Framework Programme (ICT FP7)**

**Grant Agreement No: 318497**

**Data Intensive Techniques to Boost the Real – Time Performance of Global Agricultural Data Infrastructures**


# D5.5b Prototype integration with agINFRA

| Deliverable Form | |
|---|---|
| **Project Reference No.** | ICT FP7 318497 |
| **Deliverable No.** | D5.5b |
| **Relevant Workpackage:** | WP5: Semantic Infrastructure |
| **Nature:** | Prototype |
| **Dissemination Level:** | PU = Public |
| **Document version:** | Final |
| **Date:** | 19/02/2016 |
| **Authors:** | AK, UAH, IPB |
| **Document description:** | This version of the report integrates the revisions requested in the final review report and reflects on the updates of AGINFRA as an agri-food research data hub. |

# Document History

| Version | Date | Author (Partner) | Remarks |
|---|---|---|---|
| Draft v0.1 | 01/11/2015 | UAH | Initial version, ToC |
| Draft v0.2 | 07/11/2015 | IPB, AK, UAH | RESTFUL interface section and comments |
| Draft v0.3 | 09/11/2015 | IPB | The e-infrastructure section |
| Final v1.0 | 26/11/2015 | NCSR-D | Internal review and delivery |
| Draft v1.1 | 03/02/2016 | UAH | New section (Section 3.1) on IPython Notebook |
| Draft v1.2 | 08/02/2016 | AK, FAO | New section (Section 3.6) on SemaGrow-powered AGRIS recommendation widget |
| Draft v1.3 | 08/02/2016 | AK | Revisions based on the final review report and updated AGINFRA (New Chapters 3 and 6, new Annex, updates in Sections 1 and 2) |
| Draft v1.4 | 12/02/2016 | NCSR-D | Internal review |
| Final v2.0 | 19/02/2016 | UAH | Delivered as D5.5b |

## EXECUTIVE SUMMARY

The first version of this deliverable presented a documentation of how the SemaGrow stack was integrated in the agINFRA European FP7 project (Grant agreement no: 283770; http://www.aginfra.eu/project). The agINFRA project was an Integrated Infrastructure Initiative (I3) project that worked on introducing the agricultural scientific communities into the vision of open and participatory data-intensive science.

After the end of the agINFRA project in February 2015, the infrastructure and other outcomes of the project were adopted by a consortium of initiatives all over the world, aiming to ensure their sustainability, further adoption by potential stakeholders, as well as the continuation and expansion of the work done so far. The new AGINFRA is an agricultural data e-infrastructure that aims to provide a wide variety of related services to various types of stakeholders in the agrifood sector. In this context, the following apply regarding the integration of SemaGrow components into the AGINFRA infrastructure:

1. The SemaGrow stack becomes the open source software placeholder for all AGINFRA developers;
2. The SemaGrow demo software is also registered in CIARD RING and exposed using the AGINFRA API Gateway;
3. The SemaGrow-powered VocBench is behind the AGINFRA semantic backbone;
4. The SemaGrow-powered Trees4Future/AgMIP data federation is integrated with CIARD RING;
5. The SemaGrow-powered AGRIS recommendation widget is already integrated in the operational service.

In this context, the main objective of this version of the deliverable is to describe the integration of the SemaGrow stack and individual components in the AGINFRA infrastructure so that they become more easily available to the agricultural research communities at a global level, facilitating retrieval and adoption by potential stakeholders and end users. This version of the deliverable also integrates the comments received through the Final Review Report, dated 07/01/2016, such as the elaboration of the section regarding the IPython notebook used as a demonstrator of the integrated SemaGrow/AGINFRA prototype.

# TABLE OF CONTENTS

**LIST OF FIGURES** ..... **5**

**1. INTRODUCTION** ..... **7**

1.1 Purpose and Scope ..... 7

1.2 Approach ..... 7

1.3 Big Data Aspects ..... 7

**2. THE AGINFRA PROJECT AND THE AGINFRA E-INFRASTRUCTURE** ..... **9**

2.1 The agINFRA FP7 project ..... 9

2.2 The AGINFRA global hub of agricultural research ..... 9

2.3 AGINFRA core components ..... 10

2.3.1 CIARD-RING ..... 10

2.3.2 The AGINFRA e-infrastructure ..... 11

**3. SEMAGROW INTEGRATION IN THE AGINFRA E-INFRASTRUCTURE** ..... **13**

3.1 The IPython Notebook ..... 13

3.2 Registry of running software services and components ..... 14

3.3 Repository of open source software ..... 15

3.4 VocBench as the backbone of GACS ..... 15

3.5 The SemaGrow-powered Trees4Future/AgMIP data federation ..... 16

3.6 The SemaGrow-powered AGRIS recommendation widget ..... 17

**4. THE AGINFRA RESTFUL INTERFACE** ..... **19**

4.1 AGINFRA underlying components ..... 19

4.1.1 agHarvester ..... 19

4.1.2 agCrawler ..... 19

4.1.3 agDCtoLOM ..... 19

4.1.4 agLOMtoAKIF ..... 20

4.1.5 agLOMtoRDF ..... 20

4.1.6 agTextMining ..... 20

4.1.7 agTagger ..... 20

**5. INTEGRATION PROCESS FOR THE SEMAGROW STACK** ..... **21**

**6. CONCLUSIONS** ..... **23**

**7. ANNEX: INTEGRATING SEMAGROW IN THE EVOLUTION OF AGINFRA** ..... **25**

# LIST OF FIGURES

Figure 2-1: The AGINFRA global hub of agri-food research ..... 10
Figure 3-1: Part of the AGINFRA components in CIARD RING ..... 13
Figure 3-2: The IPython notebook available through CIARD RING ..... 14
Figure 3-3: The SemaGrow open software stack in AGINFRA ..... 15
Figure 3-4: Using VocBench for the management of GACS concepts ..... 16
Figure 3-5: SemaGrow-enhanced data sources in CIARD RING ..... 17
Figure 3-6: An AGRIS mashup page displaying an AGRIS bibliographic record with related information computed making use of the SemaGrow Stack ..... 18
Figure 5-1: Workflow demonstrating the integration between AGINFRA and SemaGrow ..... 21

# 1. INTRODUCTION

## 1.1 Purpose and Scope

During the lifetime of the agINFRA project, a scientific data infrastructure was designed and developed for the agricultural sciences, facilitating the development of policies and the deployment of services that promote sharing of data among agricultural scientists and develop trust within and among their communities. The agINFRA project outcomes consist (among others) of a catalogue of datasets and services in the field of agriculture as well as a grid infrastructure that provides big computation performance in cases where users cannot execute a service using local computers or when a dataset has such a big size that will also need to take advantages of it. After the end of the project, agINFRA evolved into the AGINFRA global hub for agri-food research, integrating not only the outcomes of the agINFRA project but also additional ones, involving a number of organizations and initiatives at a global level as contributors in this effort.

This deliverable aims to describe the integration of the SemaGrow stack and its components in the case of a real agricultural research data infrastructure like AGINFRA. At the same time, it aims to highlight the role of each component in the AGINFRA infrastructure. The integration of SemaGrow in an operational, running environment showcases the way that the research contributions of SemaGrow can become integral component of real-life solutions that bring together very large and complex data volumes in order to support very diverse needs of many different users. In addition, by integrating SemaGrow in the AGINFRA e-infrastructure it is expected that access to these outcomes will be significantly facilitated and reached by a higher number of stakeholders.

## 1.2 Approach

In order to provide the context of the work described in this document, the deliverable provides a short description on the agINFRA FP7 project and its evolution into the AGINFRA global agricultural data research e-infrastructure. Based on that, the integration of SemaGrow components in the AGINFRA infrastructure is described with references to individual components, the AGINFRA RESTFUL interface, as well as a set of conclusions summarizing the previous points. The document refers to completed and ongoing aspects of the SemaGrow integration in the AGINFRA infrastructure, providing additional technical details where necessary.

## 1.3 Big Data Aspects

The AGINFRA infrastructure consists of a grid-computing environment provided by the Institute of Physics of Belgrade (IPB), Istituto Nazionale de Fisica Nucleare de Catania (INFN) and SZTAKI of Budapest, some of the agINFRA project partners. Its main objective is to provide users with the infrastructure for big data computation. Cases involving big data that could use it will be when a service provided by SemaGrow will consist of tasks that need a big computation (such as the DLO demonstrator presented in the deliverable D6.3) to be executed or big datasets that need a big storage or big computational power to be processed.

# 2. The agINFRA project and the AGINFRA e-infrastructure

In this section, an overview of the context of the work related to the SemaGrow integration in the AGINFRA infrastructure is provided. In this direction, the agINFRA project is presented along with its evolution to the AGINFRA infrastructure and its main components.

## 2.1 The agINFRA FP7 project

The agINFRA FP7 project was an Integrated Infrastructure Initiative (I3) project that introduced the agricultural scientific communities into the vision of open and participatory data-intensive science. In particular, agINFRA aimed at designing and developing a scientific data infrastructure for agricultural sciences to facilitate the development of policies and the deployment of services that promote sharing of data among agricultural scientists and develop trust within and among their communities. agINFRA worked towards eliminating existing obstacles concerning the open access to scientific information and data in agriculture and improved the preparedness of agricultural scientific communities to face, manage and exploit the abundance of relevant data that is (or will be) available and can support agricultural research.

Ultimately, agINFRA demonstrated how a data infrastructure for agricultural scientific communities can be set up to facilitate data generation, provenance, quality assessment, certification, curation, annotation, navigation and management.

During the lifetime of the project, agINFRA worked towards the following aspects:

- ***Shared e-infrastructure, tools and services****:* agINFRA developed and made available the shared e-infrastructure required for agricultural research resources (content/data) and services.
- ***Higher interoperability of data****:* agINFRA promoted a higher level of interoperability between agricultural and other data resources (e.g. through deploying a Linked Agricultural Data Layer).
- ***Improved research data services****:* agINFRA allowed for service improvement so that agricultural researchers can produce and transfer novel scientific and technological results for effective outcomes in the agricultural sector.

## 2.2 The AGINFRA global hub of agricultural research

After the end of the agINFRA project, and with contributions from various key organizations at a global level, agINFRA evolved into a global hub of research & extension information related to agriculture & food security. The AGINFRA global hub consists of the following, among others:

- a global atlas of agricultural research & extension (including institutions, people, publications, data sets, projects, courses, OERs);
- a semantic layer of processing, enriching & interlinking research information from distributed, heterogeneous sources & formats;
- a catalogue of software components (open source software stack & APIs) that anyone may use to process research information;

- a help desk service to support institutions & projects that wish to publish their research information openly;
- a set of data-rich service and application demonstrators for specific case studies, such as food safety, vitis, crop composition etc.

The following figure provides an overview of AGINFRA as a global agri-food research data hub.

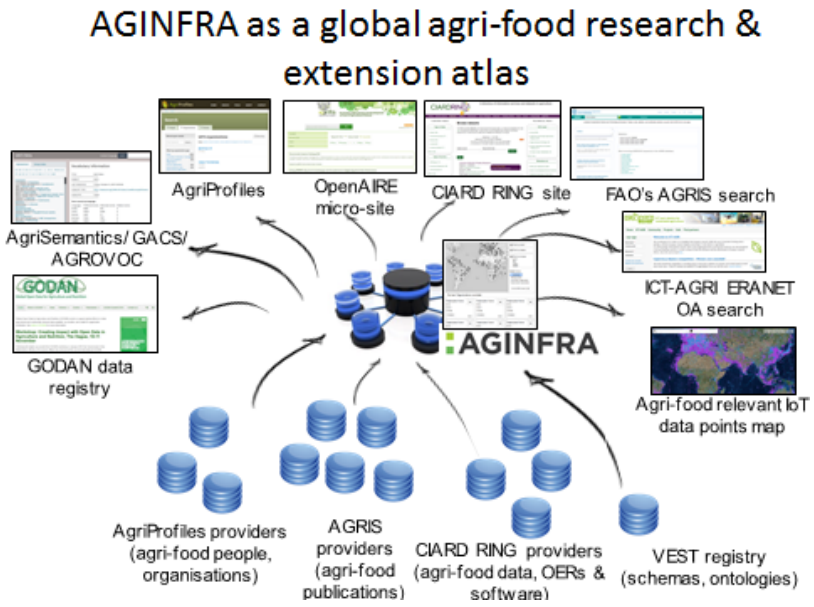


**Figure 2-1:** The AGINFRA global hub of agri-food research

AGINFRA aims to facilitate the work of various stakeholders in the agri-food research context and at a global level by identifying, describing and enhancing the accessibility of different components that are substantial for the agri-food research community.

## 2.3 AGINFRA core components

As explained earlier, the AGINFRA e-infrastructure consists of two main components: the CIARD-RING that acts as the registry of datasets and services and the e-infrastructure needed for serving the agri-food research community. These two components will be presented in details in the following sections, so that the integration of the SemaGrow stack will be put into the corresponding context.

### 2.3.1 CIARD-RING

The CIARD Routemap to Information Nodes and Gateways (CIARD RING; http://ring.ciard.info) is a project implemented within the Coherence in Information for Agricultural Research for Development (CIARD; http://www.ciard.info) initiative and is led by the Global Forum on Agricultural Research (GFAR; http://www.gfar.net).

The RING is a global directory of web-based information services and datasets for agricultural research for development (ARD). It is the principal tool created through the CIARD initiative to allow

information providers to register their services and datasets in various categories and so facilitate the discovery of sources of agriculture-related information across the world.

The RING aims to provide an infrastructure to improve the accessibility of the outputs of agricultural research and of information relevant to ARD management. The functions of the RING consist of:

- providing a map of accessible information sources with instructions on how they can be used effectively;
- providing a dataset sharing platform for agriculture;
- providing examples of services that show good practices on implementing “interoperability”;
- clarifying the level and mode of interoperability of information sources;
- providing instructions for building enhanced integrated services that repackage information in different ways.

In the context of the agINFRA project, the RING developed a machine-readable Linked Data layer to meet the following requirements:

- Datasets registered in the RING have to be found by applications;
- Applications have to be able to read all the metadata about datasets and filter datasets according to their needs
- Applications have to find enough technical metadata in the RING to:
    - Identify datasets with a specific coverage (type of data, thematic coverage, geographic coverage);
    - Identify datasets that comply with certain technical specifications (format, protocol etc.);
    - Access the dataset and get the data;

This machine-readable layer can for instance support the data aggregation workflows of external services.

### 2.3.2 The AGINFRA e-infrastructure

The agINFRA project was conceived as a sustainable data e-Infrastructure for agricultural research. The project addresses implementation of services relevant for a wide range of users: from infrastructure providers, developers of data processing and data management software, data providers, information managers, librarians, developers of high-level portals, to end-users (researchers, educators, citizens). Design of the e-Infrastructure relied on adaptation of the existing infrastructures (Grid, Cloud), their customization, and the development of new components that will ensure balance in the system and its seamless use.

At the lowest level, available hardware resources are organized into Grid sites and Cloud sites. Grid technology is dedicated to off-line processing that demands quite intensive computational power, as well as for metadata and data services able to locate and provide huge quantities of information. On the other hand, Cloud computing addresses requirements of interactive/on-line use and providing of permanent services. This layer is very extensible, and allows easy integration of new hardware

resources into the existing system, which is crucial for the infrastructure providers. On top of the Grid and Cloud infrastructure, the layer of Science Gateways (SG) provides seamless access to the distributed infrastructures for developers of processing and data management software. By definition, SG is a community-developed set of tools, applications, and data that is integrated via a portal or a suite of applications, usually through a graphical user interface, which is further customized to meet the needs of a specific community. In parallel to the two general-purpose SG implementations, an agINFRA RESTful interface is provided as well. This interface is developed within the project, and is used for cataloging, off-line processing, and management of data.

The agINFRA project Grid e-Infrastructure consisted of 4 Grid sites: AEGIS01-IPB-SCL, INFN-CATANIA, INIF-ROMA3, and SZTAKI. The infrastructure is organized around Virtual Organisations (VO). In total, 2000 CPU-cores and 900 TB of storage space are provided. In addition to this, a set of core Grid services is deployed to support the VO. The VO is operated on the EGI infrastructure, and it is supported by all Grid sites that participated in the project. The agNFRA Grid services are based on the latest version of the grid middleware software provided by EMI. Authentication and authorization of vo.aginfra.eu VO members are done by the Virtual Organization Membership Service (VOMS) instances. Job Management services are provided through a number of Cream and WMS & LB instances that are responsible for execution and control of computational jobs on the Grid infrastructure. A MyProxy (PX) service instance provides a proxy renewal mechanism that keeps proxies for the submitted jobs valid through their whole lifetime on the infrastructure. Data services are accessible through instances of Storage Elements (SE) that provide uniform access to data storage resources. Logical File Catalogue (LFC) service instance keeps track of the location of the files belonging to VO and replicas distributed in the Grid. Metadata services are provided by the AMGA metadata catalogue.

OCCI-based and Okeanos Cloud resources are provided within the agINFRA. From the hardware point of view, a server with 32 CPU cores and 128 GB of RAM, as well as 4 disks with 2 TB capacity each have been deployed to support OCCI-based infrastructure. This server has been commissioned with the latest version of the OpenStack cloud. Okeanos is an "Infrastructure as a Service" where the user can easily build her own Virtual Machines and Virtual Networks. In addition, the user can manage them, destroy them, connect to them and take a handful of other actions, all from inside a web browser. Also, the user can store files online, share them with other users and access them anytime, from anywhere in the world and access them from inside a Virtual Machine (VM).

# 3. SemaGrow integration in the AGINFRA e-infrastructure

The following sections provide information on the actual implementation of the SemaGrow components in the AGINFRA infrastructure, and more specifically on the SemaGrow components that have been integrated as well as the role of each one in the AGINFRA infrastructure. The following figure provides an overview of the various AGINFRA components, including the SemaGrow ones.


**Figure 3-1:** Part of the AGINFRA components in CIARD RING

(Source: http://ring.ciard.info/views/aginfra)

SemaGrow components, including the IPython Notebook and the SemaGrow stack, are listed at the AGINFRA software list.

## 3.1 The IPython Notebook

IPython (http://ipython.org) is a web-based interactive computational environment where one can combine code execution, text, mathematics, plots, rich media, as well as gLite user interface APIs. In the context of the SemaGrow project, an IPython notebook was used as the backbone for the

integration of the SemaGrow stack components, as well as AGINFRA RESTful API calls into a single document.

The notebook consists of 4 steps:

1) Query CIARD RING to retrieve the datasets that the user is interested in.
2) Once the user has chosen a dataset to work with, query again CIARD RING in order to know which services can work with the specific dataset.
3) Execute the service using the dataset to obtain results. If the service needs of a big computational power that cannot be provided by a regular machine (such as the one used by a typical user), this will be executed over the e-infrastructure provided by AGINFRA.
4) From the e-infrastructure and using the different methods provided by it to execute the service, the user is able to obtain a number of results.

This workflow was developed as a real use case of SemaGrow. The one chosen was the case provided by DLO, as a project partner. In this context, another notebook that would work with data provided by Trees4Future-AgMIP was developed. This service provides a file with information on the yield wheat and the average temperatures of a particular geographic zone for a range of years. In the specific demonstration, information for the top wheat producers of the last 10 years has been obtained. Once these files were obtained, data was manipulated with the use of IPython libraries in order to compare the production of five countries during these years. The results are shown in a motion chart, which is a dynamic chart to explore several indicators over time.

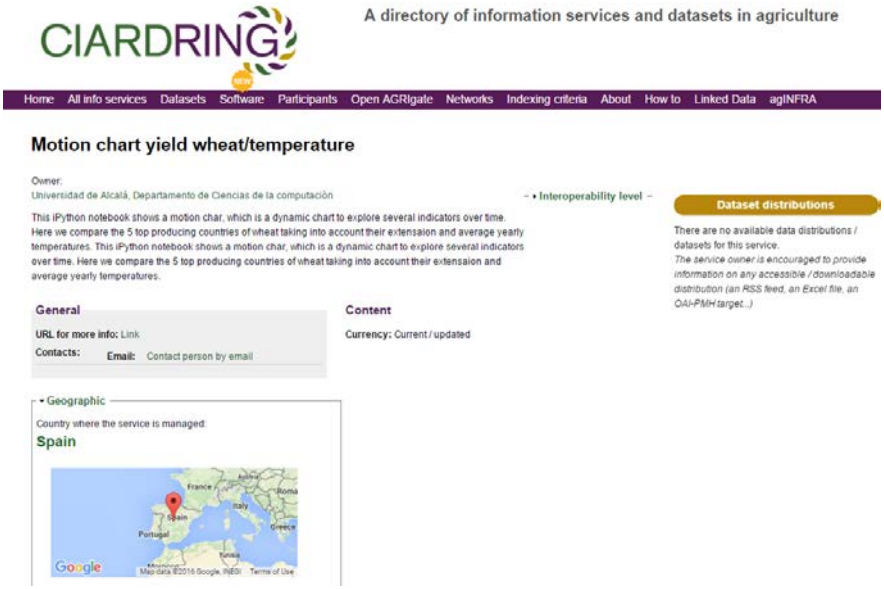


**Figure 3-2:** The IPython notebook available through CIARD RING

## 3.2 Registry of running software services and components

One of the core components of the AGINFRA e-infrastructure is the registry of services. In this context, all existing running software services and components, as well as the SemaGrow Demo

software have been registered in CIARD RING and exposed using the AGINFRA API Gateway. In this way, the SemaGrow running services can be accessed through CIARD RING and executed online. The aforementioned example of the SemaGrow IPython notebook is an example of such as running service that is currently available through AGINFRA's CIARD RING.

## 3.3 Repository of open source software

The AGINFRA approach is the development of an AGINFRA stack of technologies that will be based on the software previously developed in the context of the agINFRA FP7 project. This stack will be enriched with the existing SemaGrow stack as well as any future open source software technologies, such as the ones to be developed in the context of the BigDataEurope Horizon 2020 project (http://www.big-data-europe.eu).

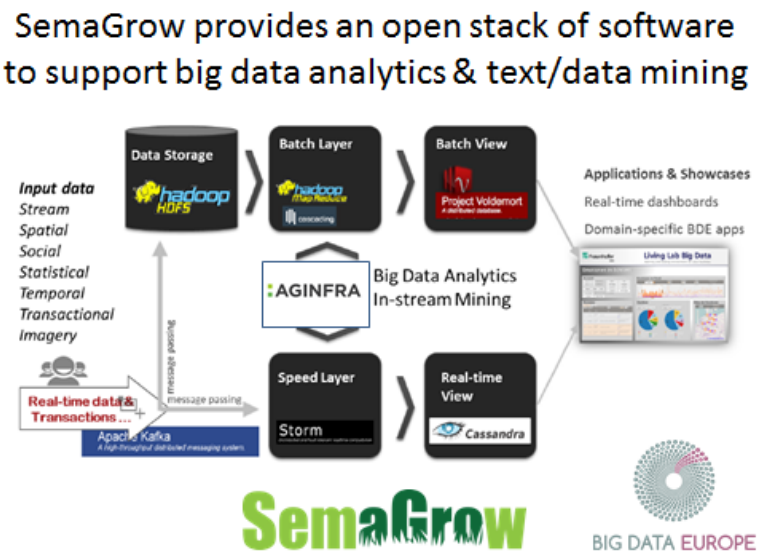


**Figure 3-3:** The SemaGrow open software stack in AGINFRA

The SemaGrow stack will become the open source software placeholder for all existing and new AGINFRA developers. All existing efforts will be taken into consideration, such as the existing GitHub repository of SemaGrow (https://github.com/semagrow / http://semagrow.github.io) and the existing but still disperse AGINFRA Github repositories hosting parts of the AGINFRA software stack. The aim of this work is to provide a centralized mean for accessing and retrieval of the SemaGrow stack components to be available through the AGINFRA hub, enhancing their availability and visibility to potential users of these components.

## 3.4 VocBench as the backbone of GACS

The VocBench evolution developed in the context of the SemaGrow project (including the ontology alignment component) is used by the Global Agricultural Concept Scheme (GACS) initiative that is working on the deployment of an aligned semantic backbone of AGINFRA. GACS is a joint effort of three major publishers and managers of agricultural research outcomes, namely the Food and

Agriculture Organization of the UN, CAB International and the National Agricultural Library of the U.S. Department of Agriculture.

This effort is aimed at improving the semantic interoperability between the thesauri maintained by these organizations, namely FAO AGROVOC (32K concepts), CABI Thesaurus (140K concepts) and NAL Thesaurus (53K concepts) by defining the core concepts broadly supported across them. The GACS initiative uses the SemaGrow-enhanced VocBench platform for the maintenance of the concept schemes and the definition of its structure.

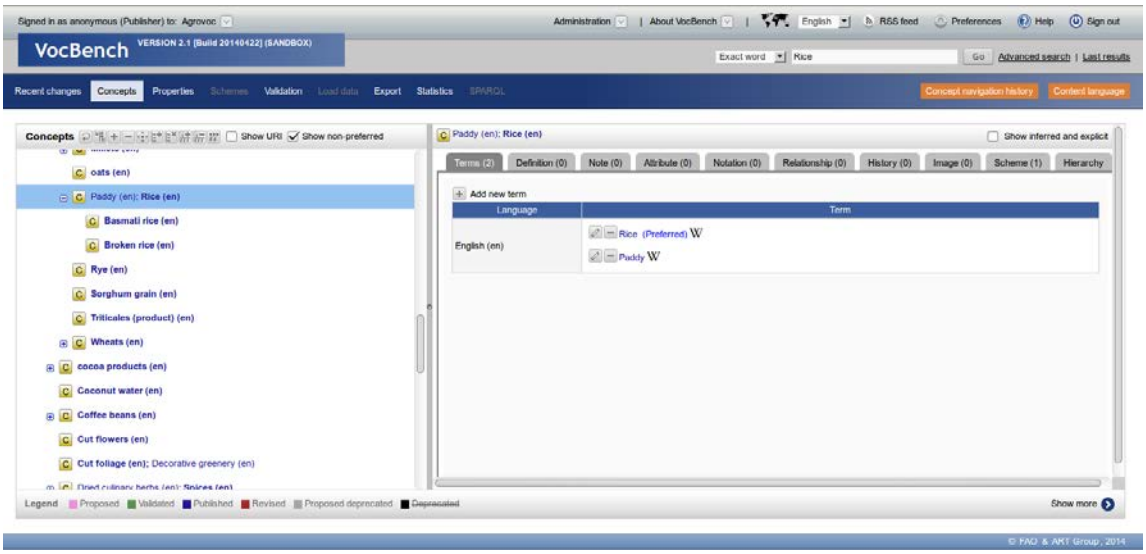


**Figure 3-4:** Using VocBench for the management of GACS concepts

## 3.5 The SemaGrow-powered Trees4Future/AgMIP data federation

The Trees4Future/AgMIP datasets, used as datasets in the context of SemaGrow's pilot trials have been included in the CIARD RING registry as individual datasets[1] by SemaGrow partners (WUR/DLO). It should be noted that these data sources are the triplified NetCDF data exposed as SPARQL endpoints and not the raw NetCDF data. This allows the specific datasets to be available not only for retrieval, but also for querying. The endpoints can be reached through CIARD RING, a core component of the AGINFRA infrastructure.

[1] http://ring.ciard.info/services-datasets?search_api_views_fulltext=semagrow&search_api_views_fulltext_1=

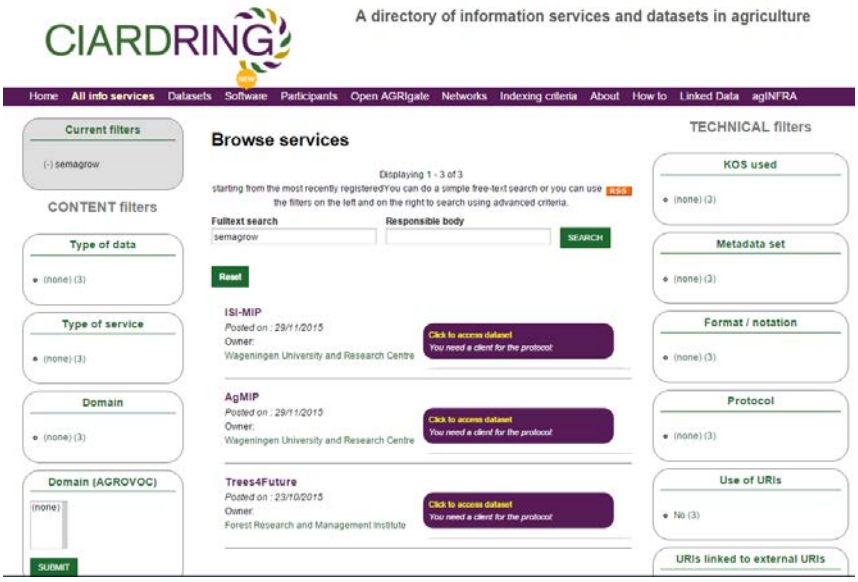


**Figure 3-5:** SemaGrow-enhanced data sources in CIARD RING

## 3.6 The SemaGrow-powered AGRIS recommendation widget

The bibliographic discovery service of FAO AGRIS (http://agris.fao.org) is powered by a SemaGrow component that intersects the AGRIS database (a core component of the AGINFRA infrastructure) with the Crawler database in order to provide recommendations of relevant resources to the users of the AGRIS service. This recommender system[2] can intersect different datasets using AGROVOC as its backbone. In SemaGrow, the AGRIS core dataset was intersected with the crawler database. This crawler database was built using a web crawler to discover resources, and the AgroTagger was used to assign AGROVOC URIs to resources discovered by the web crawler.

More information about the SemaGrow-powered recommender system can be found in the deliverable D6.2.2 “Pilot Deployment”.

[2] https://github.com/fcproj/recommender

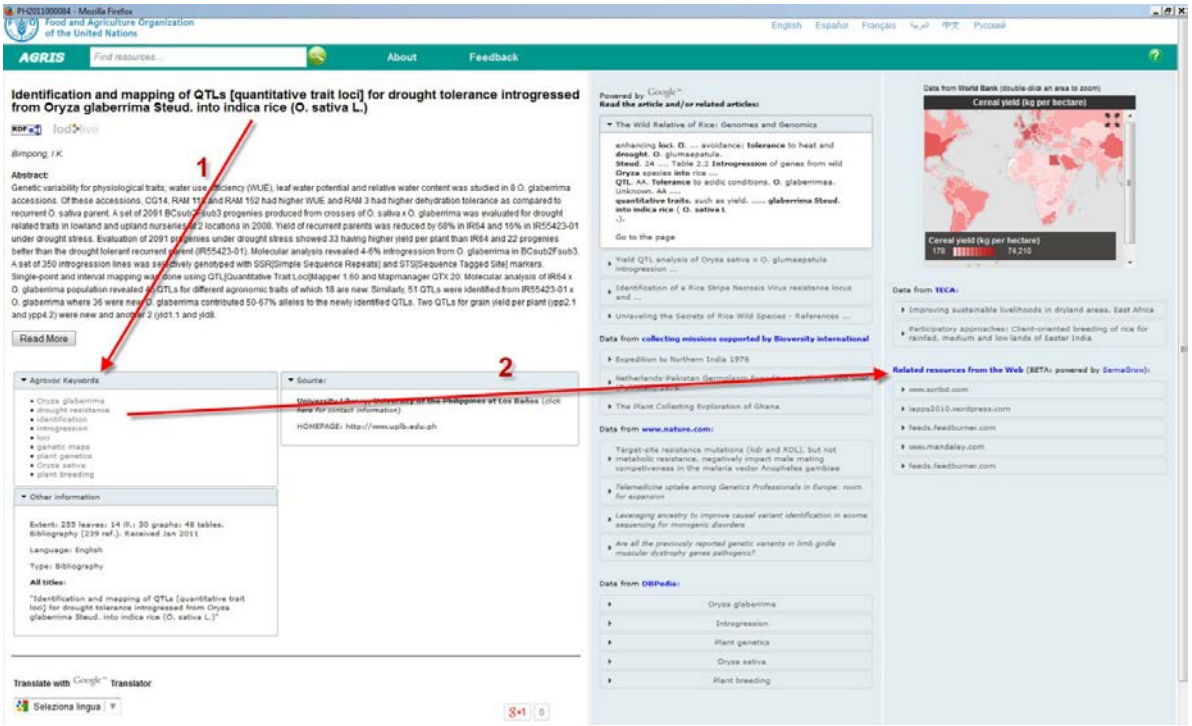


**Figure 3-6:** An AGRIS mashup page displaying an AGRIS bibliographic record with related information computed making use of the SemaGrow Stack

# 4. The AGINFRA RESTful interface

The AGINFRA RESTful interface was developed in the context of the agINFRA project and is used for cataloguing, off-line processing, and management of data. As a catalogue facility, the AGINFRA RESTful interface keeps user's configurations of Grid-ported applications, datasets' metadata information and locations, arbitrary user-defined metadata information, and allows querying of this information. Based on the information stored in the catalogue, the gateway performs off-line Grid job submissions in order to collect new datasets, transform existing datasets, and produce new datasets' metadata information relevant for end-users. Due to heterogeneity of the project infrastructure and ported applications, registered locations of datasets initially point to the different storage architectures or systems, and retrieval of datasets is limited to the storage-specific protocols and authentications mechanisms. Unification of these protocols is achieved by the AGINFRA RESTful interface, and end-users are enabled to retrieve the datasets produced within the infrastructure using a unique protocol (HTTP protocol). This interface carries out automatic replication of datasets, ensures the existence of the same dataset on different storage systems, and its exposure through the HTTP protocol.

## 4.1 AGINFRA underlying components

AGINFRA has a set of underlying components developed for different aims. A component developed in the context of the agINFRA project presents an application ported to the Grid infrastructure, and a RESTful interface on top of the application that enables configuration of application input parameters and output retrieval. The components deployed during the project are: agHarvester, agCrawler, agDCtoLOM, agLOMtoAKIF, agLOMtoRDF, agTextMining and agTagger. In the following subsections we will explained them in more details.

### 4.1.1 agHarvester

agHarvester performs harvesting of any dataset exposed via an OAI-PMH target. The module is agnostic about the type of metadata to be harvested (DC, LOM) and can support harvesting of any metadata format as this is declared in the "metadaPrefix" field of the verb "ListMetadataFormats" of an OAI-PMH target, e.g. http://aglr.agroknow.gr/organic-edunet/oai?verb=ListMetadataFormats.

### 4.1.2 agCrawler

agCrawler is a customized version of Apache Nutch (http://nutch.apache.org), a highly extensible and scalable open source Web crawler. Its main goal is to discover resources on the Web (i.e. URLs), starting from some Web sites defined by the user.

### 4.1.3 agDCtoLOM

agDCtoLOM process performs conversion of the Dublin Core metadata schema (DC) into in IEEE LOM metadata schema. This process is part of the data transformation layer. The transformation could be executed taking as input a technical binding (e.g. XSLT) of the corresponding mappings.

### 4.1.4 agLOMtoAKIF

agLOMtoAKIF performs a conversion of a set of metadata records with XML binding that follow IEEE LOM metadata format into the AKIF format. The conversion is XSL(T) based and rules are defined in single configuration file.

### 4.1.5 agLOMtoRDF

agLOMtoRDF performs conversion of a set of metadata records with XML binding that follow the IEEE LOM metadata format into RDF/XML binding.

### 4.1.6 agTextMining

agTextMining returns for a given datasets titles, authors, references and keywords. Currently, version 1.0 works with IEEE LOM records serialized as XML files. The keyword extractor uses the KEA algorithm and statistical model to calculate keywords from the text. The title and author fields are parsed detecting sudden font size changes. Finally, the references are obtained parsing numbers between brackets.

### 4.1.7 agTagger

agTagger is a keyword extractor that uses the AGROVOC thesaurus to extract keywords from the content of some URLs. Since AGROVOC is published as Linked Open Data, the agTagger can do more than extracting keywords, it can extract AGROVOC URIs. The agTagger is based on MAUI (https://code.google.com/archive/p/maui-indexer), a piece of software that automatically identifies main topics in text documents, using two different algorithms: the key-phrase extraction algorithm KEA (http://www.nzdl.org/Kea), and the machine learning toolkit WEKA (http://www.cs.waikato.ac.nz/ml/weka). To be used in the agTagger, MAUI was trained to work with AGROVOC (in English).

# 5. Integration process for the SemaGrow stack

The integration of SemaGrow with agINFRA is defined in the SemaGrow DoW as "…deal with the integration of the SemaGrow components with a real agricultural data infrastructure, the agINFRA one".

In order to facilitate the integration, this task has been divided in various phases:

- First all the services and datasets provided by the SemaGrow stack need to be registered in CIARD RING. The SemaGrow stack can also be considered as a whole service.
- Once the services and the datasets have been registered, the e-infrastructure provided by IPB can be used, so we can the work can benefit from its big computational capacity.

A workflow demonstrating the integration of SemaGrow with AGINFRA could be described as follows: Let's imagine that a user needs to use an agriculture service. The user will usually use Google for identifying and accessing these services. As CIARD RING is a portal providing agricultural services and datasets, he will access it. CIARD RING has a search engine; he will find the SemaGrow services and datasets. Once he chooses the SemaGrow service that he is interested working with, he can execute it. Finally, if this service requires big computing power, the service will be executed from the e-infrastructure.

This workflow highlights the integration of SemaGrow with AGINFRA. Taking into account that the services and datasets of SemaGrow are registered in CIARD RING. First all the datasets of SemaGrow are queried from CIARD RING, a core component of the AGINFRA e-infrastructure. Then the user benefits from CIARD RING as a catalog of services related with datasets. Last but not least, a SemaGrow service is executed in the e-infrastructure (one of AGINFRA's core components, too) that will be executed against the SemaGrow stack where the user will retrieve the information.

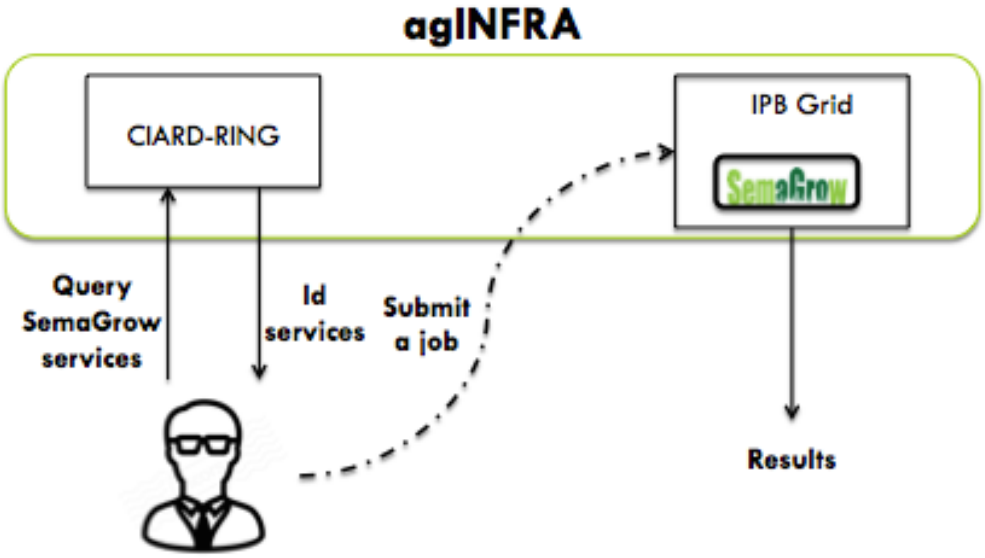


**Figure 5-1:** Workflow demonstrating the integration between AGINFRA and SemaGrow

Making the previously mentioned generic workflow more specific: The user accesses CIARD RING, searching for a service that works with temperatures and coordinates. CIARD RING will respond with a service called Trees4Future-AgMIP. This service has been developed by DLO and retrieves netCDF data regarding wheat yield and temperatures giving a set of coordinates. This service executes the SemaGrow Stack installation that federates the multiple netCDF data endpoints. As these need large-scale storage and computation, they execute on the e-infrastructure. Figure 5-1 shows the workflow described in the example.

# 6. Conclusions

This deliverable presents a documentation of the activity aimed at integrating SemaGrow with the AGINFRA e-infrastructure. It covers a variety of different aspects, such as the agINFRA FP7 project and its outcomes, the evolution of agINFRA into a global agri-food research data e-infrastructure and the actual integration of the SemaGrow stack and its components with the existing AGINFRA e-infrastructure. At the same time, the deliverable provides information on the use of the SemaGrow stack components and their role in their role in the AGINFRA e-infrastructure. Last but not least, the deliverable provides a workflow showing the integration of SemaGrow with AGINFRA in practice, as well as specific applications of the SemaGrow outcomes through the AGINFRA e-infrastructure.

AGINFRA's CIARD RING plays a central role in the process, acting as the registry where the SemaGrow outcomes have been registered so that they can be easily discovered and used by any stakeholder. This activity refers to both services and datasets developed by the SemaGrow project. At the same time, there is provision for the open source software components of SemaGrow that will play a core role in the development of an open source software repository, where the source code of these pieces of software will be stored, indexed and shared through the AGINFRA infrastructure.

Taking into account that probably the main components of the AGINFRA e-infrastructure are a catalogue of datasets and services in the field of agriculture (CIARD RING) and a shared infrastructure to solve computational problems, the integration needs to accomplish both. For SameGrow to take advantage of the first one, all the services and datasets developed during the SemaGrow project are registered at CIARD RING, including services that do not execute at the infrastructure (such as the AGRIS recommender, cf. Section 3.6). Furthermore, services that expose large-scale netCDF datasets have been integrated in the infrastructure as SPARQL endpoints and a SemaGrow federation over them, all executing in the infrastructure (cf. Section 3.5).

The next steps will further exploit the data integration facilities of the SemaGrow Stack within AGINFRA. AGINFRA is constantly enriched with new services, datasets and functionalities, supported by ongoing projects and global initiatives; in this context, it is expected that all components that consist part of its integrated outcomes, including the SemaGrow stack, will be further used and adapted in order to meet new requirements and support the global agri-food research community, which is the main stakeholder group of AGINFRA.

# 7. Annex: Integrating SemaGrow in the evolution of AGINFRA

Available as an individual file