D2.3 - DATA LINKING TECHNOLOGIES

Co-funded by the Horizon 2020
Framework Programme of the European Union

DELIVERABLE NUMBER	D2.3
DELIVERABLE TITLE	Data Linking Technologies
RESPONSIBLE AUTHOR	Panagis Katsivelis & Giannis Stoitsis (Agroknow)

GRANT AGREEMENT N.	731001
PROJECT ACRONYM	AGINFRA PLUS
PROJECT FULL NAME	Accelerating user-driven e-infrastructure innovation in Food & Agriculture
STARTING DATE (DUR.)	01/01/2017 (36 months)
ENDING DATE	31/12/2019
PROJECT WEBSITE	http://www.plus.aginfra.eu
COORDINATOR	Nikos Manouselis
ADDRESS	110 Pentelis Str., Marousi GR15126, Greece
REPLY TO	nikosm@agroknow.com
PHONE	+30 210 6897 905
EU PROJECT OFFICER	Mr. Christophe Doin

WORKPACKAGE N. | TITLE	WP2 | Data & Semantics Layer
WORKPACKAGE LEADER	Agroknow
DELIVERABLE N. | TITLE	D2.3 | Data Linking Technologies
RESPONSIBLE AUTHOR	Panagis Katsivelis (Agroknow)
REPLY TO	katsivelis.panagis@agroknow.com
DOCUMENT URL	http://www.plus.aginfra.eu/sites/plus_deliverables/D2.3.pdf
DATE OF DELIVERY (CONTRACTUAL)	30 September 2017 (M9), 30 September 2018 (M21) & 30 June 2019 (M30)
DATE OF DELIVERY (SUBMITTED)	29 September 2017 (M9), 29 November 2019 (M35, updated version) 29 January 2020 (M37, updated version)
VERSION | STATUS	3.0 | Final
NATURE	De (Demonstration)
DISSEMINATION LEVEL	PU (Public)
AUTHORS (PARTNER)	Panagis Katsivelis & Giannis Stoitsis (Agroknow)
CONTRIBUTORS	Nikos Manouselis (Agroknow), Timotheos Lanitis (Agroknow), Michail Papakonstantinou (Agroknow)
REVIEWER	Teodor Georgiev (PENSOFT)

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
0.1	Preliminary Tools and Methods review	31/05/2017	Agroknow
0.3	Harmonization with Requirements	31/07/2017	Agroknow
0.5	Silk framework assessment	07/09/2017	Agroknow
0.6	Report setup	15/09/2017	Agroknow
0.7	Report draft finalization	22/09/2017	Agroknow
0.8	Deliverable Review	27/09/2017	PENSOFT
0.9	Deliverable finalization	29/09/2017	Agroknow
1.0	Submission to the EC	30/09/2017	Agroknow
1.2	Data Linking Services kick-start	30/02/2018	Agroknow
1.5	Data Linking Services finalization	30/08/2019	Agroknow
1.6	Data Harvesting finalization	30/10/2019	Agroknow
1.9	Deliverable Review	30/11/2019	PENSOFT
2.0	Submission to the EC	29/11/2019	Agroknow
2.5	Additional registries harvesting	30/12/2019	Agroknow
2.7	Development of Data Dashboard	10/12/2019	Agroknow
2.8	Deliverable Review	13/01/2019	BfR
2.9	Deliverable Review	28/01/2019	PENSOFT
3.0	Submission to the EC	31/12/2019	Agroknow

PARTICIPANTS		CONTACT
Agroknow IKE (Agroknow, Greece)		Nikos Manouselis Email: nikosm@agroknow.com
Stichting Wageningen Research (DLO, The Netherlands)		Rob Lokers Email: rob.lokers@wur.nl
Institut National de la Recherché Agronomique (INRA, France)		Pascal Neveu Email: pascal.neveu@inra.fr
Bundesinstitut für Risikobewertung (BFR, Germany)		Matthias Filter Email: matthias.filter@bfr.bund.de
Consiglio Nazionale Delle Richerche (CNR, Italy)		Leonardo Candela Email: leonardo.candela@isti.cnr.it
University of Athens (UoA, Greece)		George Kakaletris Email: gkakas@di.uoa.gr
Stichting EGI (EGI.eu, The Netherlands)		Tiziana Ferrari Email: tiziana.ferrari@egi.eu
Pensoft Publishers Ltd (PENSOFT, Bulgaria)		Lyubomir Penev Email: penev@pensoft.net

ACRONYMS LIST

RDF	Resource Description Framework
SKOS	Simple Knowledge Organisation System
VRE	Virtual Research Environment
REST	Representational state transfer
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
FRBR	Functional Requirement for Bibliographic Records
API	Application Programming Interface
NLP	Natural Language Processing

EXECUTIVE SUMMARY

The present report is the last submitted iteration of a living document that describes progress and evolvement of the AGINFRA PLUS data linking technologies, i.e. the services and software components that were incorporated in the overall AGINFRA PLUS e-infrastructure and were responsible for providing data and relevant links to project use-cases.

As the project activities progressed, the efforts pertaining to data linking activities were generalized to define a set of services designed to bring domain-relevant data under a common repository: the AGINFRA Data Registry. The Registry was aligned with the AGINFRA Registry of Semantic Resources to enable better annotation of data assets, but also increased discoverability via semantic expansion of user queries.

The final version of the deliverable focuses on the description of the core services enabling the AGINFRA Data Registry and its surrounding family of services, namely the Data Harvesting and Data Linking services.

TABLE OF CONTENTS

TABLE OF FIGURES

1 INTRODUCTION

In general terms, data linking is the task of determining whether two object descriptions can be linked one to the other to represent the fact that they refer to the same real-world object in a given domain or the fact that some kind of relation exists between them. Quite often, this task is performed on the basis of the evaluation of the degree of similarity among different data instances describing real-world objects across heterogeneous data sources, under the assumption that the higher the similarity is between two data descriptions, the higher is the probability that the two descriptions actually belong to the same domain, refer to certain scientific workflows or processes and generally, can be used to reproduce the same research activities.

During a typical scientific lifecycle, data is gathered, processed, compared and published, but oftentimes the products of each stage cannot be directly shared to external agents for interpretation and reuse. This issue hinders communication of research tools and outcomes across scientific communities.

In the context of the Semantic Web, data linking is materialized via the Linked Data Initiative[1], which calls for datasets to provide links to other published resources, thus building the continuously expanding Linked Data Cloud[2]. However, due to lack of traction of several research communities with this vision of Linked Data, a hybrid approach has been devised, so that the engaged communities can also realize the value of de-facto community standards, common metadata schemas across data types and data discovery optimization, powered by semantic resources.

In the following subsections of the report, we describe the linking approach followed in the AGINFRA PLUS project, document the developments carried out to materialize it and present a number of services that have been powered and extended by the AGINFRA PLUS project activities. The adaption of services gave birth to the AGINFRA Data Registry, an API-accessible repository of millions of agri-food research data assets.

[1] http://linkeddata.org/

[2] http://lod-cloud.net/

2 DATA LINKING APPROACH

2.1 DATA TYPOLOGIES AND EXTENSIONS

To understand the data linking requirements of AGINFRA PLUS, we must consider the range and nature of data assets that need to be linked. Based on the requirements analysis reported in deliverable D2.1, three major types of research data were identified:

- Publications;
- Models (generalized to Models & Algorithms);
- Datasets.

The D4Science infrastructure (through which data was hosted and managed in the scope of project activities) allowed for the engineering of new typologies, as an extension to the above three, which led to instances of over-customized metadata profiles reaching the front-end of the virtual research environments. To serve the needs for customized scientific content publishing, two additional generic types of data were added to the bundle:

- Research Objects;
- Semantic Resources;
- Software & Services;
- Projects & Initiatives;
- Organizations;
- Data Sources.

Research Objects are files or metadata records or files that result from the execution of a scientific workflow and that can be used to denote the provenance or additional information about a particular research activity.

Semantic Resources are the individual objects drawn from ontologies and vocabularies (reported in deliverable D2.2), that are relevant to specific scientific domains. For instance, a semantic resource that is relevant to the food safety domain could be a vocabulary term that represents a food hazard and its hierarchical placement in a specific family of hazards. Another instance could be an ontology class describing a particular type of model, that is an extension of the generic class for risk assessment models.

Software & Services are software components, libraries and APIs developed to support scientific activities.

Projects & Initiatives include individual or collaborative enterprises that are planned to achieve a particular aim in the agri-food research sector.

Organizations refer to all organizations (e.g. research, ministries) in the agri-food domain that produce and manage domain-specific data.

Data Sources are all forms of data points that directly or indirectly provide one or several types of data services.

2.2 DATA LINKING AS PART OF A GENERIC SCIENTIFIC WORKFLOW

In the face of the scientific workflows enabled by project activities, the above content specification proved that there was significant heterogeneity of data assets across the different use-cases, as they did not present strong thematic overlaps and therefore, no notable links to one another. End-to-end linked data-powered workflows over federated data sources has not been proven as key research requirement, at least not as previous domain-specific projects had suggested in the past (eg. FP7 SemaGrow[3]). Instead, the majority of the work was centered around *formalizing de facto community standards* (as per RDA

[3] http://semagrow.eu

Agrisemantics recommendations[4]), so that each of the three project use-cases could align with or even impose new knowledge standards to its target research communities.

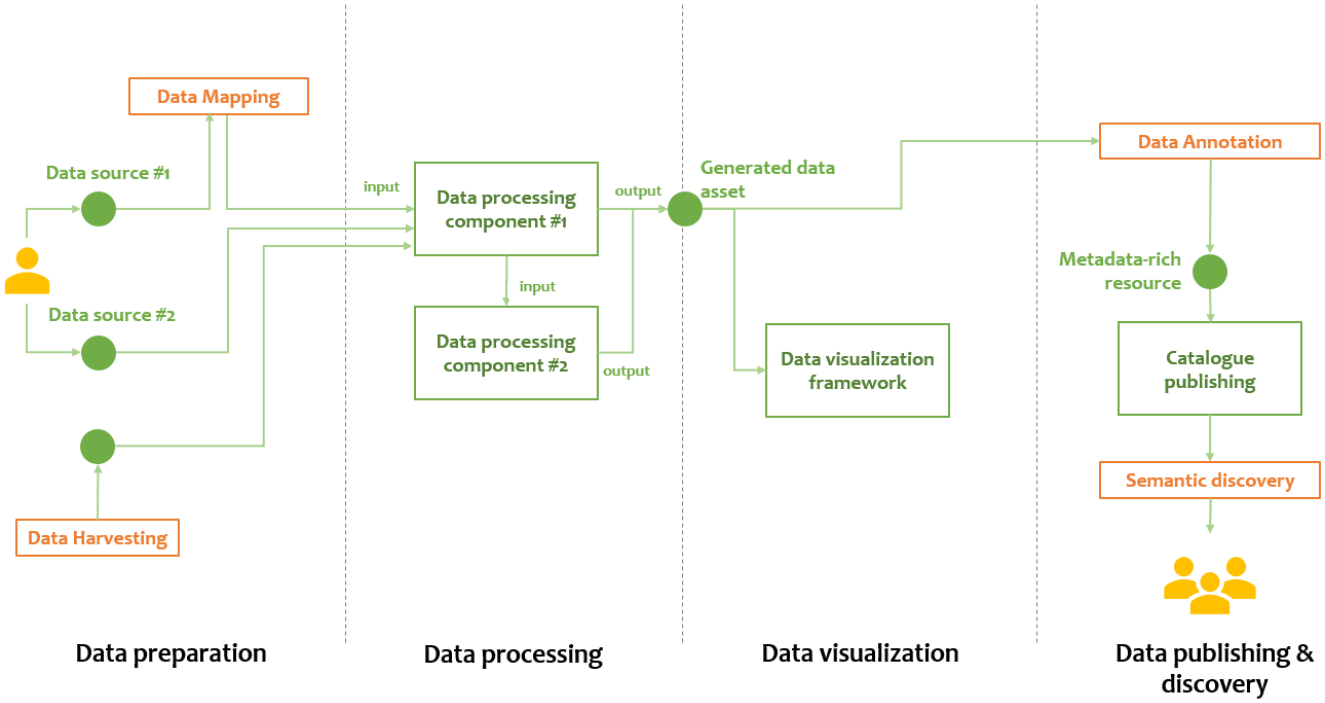

Figure 1: Data Linking services as part of a generic scientific workflow

The followed data linking approach mostly appeared in two distinct phases of each scientific workflow: the beginning (*Data Preparation phase*) and the end (*Data Publishing and Discovery phase*). This approach kept the intermediate phases unburdened by the adoption of semantic tools, which can be difficult to use, unappealing to non-experts and time-consuming when it comes to the actual scientific experimentation phase. Subsequently, attention was shifted towards practical tasks and smart integrations that can be provisioned through the application of four consecutive services (as showcased in Figure 1):

1. Data Harvesting;
2. Data Annotation;
3. Data Mapping;
4. Semantic Discovery.

The proposed services can be used to complete the missing pieces of a complete data lifecycle, from the point where data can be found in raw formats, up to when it is transformed into a reusable metadata-rich resource, available to external users and other research stakeholders.

In terms of infrastructure, data services were employed to complement the Ontological Engineering Layer of AGINFRA PLUS (documented in D2.2) in the form of external web services (as demonstrated in Figure 2). With these in place, any platform can interact with the wealth of

[4] https://www.rd-alliance.org/group/agrisemantics-wg/outcomes/39-hints-facilitate-use-semantics-data-agriculture-and-nutrition

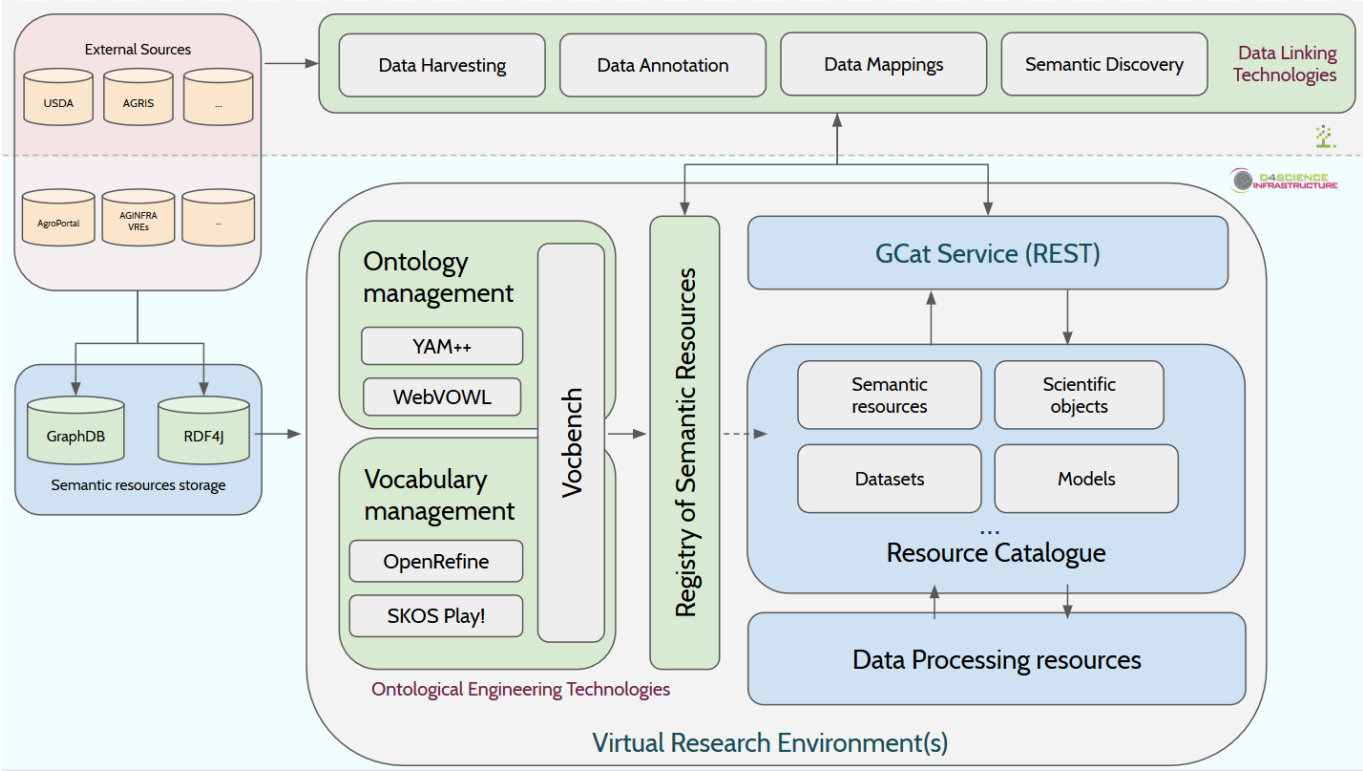

Figure 2: Data Linking services in the context of AGINFRA PLUS e-infrastructure

2.3 ADDIOTIONAL DEVELOPMENTS AND THE MAP OF THE DATA ECOSYSTEM

As part of WP2 activities, additional developments on the existing AGINFRA family of web sites[5] took place, extending their functionality to make them interoperable with the proposed Data Harvesting and Data Linking services. More specifically, the online Map of the Data Ecosystem[6] was extended to include a new web service[7] that exposes the gathered metadata records in the scope of e-ROSA activities in a machine-readable format. The metadata records refer to Organizations, Projects & Initiatives, Data Sources and Software & Services that can be found within the agri-food landscape, which were later on harvested as part of AGINFRA+ project activities.

A more detailed web service exposing Data Sources has also been developed[8], to provide a more comprehensive view on the metadata gathered as part of the Map of the Data Ecosystem.

[5] http://www.aginfra.eu/

[6] http://map.aginfra.eu/

[7] http://map.aginfra.eu/search-api-rest

[8] http://map.aginfra.eu/api/sources

3 DATA HARVESTING & LINKING SERVICES

3.1 DATA HARVESTING

In the scope of AGINFRA PLUS activities, data harvesting services were introduced in the initial requirements analysis (D2.1) as a function that would execute the ingestion of community-relevant content from external repositories and systems. As project activities progressed, it became apparent that the engaged communities were already using custom solutions and services that allowed them to infuse data into their scientific workflows. However, no uniform solution was fostered to encourage the generalization of data ingestion routines so that they can be adapted in any scientific context. At the same time, the initial data types requirements were not entirely met by the custom solutions that were put in-place, hence discouraging the completeness of the scientific value proposed by the e-infrastructure.

3.1.1 The AGINFRA Data Harvesting Workflow and the AGINFRA Data Registry

The AGINFRA Data Harvesting Workflow proposed is an extension of the initial agINFRA project "agHarvester" module[9] approach, which was intended solely for metadata harvesting of OAI-PMH targets. The new data harvesting paradigm is generalized for any possible typology of data assets that can be identified by any scientific community and be brought upon request to a desired, machine-readable format, ready to be infused into any system and context. All technical details are enumerated in D4.2.

The general functions of the workflow that can achieve this are (as depicted in Figure 3):

1. Collect;
2. Transform;
3. Enrich;
4. Curate.

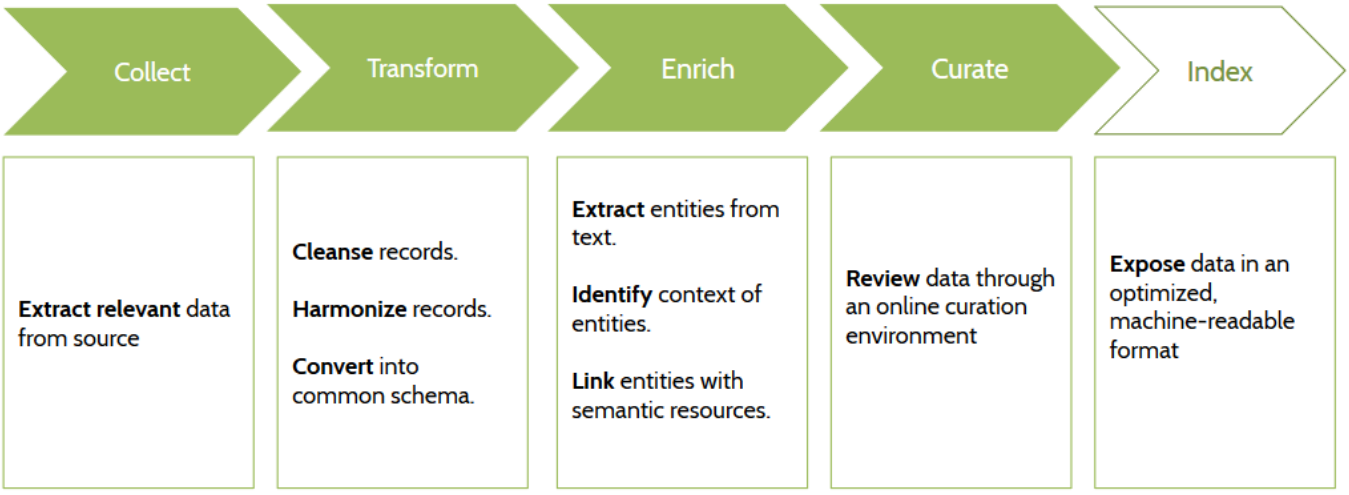

Figure 3: The four functions of the AGINFRA Data Harvesting workflow

The **"Collect" function** is responsible for extracting data from the identified source. This process is mostly appropriated for machine-readable interfaces, such as REST APIs, but it can also be adapted, with some manual intervention, to allow the importation of files into the workflow (see 3.3.2). What was a particularly novel addition to this approach is the concept of "context-aware" collection of data, which allows users to request the harvesting of data assets that are *relevant* to certain thematic requirements, according to pre-selected or engineered semantic resources. In AGINFRA PLUS, these semantic resources were derived from GACS, in the form of three distinct community-specific vocabularies, documented in

[9] https://ring.ciard.net/web-apis/aginfra-rest-api

Deliverable D2.2. The intention of each resource is to be used as query input for any identified harvesting target, so that only relevant results to the thematic scope of each community can be yielded from any source.

The **"Transform" function** is used to bring all collected data under the umbrella of a common schema. The first step towards that end is the cleaning of data records that are irrelevant to the identified thematic scope or generally malformed records that cannot be interpreted or used by the machine. The next step is the harmonization of records, in the sense that their underlying values are normalized to follow the same conventions (date fields transformed to follow a specific date format, numerical fields to use the same decimal separator etc).

The last step of the "Transform" function is for all data collections to be organized according to a common schema, that is manifested at a metadata level. As the same approach was followed by the DCAT-based typologies supported by VRE Catalogues, the Data Harvesting service proposed a FRBR[10]-inspired logic of metadata organization, so that there are basic and common metadata attributes describing each data asset, but also space for more customized metadata attributes per data type.

The **"Enrich" function** is used to generate richer or previously unidentified metadata descriptors for data assets. These descriptors are drawn from the Ontological Engineering layer and more specifically, from the semantic resources that were identified and engineered by the engaged communities. To achieve linking of data to the semantic resources, entity recognition routines are required in the textual content of each data asset and along with its identified context (eg. data type or thematic scope), linking to the appropriate semantic resource concept or class (eg. the link of a food safety risk assessment model to the class of a Hazard from Agroknow's Hazard Taxonomy).

The **"Curate" function** allows human curators to review, organise and enrich data manually. Although optional and time-consuming, this step ensures the quality of data that is generated at the end of the workflow. The tools used in this step consist of an intermediate storage component and a user interface that enables the click-through the different data records. After this step, data is forwarded to the Indexing component of the platform that is responsible for exposing it to consumer applications and users in a performance-optimized machine-readable format.

3.1.2 Service Integration

To fit the data requirements imposed by project activities, a common metadata schema was introduced, based on FRBR. This made the process of integrating new data types much easier, in a way that also conveys a basic thematic and temporal view on the underlying data assets. An example can be seen in the following example:


```
{
    "id" : "AGINFRA_c99dc88e-e27a-4483-95b3-f2378f1b7514",
    "title" : "Salmonella predictor_Growth_Salmonella spp.",
    "description" : "",
    "entityType" : "Model",
    "createdOn" : "2019-09-13T12:55:12.294439",
    "updatedOn" : "2019-09-13T12:55:14.996665",
    "dataSource" : "AGINFRA",
    "tags" : [
      "growth",
      "pork",
      "salmonella predictor",
      "openfsmr",
```


[10] https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf

```
            "salmonella spp."
          ],
          "published" : true,
          "information" : {
            "license_title" : "Creative Commons Attribution Share-Alike 4.0",
            "author" : "taras_guenther",
            "organization_created" : "2018-08-01T14:24:25.147539",
            "organization_name" : "rakip_portal",
            "organization_title" : "RAKIP_portal",
            "type" : "OpenFSMR",
            "DOLU" : "03/29/2019",
            "Item URL" : "http://data.d4science.org/ctlg/RAKIP_portal/1openfsmr09f17d42-
2e74-4067-9683-7a2969966a5d",
            "Model-Foodprocess" : "Storage",
            "Model-InpendenVariables" : "aw, pH, temp",
            "Model-Type" : "Growth",
            "PMF-Environment" : "Pork",
            "PMF-Environment-Details" : "Meat",
            "PMF-Organism" : "Salmonella spp.",
            "Software" : "Salmonella predictor",
            "Software-Accessibility" : "Public / Local installation",
            "Software-Link" :
"http://www.ifr.ac.uk/safety/SalmonellaPredictions/Salmonella_Predictions_V2.xls"
          }
        }
```


The above JSON object indicates a metadata record corresponding to a specific item posted in the AGINFRA PLUS RAKIP_portal VRE and obtained through the AGINFRA PLUS Harvesting Workflow. At the beginning of the object, eight fields denote the more conceptual properties of the item. These are:

- *id:* a unique identifier for the given item
- *title:* a name given to the item
- *description:* a brief text explaining the item
- *entityType:* one of the identified typologies of data
- *createdOn:* the date of creation of the item
- *updatedOn:* the date of last modification of the item
- *dataSource:* the high-level source where the data was pulled from (internal field to the platform)
- *tags:* keywords describing the object

The *manifestation* properties of the item are included in a separate „information“ object and they reveal that this metadata record is about a file, with specific physical characteristics, such as the license (*license_title*), author (*author*) and file type (*type*). The properties that belong inside the „information“ object can change from data type to data type, which allows the ease of introduction of new data types to the registry, as long as they can be described by the higher-level metadata properties presented above.

One last extension that has been achieved through project activities was the integration with the GCat REST Service[11].

With the above extensions, the AGINFRA Data Harvesting Workflows have been tested with all the data assets that have been produced or identified through project activities. The first stress test was performed with the ingestion of *7,524,399 records* from the AGRIS database, that were used for reference

[11] https://wiki.gcube-system.org/gcube/GCat_Service

management and conversion in PENSOFT's ReFindit tool (documented in D4.4)[12]. Later on, it was adapted on 7 more data registries. The complete list of harvested data registries is listed below:

Data registry	AGINFRA
Description	The AGINFRA Data Harvesting Workflow has been fine-tuned so that it flawlessly parses GCat items and re-indexes them without the need of the intermediate curation step. At the same time, it enabled switching of the target repository of harvested records to any D4Science Resource Catalogue, as long as the service is invoked with the User Token[13] that corresponds to a user of the respective VRE with the appropriate permissions.
Data Types harvested	• Models & Algorithms (1889) • Semantic Resources (1507) • Research Objects (199) • Datasets (27) • Software & Services (5)
Total number of records	3627

Data registry	AGRIS
Description	AGRIS is the International System for Agricultural Science and Technology, a multilingual bibliographic database that connects users directly to a rich collection of research and worldwide technical information on food and agriculture.
Data Types harvested	• Publications (7524399)
Total number of records	7524399

Data registry	CGIAR GARDIAN
Description	GARDIAN is an agricultural data search engine, launched by the Platform for Big Data in Agriculture, that makes publications and datasets produced by CGIAR discoverable.
Data Types harvested	• Publications (7526) • Datasets (7658) • Others (2324) • Research Objects (17)
Total number of records	17525

[12] https://support.d4science.org/projects/aginfraplus_wiki/wiki/D44_-_Open_Science_Publication_Technologies#Refindit-tool-enabled-for-AGRIS

[13] https://wiki.gcube-system.org/gcube/GCat_Service#gCube_Authorization_Token

Data registry	USDA
Description	The United States Department of Agriculture, also known as the Agriculture Department, is the U.S. federal executive department responsible for developing and executing federal laws related to farming, forestry, rural economic development, and food.
Data Types harvested	• Datasets (26546) • Organizations (69)
Total number of records	26615

Data registry	Zenodo
Description	Zenodo is a general-purpose open-access repository developed under the European OpenAIRE program and operated by CERN. It allows researchers to deposit data sets, research software, reports, and any other research related digital artifacts.
Data Types harvested	• Publications (26546) • Research Objects (69) • Software & Services (512) • Datasets (320) • Others (33)
Total number of records	14127

Data registry	OpenAIRE
Description	OpenAIRE is the European Open Science Infrastructure, for open scholarly and scientific communication.
Data Types harvested	• Publications (9907) • Projects (2987) • Datasets (1136) • Software & Services (339)
Total number of records	14369

Data registry	eROSA
Description	The e-infrastructure Roadmap for Open Science in Agriculture is an H2020 EU-funded project, that was concluded sucessfully in 2018. Part of the project activities was to map the existing projects, networks, data facilities and sources in the agri-

	food domain. The metadata records for the aforementioned types were harvested as part of the AGINFRA Data e-infrastructure.
Data Types harvested	• Organizations (422) • Data Sources (304) • Projects (142) • Software & Services (58)
Total number of records	926

Data registry	**Agroportal**
Description	Agroportal is a vocabulary and ontology repository for agronomy. It can be used to access and share ontologies. Apart from its usage as the main source in the Ontological Engineering Technologies layer of the project, it was also tested as a data source that can be harvested in an automated manner.
Data Types harvested	Semantic Resource (98546) • GACS (74141) • Environment Ontology (8961)
Total number of records	98546

Data registry	**Unpaywall**
Description	Unpaywall is an open database of 25,273,999 free scholarly articles. It has been tested as a source of Publications metadata, but was not further pursued. A custom codebase was developed to harvest Unpaywall contents, until it was dropped in favor of the AGIFNRA Data Harvesting Workflows.
Data Types harvested	• Publications (19)
Total number of records	19

The data assets presented above are currently indexed and constitute the **AGINFRA Data Registry**, currently made accessible through the AGINFRA Search API (documented in 3.4). The Data Registry currently includes a total of 7,700,153 resources, subject to frequent updates (currently every 12 hours), which is when the AGINFRA Data Harvesting Workflow runs without user intervention. The AGINFRA Data Registry also features a dashboard that provides statistics and access to the underlying data assets, but also to the overarching data linking services associated with the registry. The dashboard is provided as an openly accessible interface through the AGINFRA PLUS website[14] (see Figures 4 & 5).

[14] http://admin.agroknow.com/dashboard/aginfra-plus

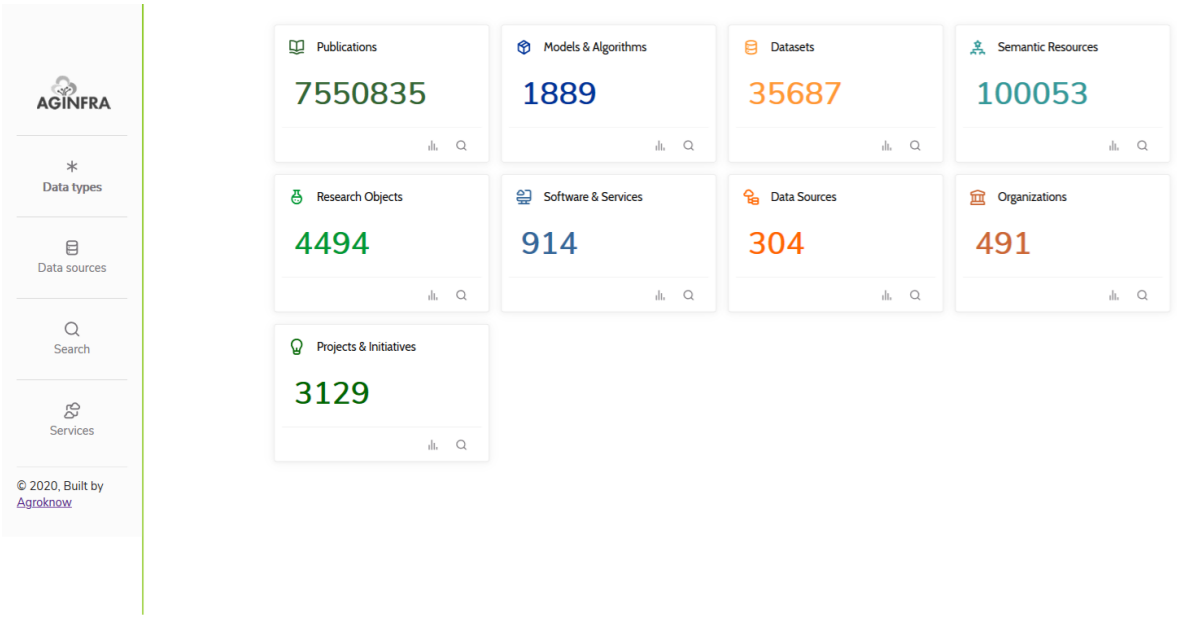

Figure 4: The data types view of the AGINFRA Data Registry Dashboard

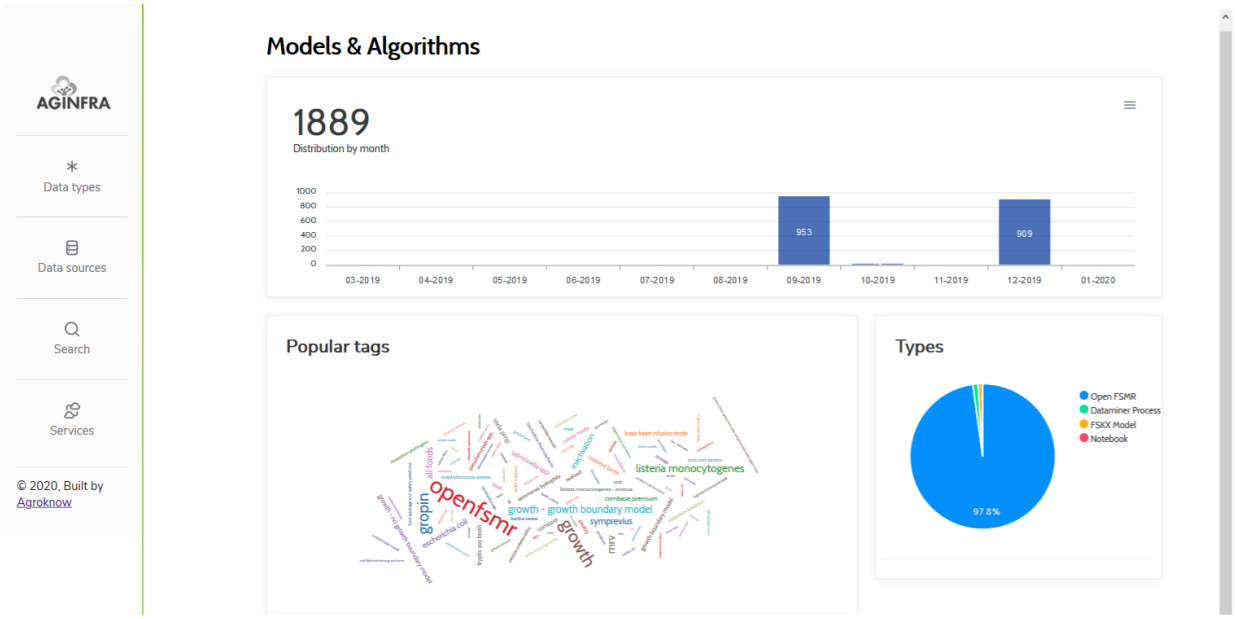

Figure 5: A detailed view of a data type in the AGINFRA Data Registry Dashboard

3.2 DATA ANNOTATION

The proposed Data Annotation service aims at identifying hierarchical relationships that data objects may have with the domain or scope of activities that they reference. Typically, every scientific use-case handles data that is largely untagged with descriptors that puts it in context of a wider domain. For instance, a data asset can simply be a CSV file, if taken out of the context of a scientific workflow. With the appropriate linking routine, said CSV file can be labelled as a food safety resource, a phenotyping experiment record or a simulation algorithm output file, ready to be reused in the other scientific workflows.

To annotate data assets with thematic descriptors, the use of SKOS concept schemes was deemed necessary. In the scope of project activities, one part of the effort was allocated to identifying and managing the appropriate SKOS classifications, while the other was focused on integrating them into a web service that would use them to suggest data links to user inputs, based on text-mining components tested in the scope of previous projects, such as the OpenMinTeD project[15]).

3.2.1 Ontological Engineering of classifications

As documented in Deliverable D2.2, a total of **7 SKOS vocabularies** were made available for reference and management to the engaged communities. Although initially provisioned by diverse sources, vocabularies were extracted, transformed and ported to the AGINFRA PLUS e-infrastructure in the form of *VocBench Projects*. Users were encouraged to manage and extend them through VocBench, but also visualize them if needed through the *SKOS Play!* interface.

3.2.2 Service Integration

To take advantage of the wealth of classifications gathered, three different scenarios were enabled through the project:

1. Manual data annotation of data assets

The case of manual data annotation became the first apparent scenario that made sense to the engaged communities. In the context of project activities, this was materialized in two different ways:

- **Use of Resource Catalogue data publishing capabilities.** In this context, VocBench-stored classifications had to be transformed into selectable keywords in the Resource Catalogues data publishing forms. To achieve this, the AGINFRA registry of semantic resources (documented in D2.2) was employed to export and feed relevant terms into the GCat service responsible for metadata profile management. The terms exported derived from the three major GACS subsets defined by the user communities and were used to extend three data types: *Dataset*, *Research Object* and *Method*.
- **The case of data annotation via dedicated KNIME workflows.** Within the project, BfR extended the FSK-ML data standard, that already provides a dedicated metadata schema for knowledge annotation. Knowledge (e.g. models or data sets) that are provided in an FSK-ML compliant format could therefore automatically inserted into the Resource Catalogues of the corresponding community VREs by applying a dedicated open source KNIME workflows that is registered as a public resource in the DataMiner. This KNIME workflows exploits the FSK-Lab KNIME extension that was developed by BfR in task 4.2 of the AGINFRA+ project as an open source software supporting the adoption of FSK-ML.

2. Data Annotation as a service through AGINFRA Semantic API.

To enable the automated annotation of data assets, a new service was introduced to the AGINFRA Data Linking services: the **AGINFRA Semantic API**.

Data Linking Service	AGINFRA Semantic API
Service description	The AGINFRA Semantic API is responsible for delivering textual analysis and semantic link suggestions in response to user input. If chained correctly, NLP services can be used in conjunction with the Search API (documented in 3.4.2) to produce links to

[15] http://openminted.eu/

	semantic resources (i.e. SKOS vocabularies) that reside in the AGINFRA Registry of Semantic Resources (documented in D2.2). Otherwise, users may invoke the "annotate" service directly to produce links to any or a specific vocabulary available.
Service endpoints	**Ngrams [POST]:** • input: [String] text to analyze, [Integer] size of output n-grams • output: [JSON Array] a list of n-grams, i.e. a contiguous sequence of String items, that can be phonemes, syllables, letters or words. **Stopwords [POST]:** • input: [String] text to analyze • output: [String] text tokens without common words, such as "a", "to" etc. **Tag [POST]:** • input: [String] text to analyse, [String] types of parts-of-speech to be fetched • output: [JSON Array] part-of-speech tags **Annotate [POST]:** • input: [String] text to analyse, [String] vocabulary to use for annotation • output: [JSON Array] list of vocabulary links
Service documentation	https://api.agroknow.com/semantic-api/swagger-ui.html#/
Hosting details	The AGINFRA Semantic API is currently hosted and maintained by Agroknow as a subscription-based REST API. The service code is available on GitHub[16].

The AGINFRA Semantic API and, specifically, the Annotate endpoint have been tested and used on the harvested content of the AGINFRA PLUS VREs, to generate asynchronously or on the spot term suggestions to all data assets harvested, or to selected assets on the interface of the Semantic Search (see 3.4).

3.3 DATA MAPPING

The proposed Data Mapping service aims at formalizing data models (or *schemas*), by enabling users to create links between elements of their datasets to related elements from reference standards or ontologies. By applying this service, users may enforce schemas to their data, transforming them into interoperable assets that can be ported and consumed by eventually more external systems and services that "understand" said schemas. Of course, such an operation is not straightforward when purely API-served and it often requires user input and supervision to define *rules* or *connections* between elements.

3.3.1 Ontological Engineering of data schemas

In the scope of the AGINFRA PLUS e-infrastructure, the definition of data schemas is available through the Ontology Management tools (presented in Deliverable D2.2, section 2.1), i.e. *VocBench, YAM++ and WebVOWL*. By using those, users can define OWL-based definitions of classes that represent real-world objects, along with their properties and links. In the context of data mapping, these definitions can be

[16] https://github.com/AGINFRA-PLUS/SemanticAPI

transformed into tabular data representations which can fully determine the structure and sometimes the actual values of data. This imposes six key rules to ontology definitions:

1. Classes correspond to data objects, not the underlying values of the objects;
2. Class properties are used to denote the elements of the data objects, that can be resolved to either primitive types or derived types;
3. All class properties need to be resolved, if they are to be used for data mapping;
4. Class properties are resolved/instantiated as rdfs:range properties[17];
5. Primitive types can be either of the 19 types listed in the XSD specification[18];
6. Derived types are essentially links to classes or individuals of classes. Said classes will in turn be resolved to the value of their rdfs:label[19] properties or their URI.

If the above key rules are followed, an ontology can be transformed into a tabular data schema, with columns representing properties and their underlying values matched to primitive data types or a controlled list of items that are either labels or URIs to specific classes.

3.3.2 Service Integration and the Data Integration Tool

Although the above scenario was not immediately realized through all project activities, the engineering of new data types progressed throughout the duration of the project, nowadays enumerating **10 different types** of data assets. Those were not clearly connected to particular metadata or data schemas, but instead followed the DCAT-based knowledge organization that was offered by the VREs' Resource Catalogues. This was the base upon which, a FRBR-based global schema was introduced, enabling the adoption of flexible metadata schemas, also known as *smart schemes*. At the same time, a series of API endpoints were launched to fit the knowledge organization paradigm. Those became part of the **AGINFRA Data Integration API**.

Data Linking Service	AGINFRA Data Integration API
Service description	The AGINFRA Data Integration API is responsible for handling data and types harvested in the AGINFRA+ data registry. It provides all necessary CRUD operations for the management of data and semantic resources, while also providing insights to the current state of the platform in terms of supported datatypes.
Service endpoints	**Entity delete [DELETE]:** • input: [String] user authorization key, [String] ID of the entity • output: [HTTP response] HTTP response that indicates the successful deletion of an entity **Semantic resource import [PUT]:** • input: [JSON Object] an object describing a semantic resource that already exists or that is new • output: [HTTP response] HTTP response that indicates the successful creation or update of semantic resources

[17] https://www.w3.org/TR/rdf-schema/#ch_range

[18] https://www.w3.org/TR/xmlschema-2/#built-in-primitive-datatypes

[19] https://www.w3.org/TR/rdf-schema/#ch_label

	Smart scheme import [PUT]: • input: [JSON Object] an object describing a data asset that already exists or that is new • output: [HTTP response] HTTP response that indicates the successful creation or update of data assets
Service documentation	https://api.agroknow.com/data-integration-api/swagger-ui.html#/
Hosting details	The AGINFRA Data Integration API is currently hosted and maintained by Agroknow as a subscription-based REST API. The service code is available on GitHub[20].

To serve the need of data schema enforcement, a front-end tool was also prototyped, utilizing the Data Linking services in-place. The Data Integration Tool[21] was put in place to showcase how tabular data can become compliant with standards and specific formats in a user-supervised manner, through a controlled, yet friendly interface. The tool functions as a data import wizard with the following steps:

1. **Data Source Selection:** The user chooses the source of data to be uploaded (file or API endpoint – currently only files are supported);
2. **Data Type Selection:** The user chooses the type of data to be uploaded (based on the data types obtained from the AGINFRA Data Integration API) (see Figure 5);
3. **Data Upload:** The user uploads the file (CSV, XLS, XLSX formats are currently supported);
4. **Data Mapping:** The user maps their file columns to the expected schema elements, according to the Data Integration API specification (see Figure 6);
5. **Data Editor:** The user proceeds to fill-in missing or erroneous values in an online spreadsheet-like interface (see Figure 7);
6. **Metadata:** The user inputs some metadata that accompany the file that they want to publish and optionally, select the data asset that they want to associate the file with;
7. **Publish:** The file is published in a target repository (by default to the AGINFRA Data Registry).

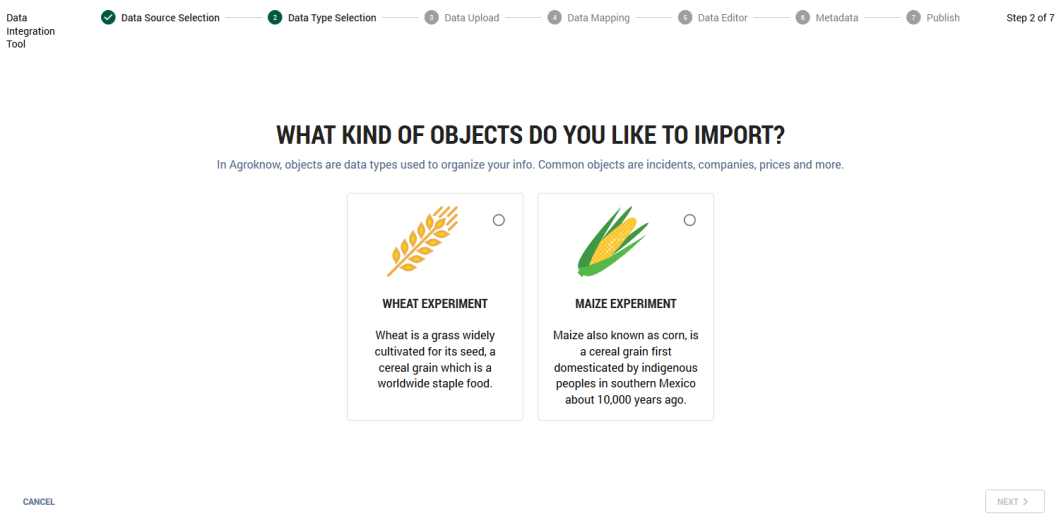

Figure 6: The second step of the Data Integration Tool

[20] https://github.com/AGINFRA-PLUS/DataIntegrationAPI

[21] http://admin.agroknow.com:3006

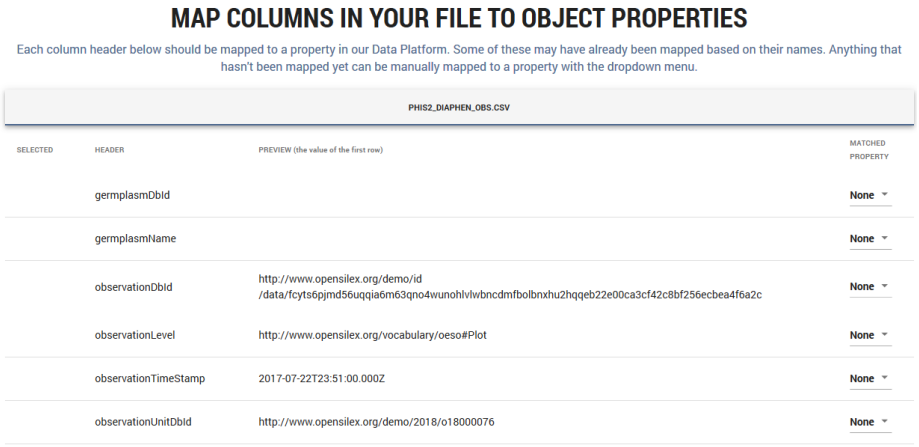

Figure 7: The Data Mapping step of the Data Integration Tool

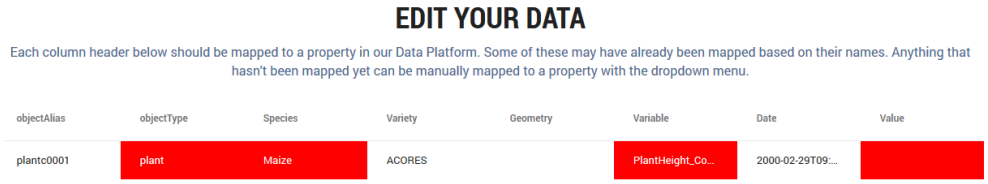

Figure 8: The Data Editor step of the Data Integration Tool

Bits and pieces of the Data Integration Tool have been tested with other research communities engaged in major research projects, such as in DFID-led GODAN Action[22], where the tool was initially used to enrich datasets with geospatial information[23] or in the Bill & Melinda Gates Foundation-funded Global Water Pathogen Project[24], where the tool was used to enforce ad-hoc community data standards to user-uploaded data[25].

In the case of AGINFRA PLUS, the Data Integration Tool has been extended in terms of functionality and it is now prototyped for two data types that were introduced to the FoodSecurity VRE[26] of the AGINFRA Gateway: MaizeExperiment and WheatExperiment. It has been configured to map tabular observation data resulting from crop experiments for wheat and maize to the respective ontologies: CO_321 and CO_322 (documented in Deliverable D2.2). The tool functionality can however be generalized for other types and reference schemes, as long as they are registered at the AGINFRA Data Integration API.

3.4 SEMANTIC DISCOVERY

The last data linking service proposed is the semantic discovery of data assets published on the Resource Catalogues of the AGINFRA PLUS VREs. Taking on performance issues, it leverages big data processing technologies (tested previously in the scope of the BigDataEurope project[27]) to achieve fast response

[22] https://www.godan.info/godan-action/about

[23] http://era.agroknow.com/godanaction

[24] http://www.waterpathogens.org/news/gwpp-water-k2p-translating-knowledge-practice-safe-sanitation

[25] http://dev.k2p.agroknow.com:3000/

[26] https://aginfra.d4science.org/group/foodsecurity/foodsecurity

[27] https://www.big-data-europe.eu/

times and high result relevance. The key technology behind the service is ElasticSearch[28], but the actual bulk of the service code is developed using Spring Boot[29].

The premise of this service was to make the discovery of data assets more accurate by making use of their semantic context, generated through the Ontological Engineering activities of the project. The overall objective was to build a discovery scenario that performs two basic operations:

1. Realization of user intent through the submitted query;
2. Semantic Expansion of the user query based on the relevant semantic resources.

The realization of user intent can be seen as a text analysis task from a machine perspective, so that the most important parts of a user query can be detected, while the unimportant parts are omitted from the second operation. In computational linguistic terms, the output should be a number of string tokens with no leading or trailing stop-words, that correspond to detected concepts from relevant semantic resources.

The *Semantic Expansion* operation accepts the detected concepts as input from the previous operation and performs a relationship lookup in the semantic tree to fetch other concepts that are connected to the original ones via parent-children or sibling relationships. The initially submitted user query is thus enriched with more terms that are semantically relevant and is used to procure more results that match one or more of the expanded terms.

3.4.1 Ontological Engineering of classifications

To build the backbone that would be used for semantic relationship lookup, again all **7 SKOS classifications** (documented in Deliverable D2.2) were used.

3.4.2 Service Integration and the Semantic Search front-end

Initially, the **AGINFRA Search API** was designed to cater the needs of the harvested content of the AGINFRA PLUS Resource Catalogues, along with all SKOS vocabularies that were harvested from the AGINFRA Registry of Semantic Resources (documented in D2.2). Nowadays, the AGINFRA Search API covers millions of data assets that were harvested as part of the AGINFRA Data Harvesting Workflows execution. A description of the services is provided below:

Data Linking Service	AGINFRA Search API
Service description	The AGINFRA Search API is an advanced search service responsible for delivering highly accurate search results over the data harvested within the AGINFRA Data Registry. By disambiguating the meaning of search queries, it detects relationships between possible results and semantic concepts or classes to further enrich the search experience with more relevant results.
Service endpoints	**Search [POST]:** • input: [JSON Object] all search parameters: ○ freetext [String]: the user query ○ apikey [String]: the user authorization key ○ page [Integer]: the page of results to get ○ pageSize [Integer]: the sige of pages

[28] https://www.elastic.co/products/elasticsearch

[29] https://spring.io/projects/spring-boot

	○ strictQuery [JSON Object]: key-value pairs of fields and an explicit value that they should match ○ aggregations [JSON Array]: array of JSON Objects defining facets to be fetched, along with the results ○ smart [Boolean]: a boolean switch that enables or disables the Semantic Search feature ○ method [String]: a selector of the Semantic Expansion lookup method. Accepted values: “children”, “parents”, “siblings” • output: [JSON Object] A JSON Object encapsulating all search results, along with the generated facets.
Service documentation	https://api.agroknow.com/search-api/swagger-ui.html#/
Hosting details	The AGINFRA Search API is currently hosted and maintained by Agroknow as a subscription-based REST API. The service code is available on GitHub[30].

With the use of the above service, a typical search scenario can be improved as following:

1. A user submits a query to the AGINFRA Search API.
2. The AGINFRA Semantic API’s Annotate endpoint is invoked which (internally):
    a. Sends a request to the Ngrams endpoint that creates n-grams from the query terms;
    b. Sends a request to the Stopwords endpoint which removes unnecessary stop-words from the generated n-grams;
    c. Sends a request to the Tag endpoint which finds the parts-of-speech from the given n-grams;
    d. Queries the Search Endpoint to collect semantic resources from the 7 vocabularies that match the generated parts-of-speech. These are the detected semantic terms of the initial query.
3. The AGINFRA API expands the detected semantic terms to their parents/children/sibling terms, according to user selection.
4. The Agroknow Search API performs a search to its underlying content with all the expanded hierarchy of terms. Results with exact free-text matches are returned first, while those that have matched terms from the Semantic Expansion operation come right afterwards. Third in order come the results which present a fuzzy match with the initial query.

A typical example of the above scenario could be the query *"models for poultry“*. This AGINFRA Search API procures **24 results** with exact free-text match for the word "poultry“, but then also starts enumerating results for models tagged with the keywords "chicken“ or "duck“, which are children terms of the family "poultry“, according to the Agroknow Product Taxonomy (see Deliverable D2.2, section 4.2.2).

To fully showcase the functionality of the Semantic Discovery, a front-end prototype[31] was developed over AGINFRA Search API (see Figure 8). The application allows users to submit their queries to the API, but also filter the results by tag, generic data type (as per 2.1) and VRE that produced the data asset. The interface is equipped with a "Semantic Expansion“ toggle, that allows users to view the difference of a simple full-text search and a semantically expanded search. In addition, users may click on the "Classifications“ button underneath each result and invoke the Annotate endpoint on the spot to view

[30] https://github.com/AGINFRA-PLUS/SearchAPI

[31] https://plus.aginfra.eu/semantic-search

the detected terms for the given item that matched the detected terms of the submitted query. This function can also be viewed as a term recommendation service on each harvested item.

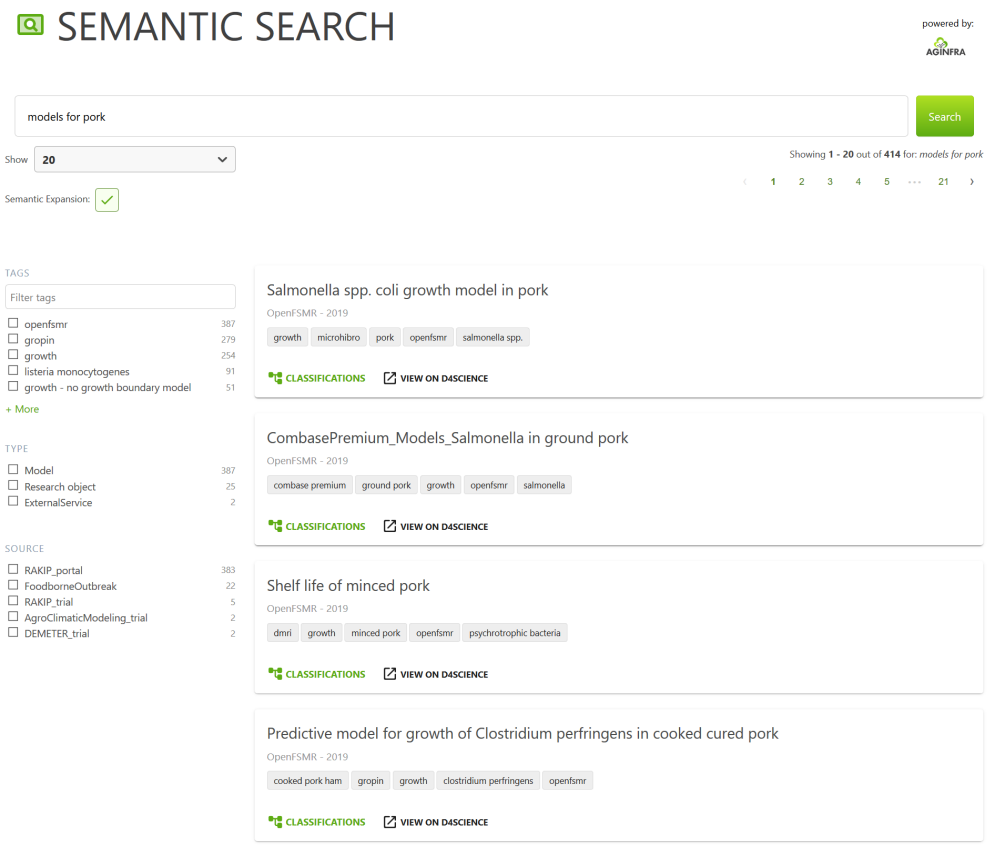

Figure 9: The front-end of the Semantic Search service

The code of the front-end of the Semantic Search service is available on GitHub[32] under an Apache v2.0 license.

[32] https://github.com/AGINFRA-PLUS/semantic-search

4 SUSTAINABILITY AND NEXT STEPS

All data harvested and linked through WP2 activities are openly available through the AGINFRA web portal and search[33]. Agroknow strategically aims at maintaining the AGINFRA data ecosystem and relevant services even after the project duration. More specifically, the sustainability and exploitation pathways of the documented Data Harvesting and Data Linking services are more thoroughly described in D8.4. A registration of the generated registries is already planned on CIARD RING.

Another step towards the exploitation of the proposed services, is the further development and extension of front-end tools associated with them, to increase their TRL and hopefully to become a reference-point for various communities in scientific data mapping and discovery. Exploitation of the Data Integration Tool is already planned in the context of the H2020-funded project BigDataGrapes[34], as a means for data integration from multiple communities engaged in various aspects of grape research and viticulture.

[33] https://plus.aginfra.eu/semantic-search

[34] http://bigdatagrapes.eu/