



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	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 (24 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	Mrs. Georgia Tzenou
WORKPACKAGE N. TITLE	WP2 Data & Semantics Layer
WORKPACKAGE LEADER	Agroknow
DELIVERABLE N. TITLE	D2.3 Data Linking Technologies
RESPONSIBLE AUTHOR	Panagis Katsivelis
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 2017 (M33)
DATE OF DELIVERY (SUBMITTED)	29 September 2017 (M9), 29 November 2019 (35)
VERSION STATUS	2.0 Final
NATURE	De (Demonstration)
DISSEMINATION LEVEL	PU (Public)
AUTHORS (PARTNER)	Nikos Manouselis, Panagis Katsivelis (Agroknow)
REVIEWERS	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

PARTICIPANTS		CONTACT
Agro-Know 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 Recherche 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 Ricerche (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 first submitted iteration of a living document that will describe progress and evolution of the AGINFRA PLUS data linking components, i.e. the services that will be incorporated in the overall AGINFRA PLUS architecture and be responsible for providing indicative links and relations between heterogeneous data assets, e.g. inclusion and relevance relations between publications, datasets and models, similarities in conceptualizations, etc.

The current version of the deliverable focuses on the description of the core linking services commonly used for semantically rich data assets, as well as, integration of a major data linking service workflow to serve as a showcase and baseline of the functionalities to be introduced in AGINFRA PLUS. It is expected that the usage of the data linking services will be significantly generalized based on the needs of the involved research communities, and will be tailored to their needs as well as their usability and ease-of-use requirements.

It is expected that, as the use cases are refined and executed, the data linking services will be updated and extended or modified accordingly. To this end, the report is treated as a living document, with regular submission to the EC of versions that report on significant changes in the respective prototypes.

TABLE OF CONTENTS

EXECUTIVE SUMMARY	6
1 INTRODUCTION	9
2 DATA LINKING APPROACH	10
2.1 DATA TYPOLOGIES AND EXTENSIONS	10
2.2 DATA LINKING AS PART OF A GENERIC SCIENTIFIC WORKFLOW	10
2.3 THE AGROKNOW DATA PLATFORM	12
3 DATA LINKING SERVICES	15
3.1 DATA HARVESTING.....	15
3.1.1 The AGINFRA PLUS Data Harvesting Workflow	15
3.1.2 Service Integration.....	16
3.2 DATA ANNOTATION	18
3.2.1 Ontological Engineering of classifications	18
3.2.2 Service Integration.....	18
3.3 DATA MAPPING.....	20
3.3.1 Ontological Engineering of data schemas	20
3.3.2 Service Integration and the Data Integration Tool.....	20
3.4 SEMANTIC DISCOVERY	23
3.4.1 Ontological Engineering of classifications	23
3.4.2 Service Integration and the Semantic Search front-end	23
4 NEXT STEPS.....	26

TABLE OF FIGURES

FIGURE 1: DATA LINKING SERVICES AS PART OF A GENERIC SCIENTIFIC WORKFLOW	11
FIGURE 2: DATA LINKING SERVICES IN THE CONTEXT OF AGINFRA PLUS E-INFRASTRUCTURE	12
FIGURE 3: THE AGROKNOW DATA PLATFORM ARCHITECTURE	13
FIGURE 4: THE FOUR FUNCTIONS OF THE AGINFRA PLUS DATA HARVESTING WORKFLOW	15
FIGURE 5: THE SECOND STEP OF THE DATA INTEGRATION TOOL.....	22

FIGURE 6: THE DATA MAPPING STEP OF THE DATA INTEGRATION TOOL	22
FIGURE 7: THE DATA EDITOR STEP OF THE DATA INTEGRATION TOOL	22
FIGURE 8: THE FRONT-END OF THE SEMANTIC SEARCH SERVICE	25

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 Initiative¹, which calls for datasets to provide links to other published resources, thus building the continuously expanding Linked Data Cloud². 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.

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;
- 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.

Research Objects are files or metadata records that are generally required to configure scientific workflows (as input), or that result from scientific workflows execution (as output). An instance of such a type could be the output files of a simulation algorithm: tabular summary of executions, provenance records and the actual output in machine-readable format.

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.

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 SemaGrow³). Instead, the majority of the work was centered around *formalizing de facto community standards* (as per RDA Agrisemantics recommendations⁴), so that each of the three project use-cases could align with or even impose new knowledge standards to its target research communities.

³ <http://semagrow.eu>

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

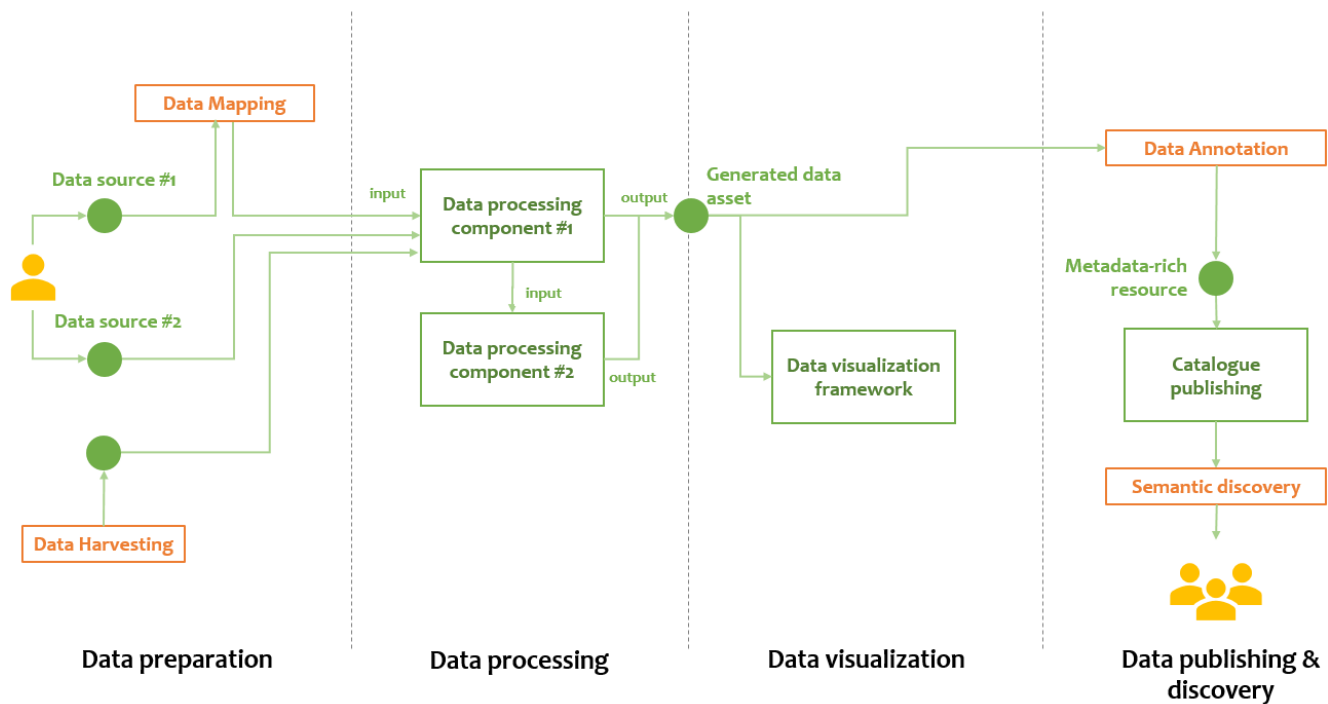


Figure 1: Data Linking services as part of a generic scientific workflow

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 four major 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). To encourage their stability and future exploitation, they were developed as *extensions* of Agroknow's Data Platform components (see 2.3), drawing their content from the knowledge space crafted by the Ontological Engineering layer.

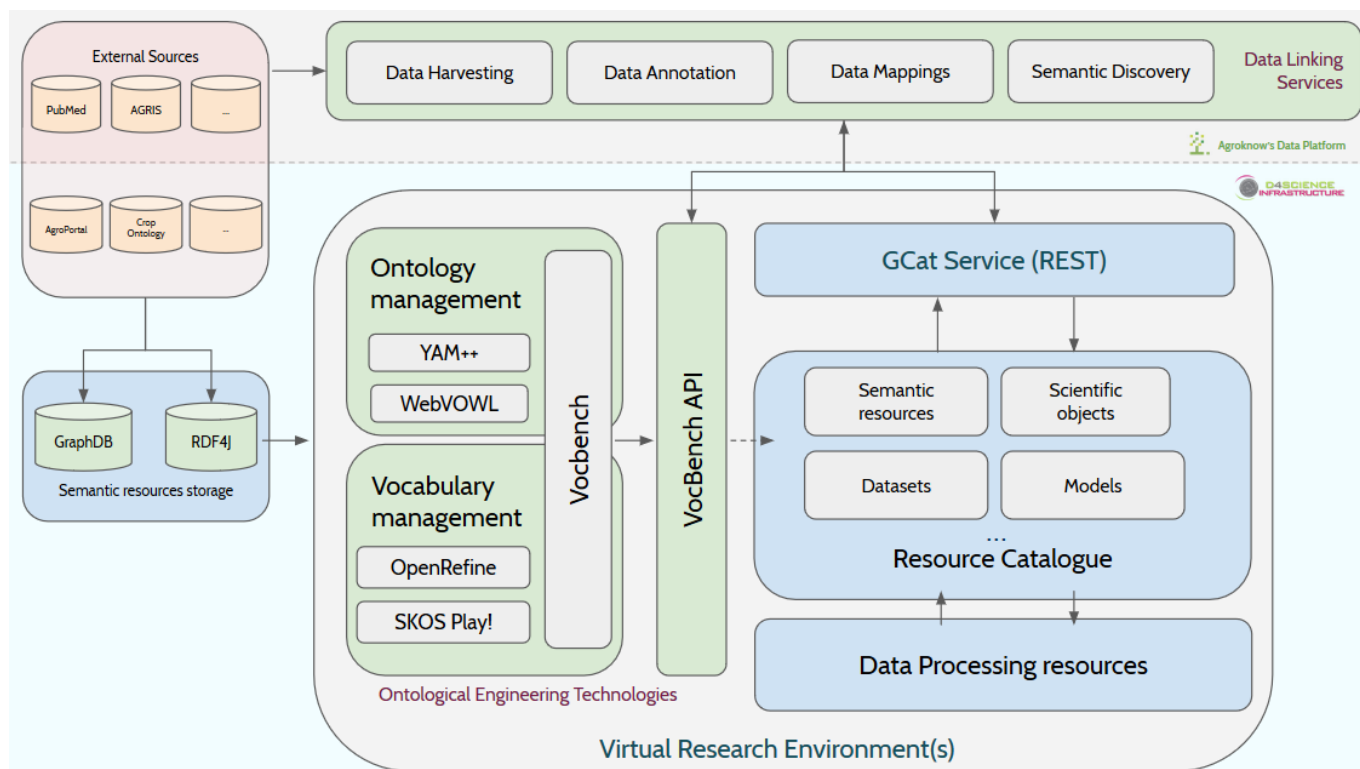


Figure 2: Data Linking services in the context of AGINFRA PLUS e-infrastructure

2.3 THE AGROKNOW DATA PLATFORM

The Agrokno Data Platform is a back-end system responsible for collecting, processing, indexing and publishing agri-food data from various data sources globally. It was first prototyped as an outcome of the initial agINFRA project⁵, mostly dealing with the collection, processing, enrichment and indexing of bibliographic metadata records. Through the years the platform was extended to support RDF storage and querying (through the SemaGrow project⁶), big data processing (through the BigDataEurope project⁷), but also text-mining components (through the OpenMinTeD project⁸).

Nowadays, the platform is organized in a microservice architecture, with different technology components handling different aspects of the data lifecycle. All of the components are interconnected using API endpoints, each responsible for storing and processing different types of data. More specifically, the platform includes:

1. the **Data Discovery** component, where one can search, extract and combine the different types of data collected using the respective API endpoints and an API key,
2. the **Data Harvesting** component, through which data is submitted to the platform through a common schema,
3. the **Data Indexing** component, which performs data transformation to an appropriate format designed for performance optimization,
4. the **Storage** component, which features various storage engine technologies, responsible for the physical archiving of data assets.

⁵ <http://aginfra.eu/>

⁶ <http://www.semagrow.eu/>

⁷ <https://www.big-data-europe.eu/>

⁸ <http://openminded.eu/>

5. the **Knowledge Classification** component, which provides *schema enforcement* to the Data Integration component. This layer consists of collections of semantic resources, tabular data models and metadata profiles that can be used to provide structure to otherwise unstructured data streams that reach the platform. All resources are exposed to an internal Semantic API that serves the other components of the platform.
6. the **Data Processing** component, which is responsible for hosting individual text mining and machine learning scripts that can be used in a variety of contexts as standalone pieces of code or sometimes even available as a service.
7. The **Data APIs** component, which is responsible for exposing and ingesting data to and from different sources, thus serving as the machine-readable interface of the platform. The two main endpoints of this component are: the **Data Integration API** which allows external users to submit or recommend data to the platform and the **Search API** which allows external users to discover assets hosted and managed by the platform.

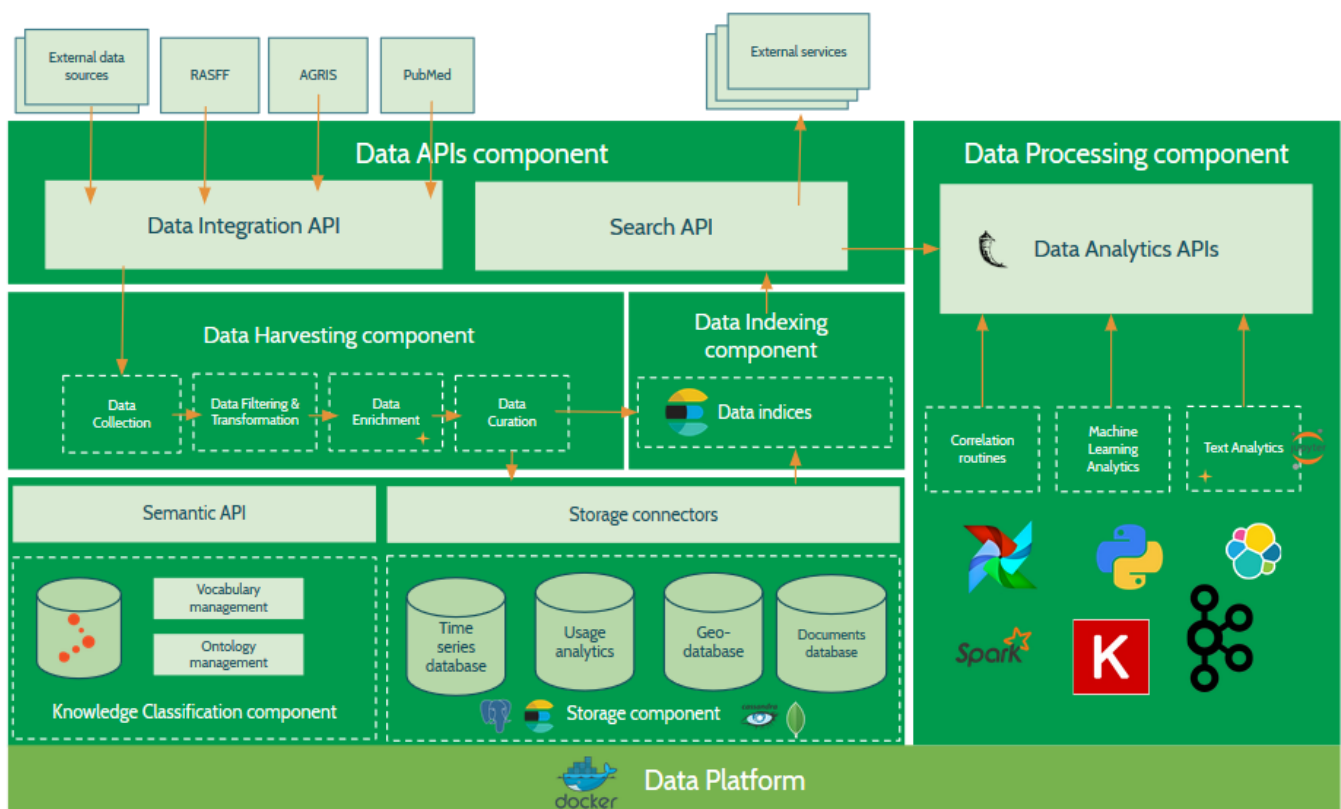


Figure 3: The Agroknow Data Platform architecture

The individual platform components are built using state-of-the-art technologies:

- Spring Boot⁹, for the core and the persistence layer of the platform;

⁹ <https://spring.io/projects/spring-boot>

- Selenium¹⁰, Crawler4J¹¹, Apache Tika¹², Apache PDFBox¹³, PHPEXcel¹⁴, Apache POI¹⁵ for harvesting and parsing of online sources and REST services;
- Apache Kafka¹⁶, for harvesting of data streams;
- Snowball¹⁷, pyjarowinkler¹⁸, Stanford CoreNLP¹⁹, NLTK²⁰, Yandex Translate API²¹ for data enrichment and text analytics;
- Drupal 7.0²² and MySQL²³ for data curation;
- Apache Tomcat²⁴, as the application server hosting the platform;
- ElasticSearch²⁵, FileBeat²⁶ and Logstash²⁷ for data indexing, performance monitoring and usage analytics;
- PostgreSQL²⁸, ElasticSearch, Apache Cassandra²⁹ and MongoDB³⁰ for storage of geodata, usage metrics, numerical values and documents storage respectively;
- Docker³¹, for platform virtualization.

¹⁰ <https://selenium.dev/documentation/en/>

¹¹ <https://github.com/yasserg/crawler4j>

¹² <https://tika.apache.org/>

¹³ <https://pdfbox.apache.org/>

¹⁴ <https://github.com/PHPOffice/PHPExcel>

¹⁵ <https://poi.apache.org/>

¹⁶ <https://kafka.apache.org/>

¹⁷ <https://snowballstem.org/>

¹⁸ <https://pypi.org/project/pyjarowinkler/>

¹⁹ <https://stanfordnlp.github.io/CoreNLP/>

²⁰ <https://www.nltk.org/>

²¹ <https://tech.yandex.com/translate/>

²² <https://www.drupal.org/drupal-7.0>

²³ <https://www.mysql.com/>

²⁴ <http://tomcat.apache.org/>

²⁵ <https://www.elastic.co/products/elasticsearch>

²⁶ <https://www.elastic.co/products/beats>

²⁷ <https://www.elastic.co/products/logstash>

²⁸ <https://www.postgresql.org/>

²⁹ <http://cassandra.apache.org/>

³⁰ <https://www.mongodb.com/>

³¹ <https://www.docker.com/>

3 DATA 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 PLUS Data Harvesting Workflow

The AGINFRA PLUS Data Harvesting Workflow proposed is an extension of the initial agINFRA project “agHarvester” module³² 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. The general functions of the workflow that can achieve this are (as depicted in Figure 3):

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

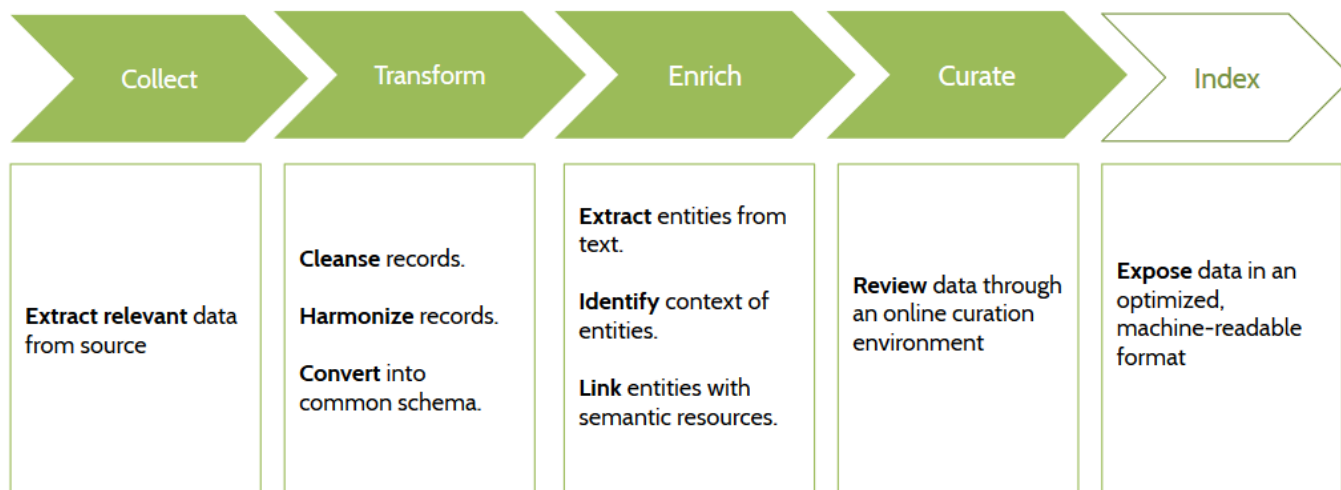


Figure 4: The four functions of the AGINFRA PLUS Data Harvesting workflow

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

³² <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 FRBR³³-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

The individual functions of the proposed harvesting workflow correspond to the existing components of Agroknow’s Data Platform, which were extended to fit the needs of the data assets tapped by project activities. The three extensions performed were:

1. Support for new data types;
2. Introduction of a common metadata schema across data types;
3. Integration with the GCat REST Service

The introduction of a common FRBR-based metadata schema to the platform 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",
```

³³ https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf

```

    "updatedOn" : "2019-09-13T12:55:14.996665",
    "dataSource" : "AGINFRA",
    "tags" : [
      "growth",
      "pork",
      "salmonella predictor",
      "openfsmr",
      "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-IndepenVariables" : "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 Agroknow Data Platform, as long as they can be described by the higher-level metadata properties presented above. Due to this, Agroknow’s Data Platform has been extended to support *Models*, *Research Objects* and *Semantic Resources*, as denoted in 2.1.

One last extension that has been achieved through project activities was the integration with the GCat REST Service³⁴. The AGINFRA PLUS 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 Token³⁵ that corresponds to a user of the respective VRE with the appropriate permissions.

With the above extensions, the AGINFRA PLUS Data Harvesting Workflow has 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 **7,441,477 records** from the AGRIS database, that were used for reference management and conversion in PENSOFT's ReFindit tool (documented in D4.4)³⁶. Nowadays, it features more than **2466 items** of the **5 different typologies** drawn from **7 VREs**, but also more than **5000 publications** drawn from PubMed, matching top level concepts from the GACS sub-vocabularies generated from project activities.

The data assets presented above are currently indexed by Agroknow's Data Platform and they are subject to frequent updates (currently every 12 hours), which is when the Data Harvesting Workflow runs without user intervention.

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.

3.2.1 Ontological Engineering of classifications

As documented in Deliverable D2.2, a total of **7 SKOS vocabularies** was 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, two different scenarios were enabled through the project:

³⁴ https://wiki.gcube-system.org/gcube/GCat_Service

³⁵ https://wiki.gcube-system.org/gcube/GCat_Service#gCube_Authorization_Token

³⁶ https://support.d4science.org/projects/aginfraplus_wiki/wiki/D44_-_Open_Science_Publication_Technologies#Refindit-tool-enabled-for-AGRIS

1. Manual data annotation of data assets in VRE Resource Catalogues;
2. Data Annotation as a service through Agroknow's Data Platform.

The case of manual data annotation became the first apparent scenario that made sense to the engaged communities, as the first data assets were published on Resource Catalogues. To enable this, VocBench-stored classifications had to be transformed into selectable keywords in the Resource Catalogues data publishing forms. To achieve this, the VocBench API (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*.

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

Data Linking Service	Agroknow Semantic API
Service description	The Agroknow 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 semantic resources (i.e. SKOS vocabularies) harvested by Agroknow's Data Platform. Otherwise, users may invoke the "annotate" service directly to produce links to any or a specific vocabulary harvested by the platform.
Service endpoints	<p>Ngrams [POST]:</p> <ul style="list-style-type: none"> 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. <p>Stopwords [POST]:</p> <ul style="list-style-type: none"> input: [String] text to analyze output: [String] text tokens without common words, such as "a", "to" etc. <p>Tag [POST]:</p> <ul style="list-style-type: none"> input: [String] text to analyse, [String] types of parts-of-speech to be fetched output: [JSON Array] part-of-speech tags <p>Annotate [POST]:</p> <ul style="list-style-type: none"> input: [String] text to analyse, [String] vocabulary to use for annotation output: [JSON Array] list of vocabulary links
Service documentation	http://52.214.72.17:9092/swagger-ui.html#/semantic-controller
Hosting details	The Agroknow Semantic API is hosted by Agroknow's Data Platform and is part of a proprietary suite of services that have been extended through AGINFRA PLUS to be exploited commercially in the future.

The Agroknow 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 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` properties³⁷;
5. Primitive types can be either of the 19 types listed in the XSD specification³⁸;
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`³⁹ 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 **16 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 gave rise to the extension of Agroknow’s Data Platform, so that it supports FRBR-based data schemas, labeled 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 **Agroknow Data Integration API**.

³⁷ https://www.w3.org/TR/rdf-schema/#ch_range

³⁸ <https://www.w3.org/TR/xmlschema-2/#built-in-primitive-datatypes>

³⁹ https://www.w3.org/TR/rdf-schema/#ch_label

Data Linking Service	Agroknow Data Integration API
Service description	<p>The Agroknow Data Integration API is responsible for handling data and types hosted by the Agroknow Data Platform. 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.</p>
Service endpoints	<p>Entity delete [DELETE]:</p> <ul style="list-style-type: none"> input: [String] user authorization key, [String] ID of the entity output: [HTTP response] HTTP response that indicates the successful deletion of an entity <p>Semantic resource import [PUT]:</p> <ul style="list-style-type: none"> 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 <p>Smart scheme import [PUT]:</p> <ul style="list-style-type: none"> 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	<p>http://52.214.72.17:9090/swagger-ui.html#/entity-controller</p>
Hosting details	<p>The Agroknow Data Integration API is hosted by Agroknow's Data Platform and is part of a proprietary suite of services that have been extended through AGINFRA PLUS to be exploited commercially in the future.</p>

To serve the need of data schema enforcement, a front-end tool was also prototyped, utilizing the services of Agroknow's Data Platform. The Data Integration Tool⁴⁰ 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 Agroknow 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;

⁴⁰ <http://52.17.48.226:3006/>

7. **Publish:** The file is published in a target repository (by default to Agrokno⁴¹’s Data Platform using the Smart scheme import endpoint).


Data Integration Tool

1 Data Source Selection
2 Data Type Selection
3 Data Upload
4 Data Mapping
5 Data Editor
6 Metadata
7 Publish

Step 2 of 7


WHAT KIND OF OBJECTS DO YOU LIKE TO IMPORT?

In Agrokno, objects are data types used to organize your info. Common objects are incidents, companies, prices and more.



WHEAT EXPERIMENT

Wheat is a grass widely cultivated for its seed, a cereal grain which is a worldwide staple food.



MAIZE EXPERIMENT

Maize also known as corn, is a cereal grain first domesticated by indigenous peoples in southern Mexico about 10,000 years ago.

CANCEL

NEXT >

Figure 5: The second step of the Data Integration Tool

MAP COLUMNS IN YOUR FILE TO OBJECT PROPERTIES

Each column header below should be mapped to a property in our Data Platform. Some of these may have already been mapped based on their names. Anything that hasn't been mapped yet can be manually mapped to a property with the dropdown menu.

PHIS2_DIAPHEN_OBS.CSV			
SELECTED	HEADER	PREVIEW (the value of the first row)	MATCHED PROPERTY
	germplasmDbId		None
	germplasmName		None
	observationDbId	http://www.opensilex.org/demo/id/data/fcyts6pmd56uqia6m63qno4vunohlvibncdmfboibnxbu2hqqeb22e00ca3cf42c8bf25ecbea4fea2c	None
	observationLevel	http://www.opensilex.org/vocabulary/oeso#Plot	None
	observationTimeStamp	2017-07-22T23:51:00.000Z	None
	observationUnitDbId	http://www.opensilex.org/demo/2018/o18000076	None

Figure 6: The Data Mapping step of the Data Integration Tool

EDIT YOUR DATA

Each column header below should be mapped to a property in our Data Platform. Some of these may have already been mapped based on their names. Anything that hasn't been mapped yet can be manually mapped to a property with the dropdown menu.

objectAlias	objectType	Species	Variety	Geometry	Variable	Date	Value
plantc0001	plant	Maize	ACORES		PlantHeight_Co...	2000-02-29T09...	

Figure 7: 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 Action⁴¹, where the tool was initially used to enrich datasets with geospatial information⁴² or in the Bill & Melinda Gates Foundation-funded Global Water

⁴¹ <https://www.godan.info/godan-action/about>

⁴² <http://era.agroknow.com/godanaction>

Pathogen Project⁴³, where the tool was used to enforce ad-hoc community data standards to user-uploaded data⁴⁴.

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 VRE⁴⁵ 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 Agroknow's Data Platform. For this particular case, the target repository where the resulting data end up has been chosen to be the FoodSecurity VRE Resource Catalogue.

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. 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

To enable a full-text search scenario, the **Agroknow Search API** was adapted over the harvested content of the AGINFRA PLUS Research Catalogues, along with all SKOS vocabularies that were harvested from the VocBench API. To integrate the two different types together to enable the Semantic Discovery scenario, an extension to the Search API has been introduced:

⁴³ <http://www.waterpathogens.org/news/gwpp-water-k2p-translating-knowledge-practice-safe-sanitation>

⁴⁴ <http://dev.k2p.agroknow.com:3000/>

⁴⁵ <https://aginfra.d4science.org/group/foodsecurity/foodsecurity>

Data Linking Service	Agroknow Search API
Service description	<p>The Agroknow Search API is an advanced search service responsible for delivering highly accurate search results over the data harvested by Agroknow's Data Platform. 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.</p>
Service endpoints	<p>Search [POST]:</p> <ul style="list-style-type: none"> input: [JSON Object] all search parameters: <ul style="list-style-type: none"> freetext [String]: the user query apikey [String]: the user authorization key page [Integer]: the page of results to get pageSize [Integer]: the size of pages 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	http://52.214.72.17:9091/swagger-ui.html#/search-controller
Hosting details	<p>The Agroknow Search API is hosted by Agroknow's Data Platform and is part of a proprietary suite of services that have been extended through AGINFRA PLUS to be exploited commercially in the future.</p>

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

1. A user submits a query to the Agroknow Search API.
2. The Agroknow NLP 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 Agroknow Search 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 Agroknow Search API procurs **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 prototype⁴⁶ was developed over Agroknow's 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 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.

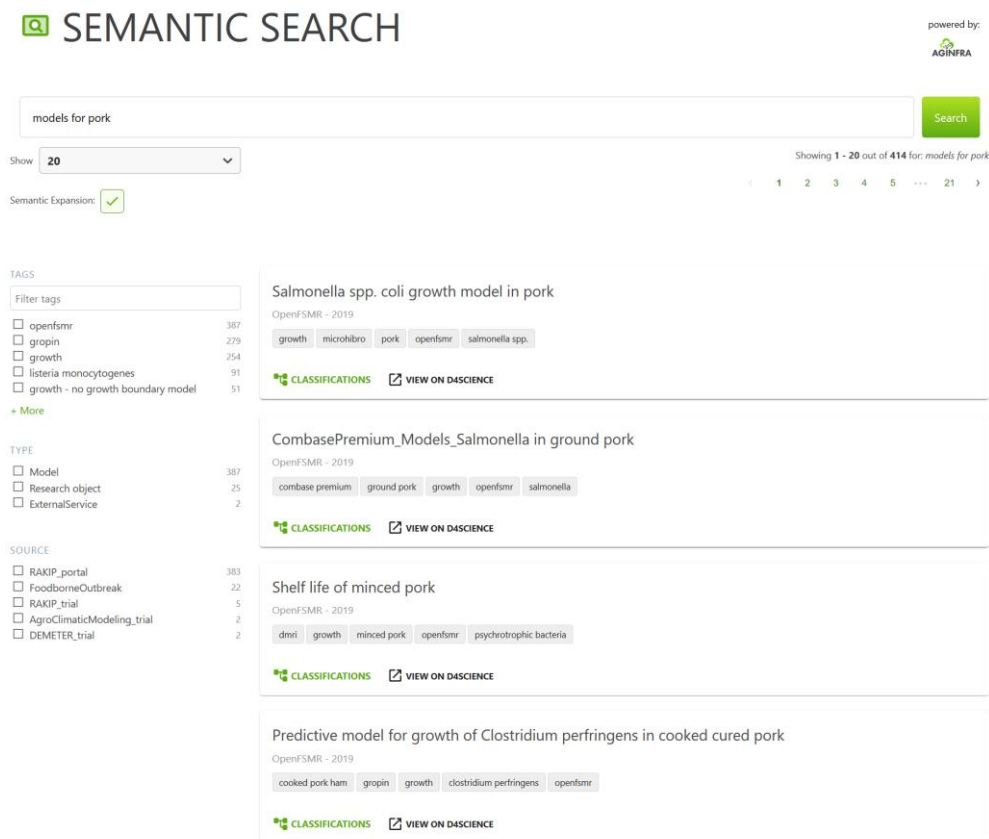


Figure 8: The front-end of the Semantic Search service

⁴⁶ <https://plus.aginfra.eu/semantic-search>

4 NEXT STEPS

The first step towards the finalization of the proposed AGINFRA PLUS Data Linking services is the improvement of their documentation and accessibility, so that their functionality and value proposition is clear to any user that is interested in performing data operations on any research data type of the agri-food sector. The end-goal of this task is the commercialization of said services through Agroknow's Data Platform, which will play an important role in the e-infrastructure's future sustainability.

In addition to the above step, more technical detail will be provided on the AGINFRA PLUS Data Harvesting Workflow, which, although based on solid technological foundation and rigorous developments, is still not a visible part of the AGINFRA PLUS scientific workflow technologies, regardless of its workflow-based nature. This more in-depth documentation will be featured as an updated version of Deliverable D4.2.

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 BigDataGrapes⁴⁷, as a means for data integration from multiple communities engaged in various aspects of grape research and viticulture.

As the evolution of the front-end tools continues, Agroknow plans on developing an AGINFRA-powered Open Science Discovery Service, hosted under the project website, that will deliver open-science content that has been published or harvested using the AGINFRA PLUS Data Linking Services. At the same online space, a data dashboard will become available for visitors to browse said content and relevant statistics more easily, while linking to the original sources for reference and re-use.

⁴⁷ <http://bigdatagrapes.eu/>