D2.2 - ONTOLOGICAL ENGINEERING TECHNOLOGIES

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

DELIVERABLE NUMBER	D2.2
DELIVERABLE TITLE	Ontological Engineering 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 and Semantics Layer
WORKPACKAGE LEADER	Agroknow
DELIVERABLE N. | TITLE	D2.2 | Ontological Engineering Technologies
RESPONSIBLE AUTHOR	Panagis Katsivelis
REPLY TO	katsivelis.panagis@agroknow.com
DOCUMENT URL	http://www.plus.aginfra.eu/sites/plus_deliverables/D2.2.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 | First submission to the EC
NATURE	De (Demonstration)
DISSEMINATION LEVEL	PU (Public)
AUTHORS (PARTNER)	Panagis Katsivelis, Nikos Manouselis (Agroknow)
REVIEWERS	Teodor Georgiev (PENSOFT)

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
0.1	Preliminary Tools review	31/05/2017	Agroknow
0.3	Harmonization with Requirements	31/07/2017	Agroknow
0.5	Tool deployment	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	Semantic Resources Identification	30/02/2018	Agroknow
1.4	Semantic Resources Ingestion Finalization	30/09/2018	Agroknow
1.5	Tools Finalization	30/06/2019	Agroknow
1.9	Deliverable Review	29/11/2019	PENSOFT
2.0	Submission to the EC	30/11/2019	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 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
OWL	Web Ontology Language
SPARQL	SPARQL Protocol and RDF Query Language
REST	Representational state transfer
JSON	JavaScript Object Notation
XML	Extensible Markup Language
FSK-ML	Food Safety Knowledge Markup Language
OEPO	Ontology of Experimental Phenotyping Objects
CO	Crop Ontology
ENVO	Environment Ontology
PPEO	Plant Phenotyping Experiment Ontology
TO	Trait Ontology
RAKIP	Risk Assessment Modelling and Knowledge Integration Platform
GACS	Global Acricultural Concepts Space

EXECUTIVE SUMMARY

The present report is the first submitted iteration of a living document that will describe progress and evolvement of the AGINFRA PLUS ontology engineering components, i.e. the services that will be incorporated in the overall AGINFRA PLUS architecture and serve the conceptualization design, requirements of the involved research communities.

The current version of the deliverable focuses on the installation and deployment of prominent ontology engineering frameworks that will serve as the baseline for producing the final AGINFRA PLUS ontology engineering components, tailored to the core needs reported by the user groups active within the project and adhering to the requirements of the specified AGINFRA PLUS use cases.

It is expected that, as the use cases are refined and executed, the components will be accordingly updated and extended. Additionally, general developments to the used baseline tools will be monitored and adopted if suitable for the purposes of AGINFRA PLUS. 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 ontology engineering prototypes.

TABLE OF CONTENTS

TABLE OF FIGURES

1 INTRODUCTION

The present document provides details on the role and functionality of Ontology Engineering components within the overall AGINFRA PLUS architecture, reports on the initial selection of the technologies to be used and summarizes the extension and training actions foreseen for the following period.

In its general sense, ontology engineering refers to the process of building formal representations of conceptualizations for a given subject / knowledge domain. Ontology engineering tools, thus, are called to provide the means for experts in the domain at hand to express accurately and intuitively their intended definitions for the concepts and relations that adequately cover the domain.

In the context of AGINFRA PLUS, ontology engineering tools face further, specialized requirements in order to be applicable to the research communities addressed by the project and be in line with the modus operandi of the modern information and knowledge exchange practices.

The following sections present the technical description of the ontology engineering components that serve AGINFRA PLUS, and report on the current content that has been generated and made available through the AGINFRA PLUS infrastructure.

2 REQUIREMENTS AND CANDIDATE TECHNOLOGIES

As analysed in the Technical Specifications report document (D2.1, submitted at M7 of the project), the Data and Semantics layer incorporates two main components pertaining to the ontological engineering aspects of the platform; namely, the *Ontology Authoring tools* and the *Vocabulary Authoring tools*.

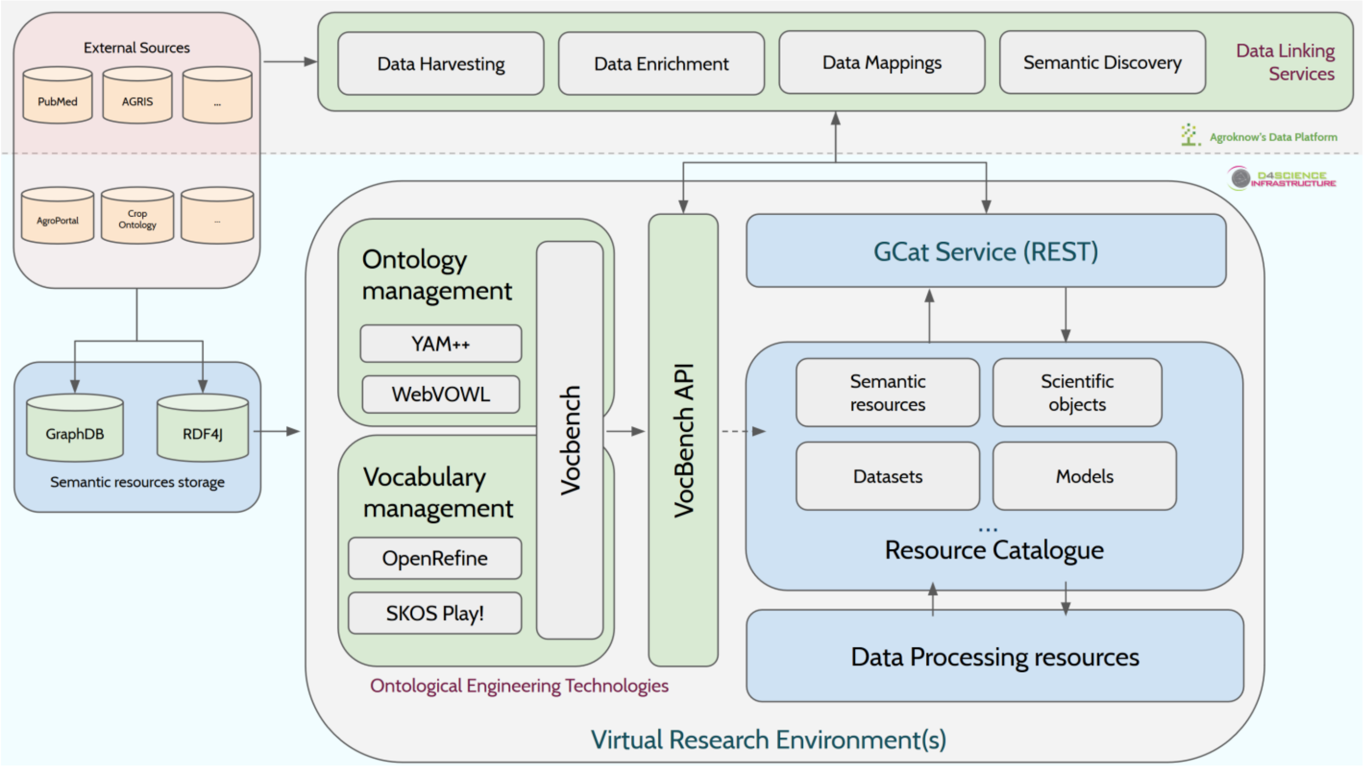


Figure 1: Ontology Engineering Technologies within the overall AGINFRA PLUS Architecture

The following subsections provide a brief overview of these frameworks that will serve the base ground for the AGINFRA PLUS ontology engineering stack.

2.1 ONTOLOGY MANAGEMENT TOOLS

2.1.1 VocBench

VocBench is an authority lists, taxonomies, thesauri and ontologies web-based tool. It is an actively maintained open-source project by the ART group of University of Rome Tor Vergata. Its latest stable version, which was published in July 2017, is 3.6.1. Its installation package is available through the project's website, while the source code is available from the project's Git page[1]. It requires some third-party software components, namely an application server (eg. Apache Tomcat[2]) and optionally the Ontotext GraphDB[3] triple store (free version) installed on an RDF4J server[4].

[1] https://bitbucket.org/art-uniroma2/vocbench3

[2] http://tomcat.apache.org/

[3] https://ontotext.com/products/graphdb/

[4] http://rdf4j.org/

In addition, VocBench provides a detailed change history for each project managed within the tool. At the same time, versioning of semantic resources is also supported, however it is limited to users with elevated privileges.

Regarding collaboration, the users assigned to a project view the same repository (in compliance with their rights), so any changes are propagated on-the-fly. There is also an e-mail notification feature included in VocBench (the mail server serving the notifications is external and configurable by the administrator of the platform).

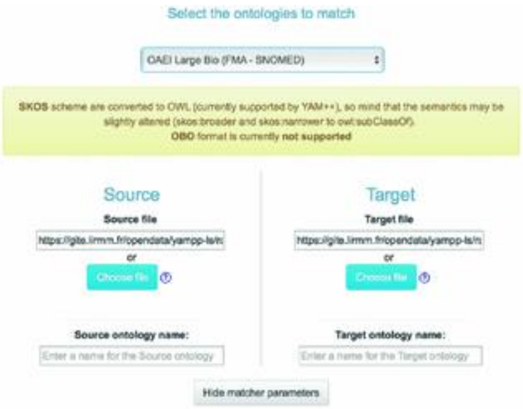
The import function of VocBench 3 accepts various file formats as input (such as Turtle, RDF/XML, JSON-LD) serializes data based on the supported standards and creates instantly manageable resources through the web interface. Regarding the export options available via VocBench, semantic resources can be downloaded in various formats (such as Turtle, RDF/XML and JSON-LD). Last but not least, VocBench supports semantic querying of its resources. It is performed via a SPARQL editor over the data of the active project that a user is working on.

Since its version 3, VocBench supports handling of ontologies expressed in OWL 2. Users may import ontologies in their VocBench project, or start developing their own, using the visual tree-like interface of the tool to define classes, relationships and instances.

2.1.2 YAM++

The *Yam++ Matcher* is a tool for ontology and thesaurus matching, developed by the Montpellier Laboratory of Computer Science, Robotics, and Microelectronics (LIRRM). Its primary function is to discover mappings between entities of two ontologies by using machine learning. The two input ontologies are: a source and a target ontology and the product of the execution is a new ontology with all the newly identified matches, which is stored in RDF/XML format locally.

A web interface for YAM++ exists online and can be used freely, with authorization provided through the project‘s website[5].



Select the ontologies to match

OAEI Large Bio (FMA - SNOMED)

SKOS scheme are converted to OWL (currently supported by YAM++), so mind that the semantics may be slightly altered (skos:broader and skos:narrower to owl:subClassOf).
OBO format is currently **not supported**

Source

Source file: https://gite.lirmm.fr/opendata/yampp-ls/r

Target

Target file: https://gite.lirmm.fr/opendata/yampp-ls/r

Source ontology name: Enter a name for the Source ontology

Target ontology name: Enter a name for the Target ontology

Figure 2: The main screen of YAM++

[5] http://yamplusplus.lirmm.fr/matcher

2.1.3 WebVOWL

WebVOWL is a web tool for ontology visualization, following the Visual Notation for OWL Ontologies[6] and available online as open-source software[7]. To create graph-like visualizations, users may provide their ontologies by uploading them as files or by providing an Internationalized Resource Identifier (IRI).

The WebVOWL environment provides additional options that allow users to filter and export pieces or the entirety of the ontology graph. Another interesting, yet experimental feature of the tool is that it now supports editing of ontologies in a visual manner.

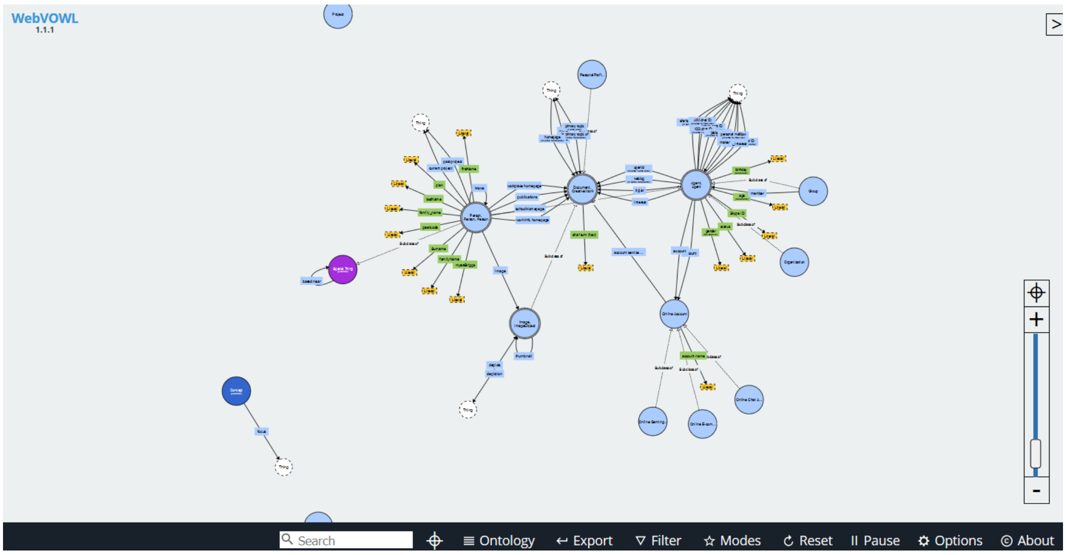


Figure 3: A example ontology visualization created with WebVOWL

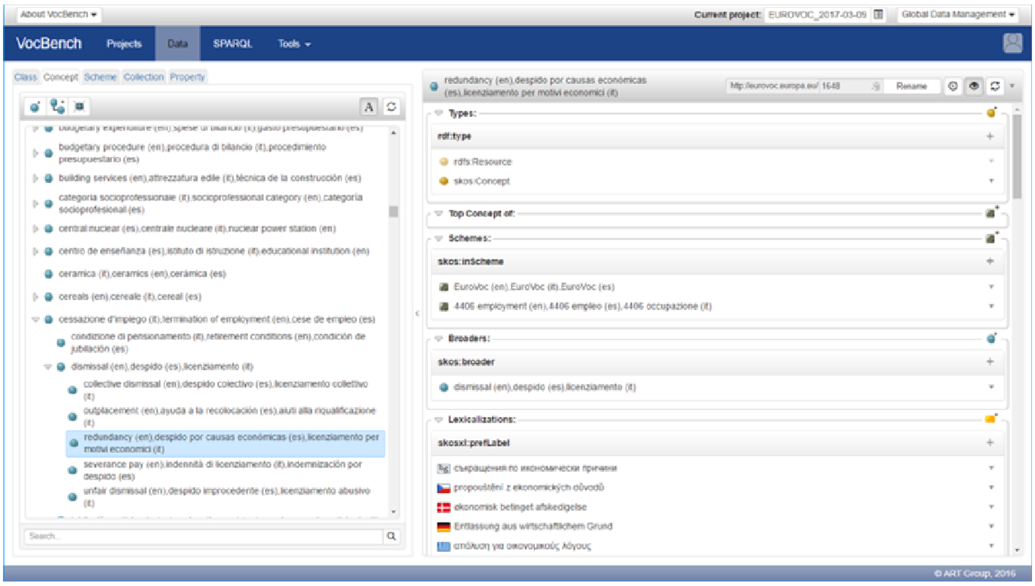
2.2 VOCABULARY MANAGEMENT TOOLS

2.2.1 VocBench

Despite its sizeable list of capabilities, the signature function of VocBench remains the handling of vocabularies expressed in SKOS or SKOS-XL in a collaborative environment that supports four languages (English, Spanish, Dutch, and Thai). Regarding the creation of terms and labels, VocBench support 44 languages in total. These features VocBench an adequate choice for maintaining community-created thesauri.

[6] http://vowl.visualdataweb.org/v2/

[7] https://github.com/VisualDataWeb/WebVOWL



budgetary procedure (en),procedura di bilancio (it),procedimiento presupuestario (es)
building services (en),attrezzatura edile (it),técnica de la construcción (es)
categoria socioprofessionale (it),socioprofessional category (en),categoría socioprofesional (es)
central nuclear (es),centrale nucleare (it),nuclear power station (en)
centro de enseñanza (es),istituto di istruzione (it),educational institution (en)
ceramica (it),ceramics (en),cerámica (es)
cereals (en),cereale (it),cereal (es)
cessazione d'impiego (it),termination of employment (en),cese de empleo (es)
condizione di pensionamento (it),retirement conditions (en),condición de jubilación (es)
dismissal (en),despido (es),licenziamento (it)
collective dismissal (en),despido colectivo (es),licenziamento collettivo (it)
outplacement (en),ayuda a la recolocación (es),aiuti alla riqualificazione (it)
redundancy (en),despido por causas económicas (es),licenziamento per motivi economici (it)
severance pay (en),indennità di licenziamento (it),indemnización por despido (es)
unfair dismissal (en),despido improcedente (es),licenziamento abusivo (it)

redundancy (en),despido por causas económicas (es),licenziamento per motivi economici (it)
http://eurovoc.europa.eu/ 1648

Types:
rdf:type
rdfs:Resource
skos:Concept

Top Concept of:

Schemes:
skos:inScheme
EuroVoc (en),EuroVoc (it),EuroVoc (es)
4406 employment (en),4406 empleo (es),4406 occupazione (it)

Broaders:
skos:broader
dismissal (en),despido (es),licenziamento (it)

Lexicalizations:
skosxl:prefLabel
bg съкращения по икономически причини
propouštění z ekonomických důvodů
økonomisk betinget afskedigelse
Entlassung aus wirtschaftlichem Grund
απόλυση για οικονομικούς λόγους

Figure 4: VocBench Concept Editor

2.2.2 OpenRefine

OpenRefine is an open-source software that is used for data cleanup and transformation. The tool allows users to:

- import and filter/facet their tabular data,
- mass-edit data by cell, column, row or even by regular expressions,
- reconcile data values against other data sources, thus providing external database IDs to datasets

The import and export functions of OpenRefine support many file formats, such as TSV, CSV, text files, XML, JSON, Google Spreadsheets (for import purposes). To also support RDF/XML or Turtle for exporting purposes, users can install the OpenRefine **RDF Extension**[8], which also provides reconciliation against external SPARQL endpoints.

Combining the strengths of OpenRefine's tabular data transformation workflows with RDF provides some interesting possibilities. Specifically, one of the identified scenarios is the transformation of lists of terms or concepts in tabular form into fully-fledged **SKOS controlled vocabularies**. After importing their data, users can create their own „RDF Skeleton“, where columns can be mapped to literals, RDF properties and classes. Once the skeleton is ready, users can export their data in RDF/XML or Turtle, obtaining a full RDF version of their initial tabular dataset in just a few clicks.

Regarding access to the tool, OpenRefine does not feature user authentication or authorization of its resources. Data stored in OpenRefine are organized in projects, where users can easily load more data or perform data transformation operations on the spot.

2.2.3 SKOS Play!

SKOS Play! is a free tool which renders and visualises thesauri, taxonomies or controlled vocabularies expressed in SKOS.

Users may invoke the visualization functions of the tool by providing their resources as a file (supported formats: RDF/XML, Turtle, TriG) or by providing a URL to their SPARQL endpoint or a website featuring

[8] https://github.com/stkenny/grefine-rdf-extension/blob/orefine/README.md

RDFa attributes. The tool will then parse the provided resources and detect KOS hierarchies and visualize them in three different modes: Tree, Square, Circle.

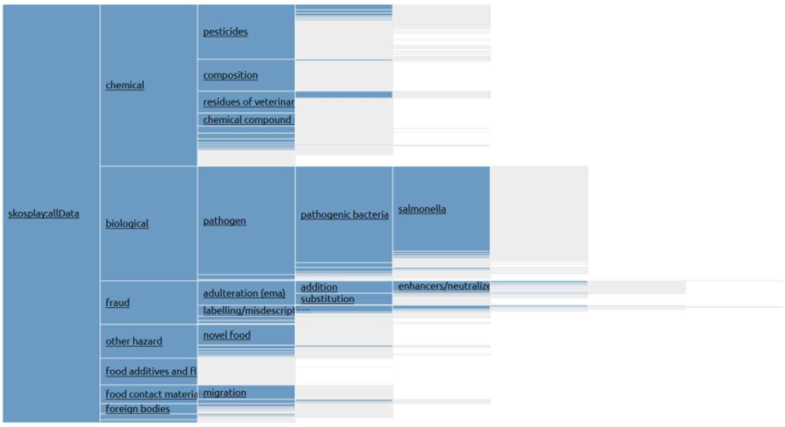


Figure 5: Square visualization created with SKOS Play!

Additional options include editing the rendering order of terms (eg. in alphabetical order), selecting the language of terms to be visualized, but also choosing the output format, which can be in HTML or PDF. Another interesting feature of *SKOS Play!* is the generation of code for autocomplete fields based on the provided thesaurus.

As most of the candidate tools, *SKOS Play!* functions in user-agnostic mode, that doesn't require any form of authentication and doesn't provide any form of usage history and authorization to previously visualized resources.

3 INTEGRATION STATUS

3.1 TOOL INTEGRATION STATUS

All tools described in the previous section have been deployed on the D4Science infrastructure and have been made available as integrated resources within the VREs, but also as standalone tools. The level of integration of each tool is documented below.

3.1.1 VocBench

VocBench's stack suggests a purely decoupled logic that has also been adopted in its integration plan with the D4Science. The front-end of the tool (web interface) is written in Angular 2 and TypeScript, based on the typical model-view-controller (MVC) approach. Its back-end is divided into two layers: the *server layer* and the *storage layer*. The storage layer consists of an RDF4J server and a GraphDB (version 8.5.0) triplestore. The server layer consists of Semantic Turkey server, which is a Java wrapper, that exposes the contents of the storage layer through a JSON API[9].

For the needs of the project, VocBench 3 (version 3.6.0) has been selected, featuring the above bundle that has been tested and deployed on D4Science infrastructure. The Virtual Machine (VM) hosting the instance featured 8 CPU cores, 32 GB RAM and 100 GB of disk space.

In addition, the tightest integration scenario among the identified tools was followed, leveraging the **gCube Authorization framework**[10] to authenticate users. In its current installation within the VREs, users may register and login to VocBench seamlessly, without the need for additional credentials[11].

To achieve this, an extension to the Authentication Layer of VocBench was developed[12], so that visiting users would automatically be registered/logged-in according to their gCube User Token (available within the context of a VRE). The tool is also available outside of the VREs[13], but it handles user accounts individually, following the typical credentials-based registration process.

In its core functionality, VocBench relies on the Roles/Users structure to organize its users and their permissions. There are multiple predefined user roles as presented in the following figure.

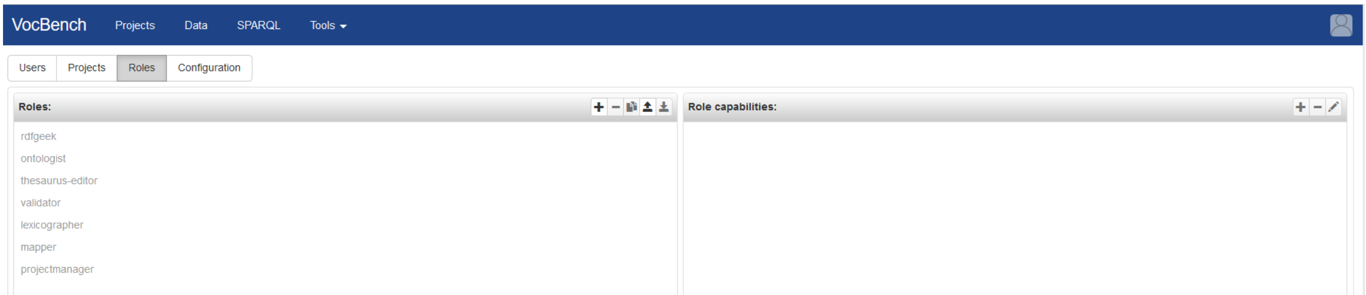


Figure 6: VocBench Role Management

The rights of each user role are defined explicitly by the administrator, and the users are subsequently and mandatorily assigned to one or more roles and one or more projects. In the context of AGINFRA PLUS activities, the typical usage scenario for VocBench was that users of a given VRE would need access to

[9] https://st-vocbench1.d4science.org

[10] https://dev.d4science.org/authorization

[11] https://aginfra.d4science.org/group/aginfraplus/vocbench

[12] https://github.com/AGINFRA-PLUS/vocbench-auth

[13] https://vocbench1.d4science.org/

more than one semantic resource (project), that might have been initiated by users of another VRE. To satisfy the need of horizontal access to different VocBench projects across VREs, the following scenario was followed:

- A user visits VocBench for the first time through VRE A. Their account is created, but they cannot access any VocBench projects.
- VocBench administrator is notified of the new account via e-mail.
- VocBench administrator visits VocBench and enables the account.
- VocBench administrator assigns the user to the appropriate VocBench project which has been maintained by users of VRE B, with the appropriate permissions.
- The user is now free to navigate through VocBench projects that have been assigned to her and perform actions tied to her permissions.

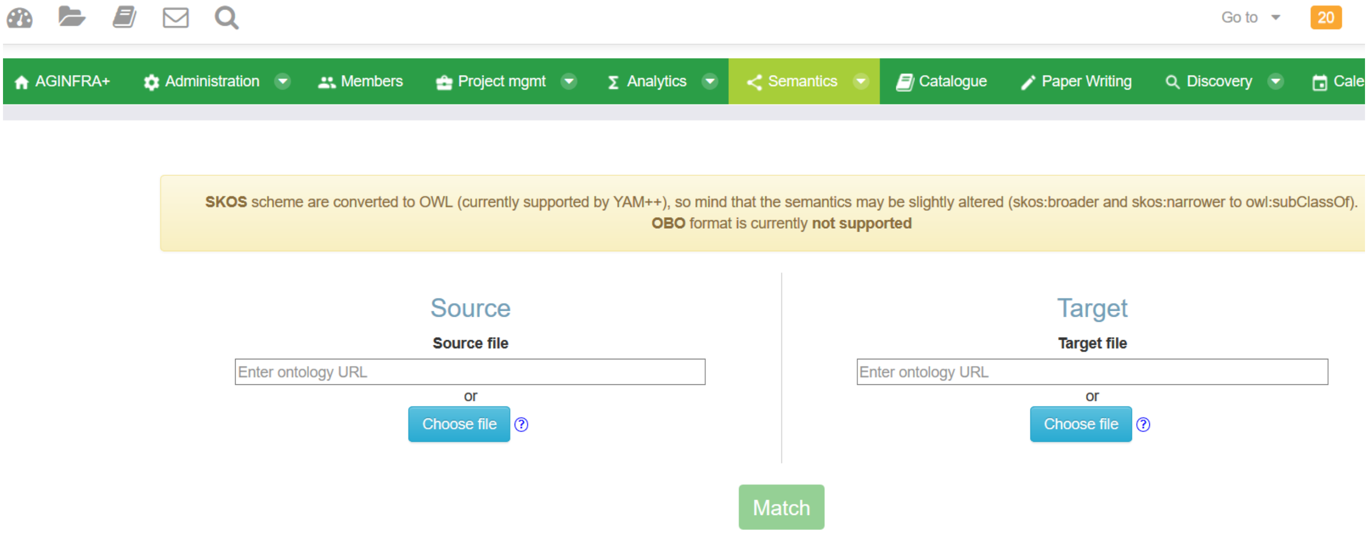
3.1.2 YAM++

YAM++ is available for download online as a Java Maven project[14] that can be built into a JAR library and can be easily invoked through scripting languages with the two required input parameters: a source ontology file and a target ontology file (both in RDF/XML format).

Since the library itself does not come with its own interface, one has been developed with resemblance to the one provided online by LIRRM, featuring the appropriate forms for file uploads. An intermediate PHP layer has also been developed to handle the uploaded ontology files, invoke the underlying library and expose the generated ontology as a downloadable file.

The full package has been wrapped and deployed on a D4Science-provisioned Virtual Machine featuring: 8 CPU cores, 16 GB of RAM and 20 GB of disk space.

YAM++ has been made accessible through the AGIFNRA+ Gateway[15], often under the Ontology Matching label. Thus, user authentication and authorization is handled on a VRE level, without any further need for user credentials.



SKOS scheme are converted to OWL (currently supported by YAM++), so mind that the semantics may be slightly altered (skos:broader and skos:narrower to owl:subClassOf).
OBO format is currently **not supported**

Figure 7: YAM++ integration within the AGINFRA PLUS VRE

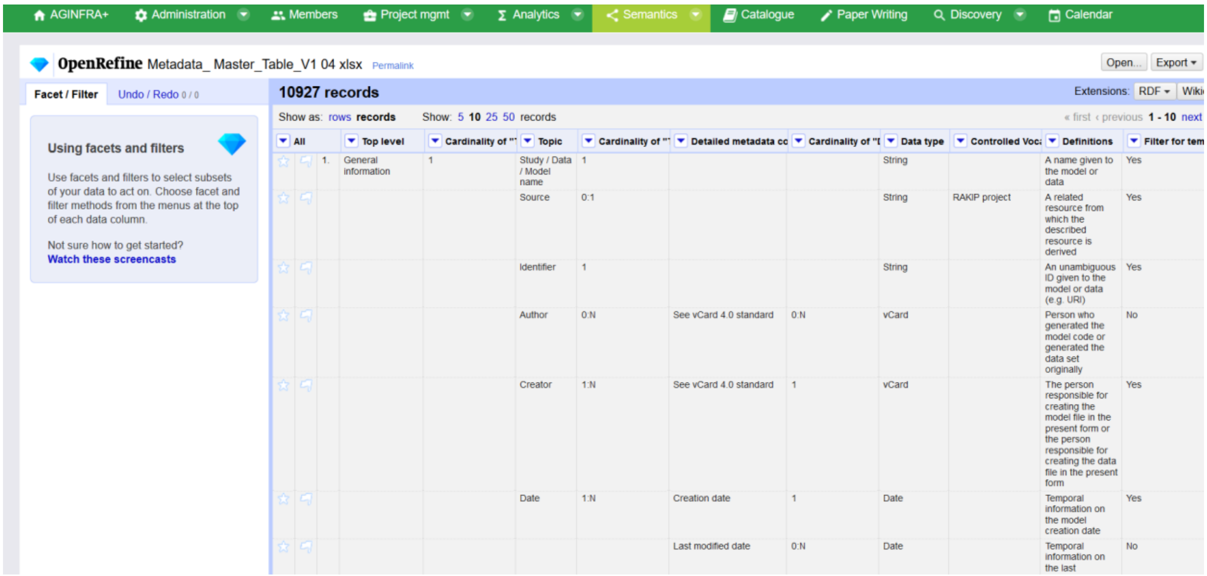
[14] https://gite.lirmm.fr/opendata/yampp-ls

[15] https://aginfra.d4science.org/group/aginfraplus/ontology-matching

3.1.3 OpenRefine

OpenRefine is available for download online on GitHub[16] as a Java Project that can be run via command line in Mac OS and Linux environments. The installation that took place in D4Science infrastructure required a Virtual Machine with 2 CPU cores, 4 GB of RAM and 8 GB of disk space.

The OpenRefine RDF extension was also installed in the project‘s „/extensions“ folder, thus enabling the RDF export scenario indicated in the previous chapter.



OpenRefine Metadata_ Master_Table_V1 04 xlsx Permalink

10927 records

Show as: rows **records** Show: 5 **10** 25 50 records « first ‹ previous **1 - 10** next

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

All	Top level	Cardinality of "'	Topic	Cardinality of "'	Detailed metadata co	Cardinality of "I	Data type	Controlled Voc:	Definitions	Filter for tem
1.	General information	1	Study / Data / Model name	1			String		A name given to the model or data	Yes
			Source	0:1			String	RAKIP project	A related resource from which the described resource is derived	Yes
			Identifier	1			String		An unambiguous ID given to the model or data (e.g. URI)	Yes
			Author	0:N	See vCard 4.0 standard	0:N	vCard		Person who generated the model code or generated the data set originally	No
			Creator	1:N	See vCard 4.0 standard	1	vCard		The person responsible for creating the model file in the present form or the person responsible for creating the data file in the present form	Yes
			Date	1:N	Creation date	1	Date		Temporal information on the model creation date	Yes
					Last modified date	0:N	Date		Temporal information on the last	No

Figure 8: OpenRefine integration within the AGINFRA PLUS VRE

OpenRefine has been made accessible through the AGINFRA PLUS Gateway[17]. Thus, user authentication and authorization is handled on a VRE level, without any further need for user credentials.

3.1.4 WebVOWL

WebVOWL is available on GitHub[18] as an open-source JavaScript project based on Node.js. To deploy it, installing of its dependency packages is required through *npm*, a well-known JavaScript package manager. Its deployment on D4Science infrastructure required a Virtual Machine with 4 CPU cores, 16 GB of RAM and 40 GB of disk space.

WebVOWL has been made accessible through the AGINFRA PLUS Gateway[19]. User authentication and authorization is handled on a VRE level, without any further need for user credentials.

3.1.5 SKOS Play!

SKOS Play! is available on GitHub as an open-source Java Maven project[20], that can be deployed on an application server like Tomcat. For its deployment on D4Science infrastructure, the selected Virtual Machine featured: 2 CPU cores, 4 GB of RAM and 50 GB of disk space.

[16] https://github.com/OpenRefine/OpenRefine

[17] https://aginfra.d4science.org/group/aginfraplus/openrefine

[18] https://github.com/VisualDataWeb/WebVOWL

[19] https://aginfra.d4science.org/group/aginfraplus/ontology-visualization

[20] https://github.com/sparna-git/skos-play

SKOS Play! has been made accessible through the AGINFRA PLUS Gateway[21]. User authentication and authorization is handled on a VRE level, without any further need for user credentials.

3.2 THIRD PARTY COMPONENTS & SOURCES

3.2.1 Ontotext GraphDB

As mentioned in the previous chapter, the deployment of an Ontotext's GraphDB[22] instance accompanied the installation of VocBench on D4Science infrastructure. Although it is optional, the use of GraphDB facilitates loading and browsing hierarchical representations within the tool and is generally well-suited for large quantities of data (files of hundreds of MBs or even GBs[23]). Integrating GraphDB with the stack required extra configurations within the hosting environment, namely memory tweaks that allowed for better performance.

Once integrated, GraphDB can be used as a storage for project creation in VocBench. Users can generally choose between the local RDF4J store or any GraphDB instance that they can use as storage. In the case of AGINFRA PLUS semantic resources, the deployed instance of GraphDB was used to store the largest of ontologies and controlled vocabularies (eg. GACS), as users experienced difficulties in navigating inside large hierarchies.

3.2.2 Agroportal

AgroPortal[24] is an online portal hosting ontologies and controlled vocabularies dedicated to the agronomic and plant domains. The AgroPortal project aims to offer a reference ontology repository for agronomy, reusing the NCBO BioPortal[25] technology. It is developed and maintained by the Montpellier Laboratory of Computer Science, Robotics, and Microelectronics (LIRRM).

In several cases of the project duration, Agroportal has proven to be a valuable source of semantic resources, often tracking changes to reference vocabularies and ontologies and providing them without any notable downtime in a variety of different formats. Some of the key semantic resources used within project activities were drawn from Agroportal and imported on VocBench seamlessly.

3.2.3 Agroknow Data Platform

The Agroknow Data Platform is a back-end system responsible for collecting, processing, indexing and publishing agri-food data from various data sources globally. 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:

- 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,
- the **Data Integration** component, through which data is submitted to the platform in a schema-agnostic manner,

[21] https://aginfra.d4science.org/group/aginfraplusdev/vocabulary-visualization

[22] https://graphdb-vocbench1.d4science.org

[23] http://vocbench.uniroma2.it/doc/sys/#configuring_vocbench_and_graphdb_for_large_quantities_of_data

[24] http://agroportal.lirmm.fr/

[25] https://bioportal.bioontology.org/

- the **Data Indexing** component, which performs data transformation to an appropriate format designed for performance optimization,
- the **Storage** component, which features various storage engine technologies, responsible for the physical archiving of data assets.
- 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 descriptions that can be used to provide structure to otherwise unstructured data streams that reach the platform.
- 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.
- 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 data to the platform and the **Search API**[26] which allows external users to discover assets hosted and managed by the platform.

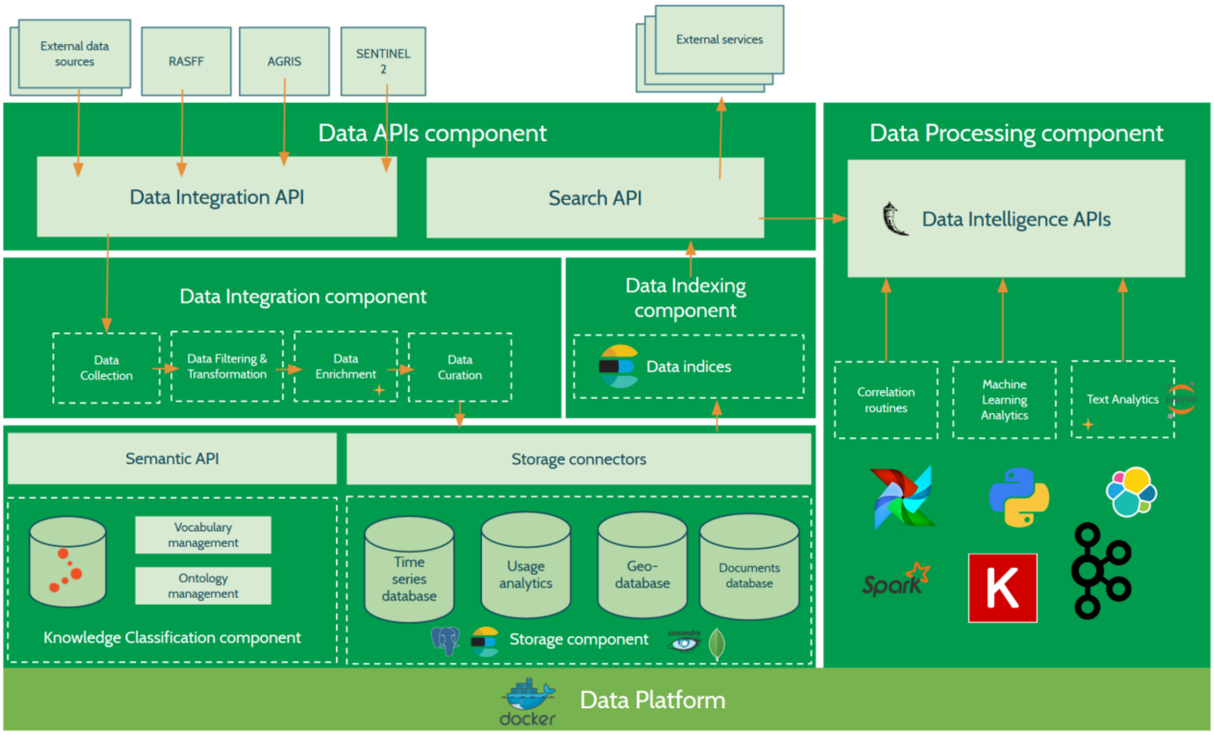


Figure 9: The Agroknow Data Platform architecture

The Knowledge Classification component of the Agroknow Data Platform consists of several semantic resources that range from ad-hoc community terminology to well-known domain-specific thesauri such as AGROVOC or GACS. Today, it hosts more than **20 reference ontologies and vocabularies**, with a total of **185487 concepts and classes**.

[26] http://148.251.22.254:8080/search-api-1.0/swagger-ui.html

4 SEMANTIC RESOURCES

During the project duration, a variety of semantic resources was identified by the engaged use-case communities. Those resources shaped the information space of the majority of the scientific objects produced by the project, either by explicitly stating the data models used per case or by being used for metadata annotation purposes. The most outstanding cases are listed below.

4.1 ONTOLOGIES

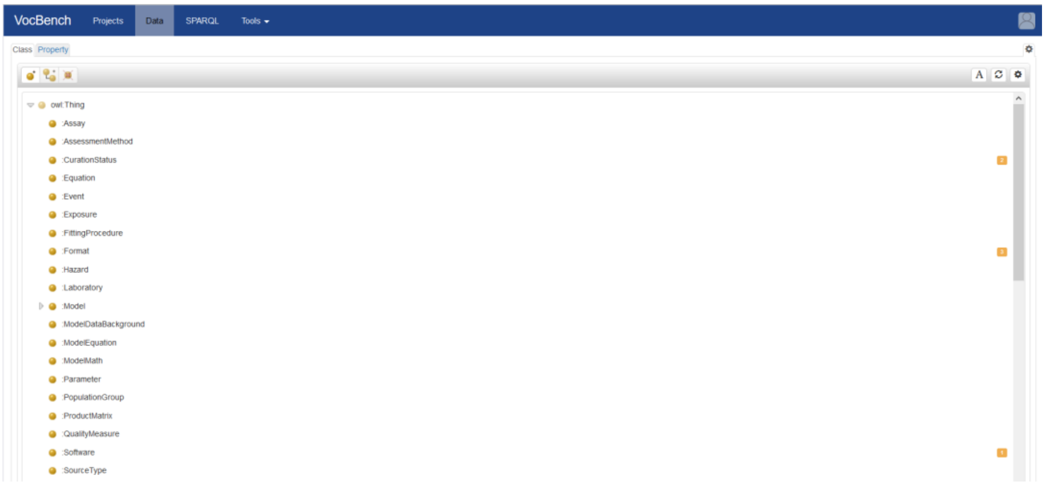
4.1.1 FSK-ML Metadata schema

The Food Safety Knowledge Markup Language (FSK-ML) is developed by the Federal Institute for Risk Assessment (BfR) in Germany, as a means to harmonize the exchange of food safety knowledge. It provides a full set of specifications for accurately describing food safety models in a way that they can become interoperable with various systems and contexts. The most common use of this language is the encoding of FSKX container files that can be used to encapsulate food safety model files, their simulations configuration and metadata[27].

Before the project, the FSK-ML metadata specification was documented in tabular form with the use of MS Excel. Different versions of the specification were directly associated with different XLSX files that were far from being standardized or controlled from a single point or reference system. This discouraged maintenance of the specification by external actors and in the long run, the uptake of it as a standard by the other potential food safety communities.

Semantic resource title	FSK-ML ontology
Semantic resource description	Towards the standardization of the FSK-ML Metadata schema, a collaborative VocBench project was set up to model the metadata schema. An OWL representation of schema classes was defined to depict hierarchy and relationships between them, such as the core inheritance relationship between the generic version of the schema and its adaptations per food safety assessment case (via the *rdfs:subClassOf* property). Another case had to do with the instantiation of specific FSK-ML classes to items from other well-known knowledge classifications, but also ad-hoc community terminology.
# of classes	52
# of individuals	15
# of properties	71
Hosting details	Today, the “FSK-ML ontology” project is an open VocBench project (see Figure 9) with **6 maintainers**.

[27] https://foodrisklabs.bfr.bund.de/fsk-ml-food-safety-knowledge-markup-language/



owl:Thing
:Assay
:AssessmentMethod
:CurationStatus
:Equation
:Event
:Exposure
:FittingProcedure
:Format
:Hazard
:Laboratory
:Model
:ModelDataBackground
:ModelEquation
:ModelMath
:Parameter
:PopulationGroup
:ProductMatrix
:QualityMeasure
:Software
:SourceType

Figure 10: FSK-ML ontology classes view in VocBench

4.1.2 ONTOLOGY OF EXPERIMENTAL PHENOTYPING OBJECTS (OEPO)

Semantic resource title	Ontology for Experimental Phenotypic Objects (OEPO)
Semantic resource description	The Ontology for Experimental Phenotypic Objects (OEPO)[28] is an ontology developed by Institut national de la recherche agronomique (INRA), which allows assigning types to objects involved in phenotyping experiments and defining specialization hierarchy between them according to the specificities of the installations and experiments. OEPO was identified as an ontology to be used for scientific experiments annotation and data integration in food security use-case scenarios. It is actively used as a reference ontology in PHIS, an ontology-driven information system designed for plant phenomics[29], which is currently developed and maintained by INRA.
# of classes	110
# of individuals	-
# of properties	60
Hosting details	OEPO was initially located in Agroportal, from where it was extracted and imported in VocBench, to be used and extended by the community, under the **OEPO Project**.

4.1.3 CROP ONTOLOGY

Crop Ontology is a project of the Integrated Breeding Platform, targeting agricultural data annotation, led by Biodiversity International in collaboration with the CGIAR[30]. The Crop Ontology project aims to compile

[28] http://agroportal.lirmm.fr/ontologies/OEPO

[29] http://www.phis.inra.fr/

[30] http://agroportal.lirmm.fr/projects/CO

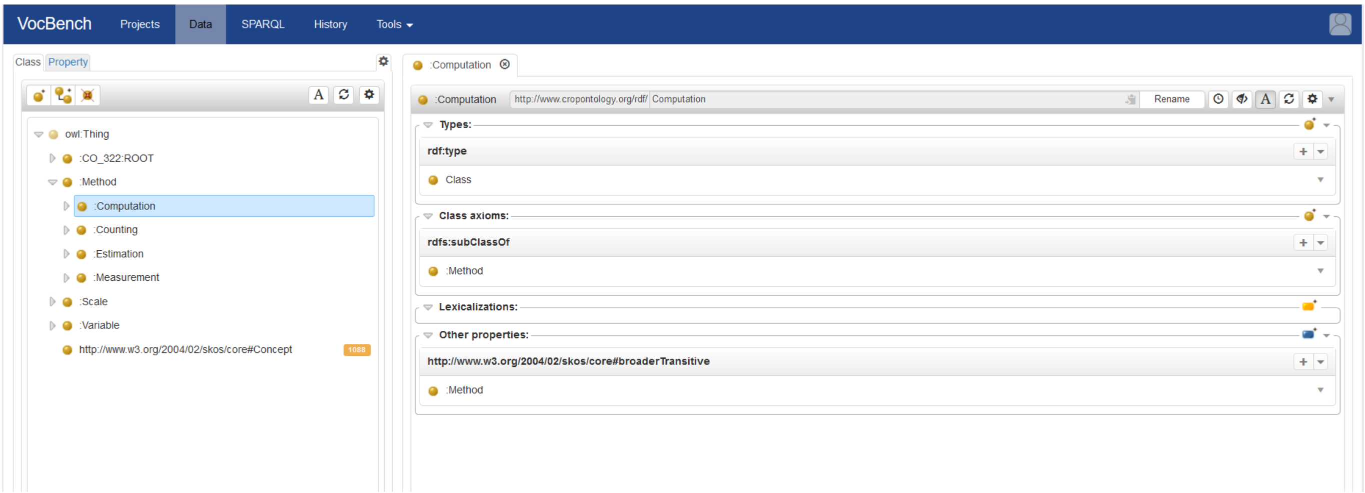
and harmonize trait and measurement information across crop species. Different instances of the Crop Ontology specification exist per species. In the scope of AGINFRA PLUS activities, three of them were picked and imported to VocBench, to be used for annotation and definition of data structures.

Semantic resource title	Crop Ontology – Wheat (CO_321)
Semantic resource description	The Wheat Ontology is a sub-project of the Crop Ontology project, defining traits, methods and scales for wheat crops.
# of classes	1899
# of individuals	1884
# of properties	3
Hosting details	CO_321 was initially located in Agroportal, from where it was extracted and imported in VocBench, to be used and extended by the community, under the **CO-WHEAT Project**.

Semantic resource title	Crop Ontology – Maize (CO_322)
Semantic resource description	The Maize Ontology is a sub-project of the Crop Ontology project, defining traits, methods and scales for maize crops.
# of classes	1103
# of individuals	1088
# of properties	3
Hosting details	CO_322 was initially located in Agroportal, from where it was extracted and imported in VocBench, to be used and extended by the community, under the **CO-MAIZE Project** (see Figure 10).

Semantic resource title	Crop Ontology – Vitis (CO_356)
Semantic resource description	Grape ontology including OIV and biodiversity descriptors, which was created by INRA in July 2017. It defines traits, methods and scales.
# of classes	814
# of individuals	-
# of properties	3
Hosting details	CO_356 was initially located in the Crop Ontology Curation tool[31], from where it was extracted and imported in VocBench, to be used and extended by the community, under the **CO-VITIS Project**.

[31] http://www.cropontology.org/ontology/CO_356/Vitis



owl:Thing
:CO_322:ROOT
:Method
:Computation
:Counting
:Estimation
:Measurement
:Scale
:Variable
http://www.w3.org/2004/02/skos/core#Concept 1088

:Computation http://www.cropontology.org/rdf/ Computation

Types:
rdf:type
Class

Class axioms:
rdfs:subClassOf
:Method

Lexicalizations:

Other properties:
http://www.w3.org/2004/02/skos/core#broaderTransitive
:Method

Figure 11: View of a Maize Ontology class in VocBench

4.1.4 ADDITIONAL ONTOLOGIES

Additional ontologies that were identified and imported in the AGINFRA PLUS semantic resources pool of VocBench are the following:

Semantic resource title	Environment Ontology (ENVO)
Semantic resource description	ENVO is an ontology for the concise, controlled description of environmental entities such as ecosystems, environmental processes, and environmental qualities[32]. It closely interoperates with a broad collection of other OBO ontologies.
# of classes	8969
# of individuals	55
# of properties	165
Hosting details	ENVO was initially located in Agroportal, from where it was extracted and imported in VocBench, to be used and extended by the community, under the **ENVO Project**.

[32] http://agroportal.lirmm.fr/ontologies/ENVO

Semantic resource title	Phenotyping Quality Ontology (PATO)
Semantic resource description	Ontology that defines properties of phenotypes (eg. ectopic, high temperature, fused, small)[33].
# of classes	2746
# of individuals	-
# of properties	23
Hosting details	PATO was initially located in Agroportal, from where it was extracted and imported in VocBench, to be used and extended by the community, under the **PATO Project**.

Semantic resource title	Plant Ontology (PO)
Semantic resource description	The Plant Ontology that defines links between plant anatomy, morphology, growth and development to plant genomics[34].
# of classes	2021
# of individuals	-
# of properties	133
Hosting details	PO was initially located in Agroportal, from where it was extracted and imported in VocBench, to be used and extended by the community, under the **PO Project**.

Semantic resource title	Plant Phenotyping Experiment Ontology (PPEO)
Semantic resource description	The Plant Phenotyping Experiment Ontology, PPEO, is an implementation of the Minimal Information About Plant Phenotyping Experiment[35].
# of classes	31
# of individuals	11
# of properties	86
Hosting details	PPEO was initially located in Agroportal, from where it was extracted and imported in VocBench, to be used and extended by the community, under the **PPEO Project**.

[33] http://agroportal.lirmm.fr/ontologies/PATO

[34] http://agroportal.lirmm.fr/ontologies/PO

[35] http://agroportal.lirmm.fr/ontologies/PPEO

Semantic resource title	Trait Ontology (TO)
Semantic resource description	A controlled vocabulary formatted in OWL that describes the phenotypic traits in plants.
# of classes	5116
# of individuals	-
# of properties	149
Hosting details	TO was initially located in Agroportal, from where it was extracted and imported in VocBench, to be used and extended by the community, under the **TO Project**.

4.2 VOCABULARIES

4.2.1 RAKIP Controlled Vocabularies

Semantic resource title	RAKIP Vocabularies
Semantic resource description	To further build on the food safety modelling case presented in 4.1.1, the food safety community identified a list of vocabularies that also linked to the FSK-ML metadata schema. These vocabularies were used to categorize and annotate food safety models in FSKX format. Their contents were drawn by FAO's CODEX Alimentarius[36] and EFSA's Standard Sample Description[37]. Again, the typical maintenance workflow included the storage and editing of said vocabularies in local MS Excel spreadsheets, impeding interoperability and standardization. To change the above state, the RAKIP controlled vocabularies were imported in OpenRefine. With the definition of RDF Skeletons per vocabulary, they were transformed into fully-fledged SKOS thesauri, that were later on imported into a new VocBench project.
# of vocabularies	47
# of terms	16270
Hosting details	RAKIP Vocabularies now exists as one VocBench project, maintained by **8 users.**

[36] http://www.fao.org/fao-who-codexalimentarius/en/

[37] https://efsa.onlinelibrary.wiley.com/doi/pdf/10.2903/sp.efsa.2015.EN-918

4.2.2 Agroknow's Product Taxonomy

Semantic resource title	Agroknow's Product Taxonomy (FDK_Products)
Semantic resource description	A key classification used to categorize food safety resources is the food product that they refer to. Reference classifications on this area are FoodEx2[38] and Foodvoc[39] , maintained by the European Food Safety Agency (EFSA) and the University of Wageningen respectively. The Agroknow Data Platform provides a new, merged version of these two classifications, extended with terms identified in global food safety incident data, such as product recalls, border rejections and disease outbreaks, that are announced daily by well-known domain agents. The Agroknow Product Taxonomy is an actively maintained project, that is frequently infused with new terms that are checked and curated periodically by expert editors. Nowadays, the taxonomy is used commercially by FOODAKAI[40], a data analytics application for food companies, but is a public resource, open for curation to food safety experts. The current version of the Taxonomy features a full hierarchy of **15607 terms**, some of which in various languages. All terms have been imported and maintained as a project in VocBench.
# of terms	15607
Hosting details	FDK_Products now exists as a VocBench project, maintained by **8 users.** It is also ingested in the Agroknow Data Platform and made available through the platform's Semantic API endpoints.

4.2.3 Agroknow's Hazard Taxonomy

Semantic resource title	Agroknow's Product Taxonomy (FDK_Hazards)
Semantic resource description	Another characteristic of food safety resources is the hazard that they indicate. The Agroknow Data Platform also features a full classification of hazards based on 9 generic groups: *allergens, biological, chemical, food additives and flavourings, food contact materials, foreign bodies, fraud, organoleptic aspects and other hazards*. This classification is used commercially in FOODAKAI but is also available as a public resource for food safety experts.
# of terms	2380
Hosting details	The vocabulary is generally hosted by the Agroknow Data Platform and is made available through the platform's Semantic API endpoints. FDK_Hazards now also exists as a VocBench project, maintained by **8 users.**

[38] https://www.efsa.europa.eu/en/data/data-standardisation

[39] http://www.foodvoc.org

[40] https://www.foodakai.com

4.2.4 GACS

Semantic resource title	GACS
Semantic resource description	During the project duration, a use case-wide need was recorded for metadata annotation with some reference vocabulary. GACS was identified early on as a valuable resource that combines concepts from three different prestigious sources: the AGROVOC multilingual agricultural thesaurus[41] (35000 concepts), the CAB Thesaurus[42] (140000 concepts) and the NAL Thesaurus[43] (53000 concepts). For this purpose, GACS was imported as an individual project in VocBench. Due to its volume, the storage back-end had to be tweaked accordingly, hence the need for GraphDB was highlighted, along with some virtual machine configuration.
# of classes	12
# of individuals	584881
# of properties	3
Hosting details	The vocabulary is generally hosted by the Agroknow Data Platform and is made available through the platform's Semantic API endpoints. FDK_Products now also exists as a VocBench project, maintained by **8 users.**

To make the best use of the newly imported thesaurus, each use case community undertook the task of defining the most important terms that corresponded to its domain. By accessing and browsing the GACS project on VocBench, three major subsets of it were identified, each existing as a separate VocBench project:

GACS – Agroclimatic Modeling

The community of Agroclimatic Modeling hand-picked a group of top-level terms from GACS, to produce a complete vocabulary of **3653 children terms**. These terms exist as separate project on VocBench.

GACS – Food Safety

The Food Safety community picked a complete vocabulary of **2784 children terms**. These terms exist as separate project on VocBench.

GACS – Food Security

Lastly, the Food Security community, picked a vocabulary of **753 terms**, that exist as a separate project on VocBench.

[41] http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus

[42] https://www.cabi.org/cabthesaurus/

[43] https://agclass.nal.usda.gov/

4.2.5 Greek Grape Varieties

Semantic resource title	Greek Grape Varieties
Semantic resource description	As part of the project semantic annotation workflows, a controlled vocabulary for grape varieties was also deemed necessary. The Greek Grape Varieties is a controlled list of definitions of national and international grape varieties cultivated in Greece, along with their translations and hierarchical links. This vocabulary was chosen to enhance the adaption of the CO_356 Vitis ontology with additional linked information to the OIV classifications[44] that the former ontology provides.
# of terms	27
Hosting details	The vocabulary is generally hosted by the Agroknow Data Platform and is made available through the platform's Semantic API endpoints.

[44] http://www.oiv.int/public/medias/5119/code-2017-en.pdf

5 NEXT STEPS

Although the knowledge space of the AGINFRA PLUS data assets and scientific objects has been defined, the pool of semantic resources that is currently stored on D4Science infrastructure is not immediately open to discovery and navigation. To amend this, Agroknow will work on creating an online dashboard that will allow users to navigate through the different semantic resources and their metadata, while at the same time providing documented programmatic access to them.

These semantic resources would be harvested by the Agroknow Data Platform in a way that they feed its Knowledge Classification component. By integrating some of the project-generated semantic resources, links with platform data can be produced, thus enriching it and increasing its commercial value. At the same time, existing mining and classification modules built by related projects like SemaGrow, OpenMinted and the original AGINFRA can be further infused with data and classifications, thus increasing their performance and supported domains.

To encourage uptake of identified semantic resources, interoperability with external systems is of paramount importance. The connection with Agroportal is already clear through the adoption of many of its hosted resources, but additional integration scenarios with the Agroknow Data Platform are already planned. In addition, publishing of commercial vocabularies as open semantic resources through the AGINFRA PLUS infrastructure enables their scientific validation and curation. Aspects of such open science-enabled refinement of specifications will continue to be pursued even after the duration of the project.

Adoption of said resources can also be tested in other domains that are relevant to the themes touched by AGINFRA PLUS and which are currently served by the Agroknow Data Platform. The case of the Global Water Pathogens Project (GWPP)[45] and its evolution to a Bill & Melinda Gates Foundation-funded project[46] provides the space to experiment with knowledge exchange between different scientific domains. The main hypothesis is that knowledge classification of food safety risk assessment modelling data assets can be generalized to be used in the domain of water safety as well.

[45] http://www.waterpathogens.org

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