D2.1 - DATA & SEMANTICS LAYER: TECHNICAL SPECIFICATIONS REPORT

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

DELIVERABLE NUMBER	D2.1
DELIVERABLE TITLE	Technical Specifications Report
RESPONSIBLE AUTHOR	Pythagoras Karampiperis (Agroknow), Antonis Koukourikos (Agroknow)

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

WORKPACKAGE N. | TITLE	WP2 | Data and Semantics Layer
WORKPACKAGE LEADER	Agroknow
DELIVERABLE N. | TITLE	D2.1 | Technical Specifications Report
RESPONSIBLE AUTHOR	Pythagoras Karampiperis (Agroknow), Antonis Koukourikos (Agroknow)
REPLY TO	akukurik@agroknow.com
DOCUMENT URL	http://www.plus.aginfra.eu/sites/plus_deliverables/D2.1.pdf
DATE OF DELIVERY (CONTRACTUAL)	30 June 2017 (M6), 30 June 2018 (M18, Updated version)
DATE OF DELIVERY (SUBMITTED)	26 July 2017(M7), 29 June 2018 (M18, Updated version)
VERSION | STATUS	V2.0 | Final
NATURE	R(Report)
DISSEMINATION LEVEL	PU(Public)
AUTHORS (PARTNER)	Pythagoras Karampiperis (Agroknow), Antonis Koukourikos (Agroknow)
REVIEWERS	Teodor Georgiev (PENSOFT)

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
0.1	Document Structure	01/05/2017	Pythagoras Karampiperis/ Antonis Koukourikos (Agroknow)
0.2	1st draft of Sections 1-2	08/05/2017	Pythagoras Karampiperis/ Antonis Koukourikos (Agroknow)
0.3	Analysis of data assets	22/05/2017	Pythagoras Karampiperis/ Antonis Koukourikos (Agroknow)
0.4	1st draft of functional specs	05/06/2017	Pythagoras Karampiperis/ Antonis Koukourikos (Agroknow)
0.5	Review of use cases	19/06/2017	Pythagoras Karampiperis/ Antonis Koukourikos (Agroknow)
0.6	2nd draft of functional specs	07/07/2017	Pythagoras Karampiperis/ Antonis Koukourikos (Agroknow)
0.7	Finalization of architecture, Draft finalization	21/07/2017	Pythagoras Karampiperis/ Antonis Koukourikos (Agroknow)
0.8	Deliverable Review	27/07/2017	Pythagoras Karampiperis/ Antonis Koukourikos (Agroknow)
0.9	Deliverable finalization	28/07/2017	Pythagoras Karampiperis/ Antonis Koukourikos (Agroknow)
1.0	Submission to the EC	31/07/2017	Pythagoras Karampiperis/ Antonis Koukourikos (Agroknow)
1.1	Use case re-evaluation	28/02/2018	Pythagoras Karampiperis/ Antonis Koukourikos (Agroknow)
1.2	Architectural updates	30/04/2018	Pythagoras Karampiperis/ Antonis Koukourikos (Agroknow)
1.5	Report updates	31/05/2018	Pythagoras Karampiperis/ Antonis Koukourikos (Agroknow)
1.8	Draft finalization	22/06/2018	Pythagoras Karampiperis/ Antonis Koukourikos (Agroknow)
2.0	Submission to the EC	29/06/2018	Pythagoras Karampiperis/ Antonis Koukourikos (Agroknow)

PARTICIPANTS		CONTACT
Agro-Know IKE (Agroknow, Greece)	Agroknow	Nikos Manouselis Email: nikosm@agroknow.com
Stichting Wageningen Research (DLO, The Netherlands)	ALTERRA WAGENINGEN UR	Rob Lokers Email: rob.lokers@wur.nl
Institut National de la Recherché Agronomique (INRA, France)	INRA SCIENCE & IMPACT	Pascal Neveu Email:pascal.neveu@inra.fr
Bundesinstitut für Risikobewertung (BFR, Germany)	BfR Bundesinstitut für Risikobewertung	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)	National and Kapodistrian University of Athens	George Kakaletris Email: gkakas@di.uoa.gr
Stichting EGI (EGI.eu, The Netherlands)	egi	Tiziana Ferrari Email: tiziana.ferrari@egi.eu
Pensoft Publishers Ltd (PENSOFT, Bulgaria)	PENSOFT	Lyubomir Penev Email: penev@pensoft.net

ACRONYMS LIST

DCAT	Data Catalog Vocabulary
RDF	Resource Description Framework
SKOS	Simple Knowledge Organisation System
OWL	Web Ontology Language
SPARQL	SPARQL Protocol and RDF Query Language
ODBC	Open Database Connectivity
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
REST	Representational state transfer

EXECUTIVE SUMMARY

Deliverable D2.1 aims to describe the components and architecture of the AGINFRA+ Data and Semantics layer, positioning it into the overall AGINFRA+ framework and highlighting the alignment of its functionalities with the requirements of the end-user communities represented in the project.

At each of its iteration, the specification is based on the analysis of the latest use case requirements analysis, restructuring the layer if necessary and reporting on relevant and immediately applicable technical solutions for fulfilling these requirements.

As the use cases are implemented, the specification and the corresponding architecture will evolve towards new requirements that arise and towards improving and extending the tools based on user feedback.

Therefore, the report is treated as a living document, with regular submissions to the EC of versions that describe significant changes in design and/or functionality.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The present document reports on the technical requirements for the Data & Semantics Layer, posed by the AGINFRA+ Use Cases. In broad terms, the data to be managed within the AGINFRA+ framework can be classified in the following three categories:

a. Datasets: the category comprises the assets that are subject to the application of algorithms, processing and analysis techniques employed by the different communities on their respective use cases.
b. Models: these are data assets that model and encapsulate the aforementioned algorithmic and processing resources, commonly software modules or systems accompanied by a description/documentation.
c. Publications: these include scientific publications along with their metadata.
d. Semantic Resources: these include ontologies, vocabularies, thesauri and any other resource used to describe and categorize data assets of all four categories.

The following subsections briefly present the main categories of data assets targeted by the use cases of each AGINFRA+ community. For further details, the reader is encouraged to refer to the respective project deliverables (D5.1, D6.1, and D7.1).

1.1 AGRO-CLIMATIC AND ECONOMIC MODELLING COMMUNITY DATA ASSETS

The use cases reported in D5.1 mainly pose demands on the processing and visualisation layers of the AGINFRA+ framework. However, the envisioned pilots combine a multitude of data in different modalities and provided by different entities and organizations. Namely, while the basis for the examined use cases is the AgroDataCube collections, the execution of the experiments require access and linking to soil and crop databases, weather and climate services at the regional, national and global scale, as well as, raw Sentinel and MODIS data.

For the data and semantics layer, these translate to a need for establishing programmatic connectivity with the relevant repositories and services indicated by the use case partners, as well as, the establishing of mechanisms for annotating these data in order to define their applicability and usage on the environmental models to be executed.

1.2 FOOD SAFETY RISK ASSESSMENT COMMUNITY DATA ASSETS

The main vision of the Food safety risk assessment use cases comprises to the creation of food safety *knowledge bases*, i.e. repositories for food safety models along with software implementations of the latter. Towards this vision, the community needs to establish information exchange formats, have access to tools and components that allow the definition of standardized descriptive semantic specifications for the food safety models, and have access to platforms and mechanisms that allow the collaboration on the creation of these descriptions, the discovery and retrieval of others' contributions and the linkage of specific data assets with other ones.

1.3 FOOD SECURITY COMMUNITY DATA ASSETS

The data layer challenges pertaining to the use cases of the food security community are mainly related to the heterogeneity of the involved datasets, as well as, the lack of semantic information associated with most of the involved data sources. These use cases will thus extensively need the means to produce or access semantic resources for describing and annotating their data.

2 DATA AND SEMANTICS LAYER REQUIREMENTS

2.1 AGINFRA+ DATA AND SEMANTICS LAYER FUNCTIONAL REQUIREMENTS

Driven by the exploratory analysis of the data requirements for the different use cases of each community, the following core operations relevant to the data and semantics layer are identified.

2.1.1 Creation and maintenance of semantic resources

The bulk of the available data are not self-descriptive and do not carry a formalized description. Given that the inclusion of semantics is critical for maximizing the utility of the data assets across all use cases, an important requirement for the AGINFRA+ data and semantics layer is the ability to create or access semantic specifications. Depending on the domain and the application domain, the semantic resources range from simple vocabularies or taxonomies to full-fledged ontologies defining complex relations.

2.1.2 Data annotation

Based on the available semantic resources, the relevant users should be able to assign metadata descriptions to their data. To this end, the data & semantic layer services should incorporate mechanisms for providing standardized descriptions of the contributed data assets. The components responsible for these functionalities should also incorporate external schemas/definitions, allow revision and history tracking, and foresee changes in the semantic resources used (which would potentially result to the need for updating the existing annotations using these resources).

2.1.3 Data ingestion

In order to execute the intended research processes and workflows, users must be able to efficiently make the targeted datasets available to the models to be executed. However, the incorporation of every relevant data asset within the AGINFRA+ VREs is not a feasible solution, due to both licensing and practical reasons. To this end, the data and semantics layer must establish the communication mechanisms with external repositories and services that provide usable data assets and import or generate descriptions and access information for them, in order to be findable and accessible within the AGINFRA+ infrastructure.

2.1.4 Data linking and mapping

One of the major objectives of AGINFRA+ is to facilitate the exchange of actionable knowledge between researchers, by making discoverable and usable their individual research results. To this end, the data and semantics layer should incorporate the mechanisms for discovering data assets of different modalities (datasets, models, publications) that are applicable to a specific research activity. That is, a model should be linked to publications related to its mechanisms; these publications should be linked to the models they employ and the datasets over which experiments were carried out, and so on.

Furthermore, AGINFRA+ must employ the mechanisms for specifying similarities between different description models for characterising the relevant data assets. Hence, there is a requirement for obtaining *alignments* between ontologies defining the conceptualisations used for classifying and describing the different assets.

2.1.5 Data discovery

A critical requirement for carrying out the AGINFRA+ use cases, as well as, constituting the overall framework practical and extensible, is the incorporation of efficient browsing, searching and filtering mechanisms over the available data. Ideally, these functions go beyond basic text or property matching,

but rather, use the enriched semantics provisioned by the previous operations in order for the users to retrieve more suitable and extensive results from their searches.

2.2 AGINFRA+ DATA AND SEMANTICS LAYER NON-FUNCTIONAL REQUIREMENTS

The following table showcases an initial summary of the non-functional requirements for the components constituting the AGINFRA+ data and semantics layer. As the systems are being made available, the set of non-functional requirements will be updated in order to be in sync with the user experience and satisfaction.

Requirement	Relevant Operation(s)	Description
Collaboration	Creation and Maintenance of semantic resources Data annotation	Community users should be able to collaboratively edit/review/comment on the relevant information
Multilinguality	Creation and Maintenance of semantic resources Data annotation	Interfaces and entity definitions should be able to use the language preferred by the end users
Uptime	Data ingestion	The tools monitoring external sources should guarantee that there is no information loss or obsolescence
Opaqueness	Data ingestion Data Linking	Introduction of additional resources should be made without changes to the operational flow of the end users
Feedback	All	Possible failures or ambiguities should be communicated to the users in a precise manner

Table 1: Data and Semantic Layer non-functional requirements

3 DATA & SEMANTICS LAYER ABSTRACT ARCHITECTURE

Taking into account the functional requirements posed by the needs of the AGINFRA+ use cases, the abstract architecture depicted in Figure 1 is proposed for the data and semantics layer of the infrastructure.

The complete AGINFRA+ solution is centred around the Virtual Research Environments (VREs) provided by the D4Science infrastructure, as they constitute an ideal testbed for configuring, integrating and deploying the AGINFRA+ assets and components under a common environment.

Consequently, the data and semantic layer is designed to maximise the reuse of services and components already available in a VRE, while new tools and services incorporate the appropriate interfaces and communication mechanisms with components of the generic D4Science framework.

As discussed in Section 2, AGINFRA+ will provide to the targeted communities a *Semantic Authoring Service*, for creating, importing, editing and managing semantic resources (e.g. ontologies, thesauri, taxonomies, etc.). The produced resources are stored in a service-specific repository, which does not necessarily follow the standards and interfacing mechanisms of the VRE. The same repository persistently stores the links and mappings discovered and defined by the *AGINFRA+ Linking and Mapping Service*. Finally, a third service producing semantic resources is the *Data Transformation Service*, responsible for the generation of semantic representations from legacy formats and formalisations.

The resources included in the Semantic Resources repository are subsequently annotated and incorporated in a dedicated D4Science catalogue. VRE users that have access to the catalogue are able to browse it and visualise the resources via the *AGINFRA+ Ontology Visualisation Service*.

Ontologies, thesauri and vocabularies that are known to the system can of course be directly used to annotate data assets. Furthermore, the concepts and properties defined by them are added to the schemas used by the VRE and can be used by the VRE's Annotation Service for describing new data assets ingested in the VRE or extending existing descriptions for already available assets. The assets and their respective metadata are then added to the appropriate VRE catalogue (for datasets, models or publications).Given the openness of the system and the seamless incorporation of additional schemas and ontologies, the user will be free to annotate the resources from different aspects (scope, provenance, access and usage rights, scientific and mathematic characteristics). Different users may also use different metadata models for the same resource, thus collaboratively providing a cross-domain and cross-community description for a given asset.

The initial integration point is the Authentication and Authorisation service. The AGINFRA+ components use the VRE AA layer for verifying authenticated users or are exposed within the VRE to users that have access permissions to a given service.

The annotation process can be applied either to assets actually transferred and stored in the internal VRE repositories, or to external resources for which metadata and location identifiers are accessible. These can be discovered using the relevant Search and Discovery Services exposed through the VRE and operating over external repositories with a known and well-defined access point.

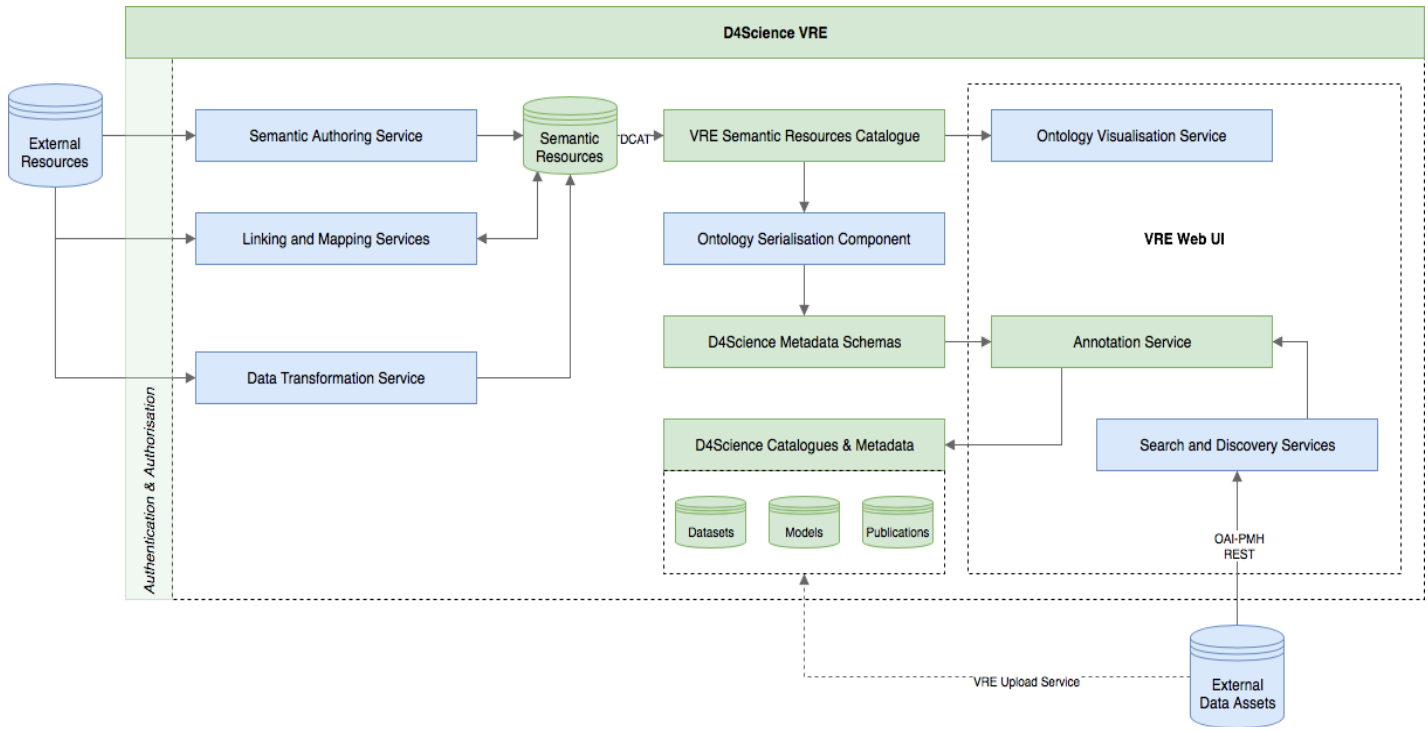

Figure 1: Data and Semantics Layer Abstract Architecture

All functionality and components are available to authorised VRE users. The tools are either fully integrated in the VRE and thus accessible only to registered and logged in users, or – primarily in the case of the more complex authoring system –interfacing with the VRE's Authentication & Authorization Service which is used to access the frameworks.

4 DATA& SEMANTICSLAYERRELEVANTTOOLS &TECHNOLOGIES

This section provides a brief overview of the technology landscape for the tools and services used in the context of the AGINFRA+ Data & Semantics Layer, as presented in Section 3.

4.1 ONTOLOGY & VOCABULARY AUTHORING

The existence and powerful yet intuitive and user-friendly tools for editing semantic resources is crucial for spreading the usage of semantic web technologies. To this end, several efforts to systematize the discovery of such tools and services are being carried out by organisations like the World Wide Web Consortium (W3C)[1] and the Research Data Alliance (RDA)[2]. Some prominent tools and services with a fairly extensive user base and support are described below.

VocBench[3] is a vocabulary and ontology management web-based tool. Its original purpose was the management of the AgroVoc thesaurus. It is actively maintained by the ART group of University of Rome Tor Vergata. Its latest stable version is 3.0.2, with the release of VocBench 3 planned for the end of July 2017.VocBench is an actively maintained open-source project. Its installation package is available through the system's website, while the source code is available from the system's Git page[4]. It requires some third-party software components, namely the Apache Tomcat application server[5], a MySQL server[6], and optionally the Ontotext GraphDB[7] triple store installed on an RDF4J server[8]. All the external components are open-source or free-to-use software.

VocBench is a full-feature management platform, allowing the creation, import, update and merge of vocabularies and ontologies. The adopted UI layout is the standard for systems of this nature, with a navigation panel to the right where the available concepts and properties are presented in a tree structure and a main panel where the edit/updated operations are performed.

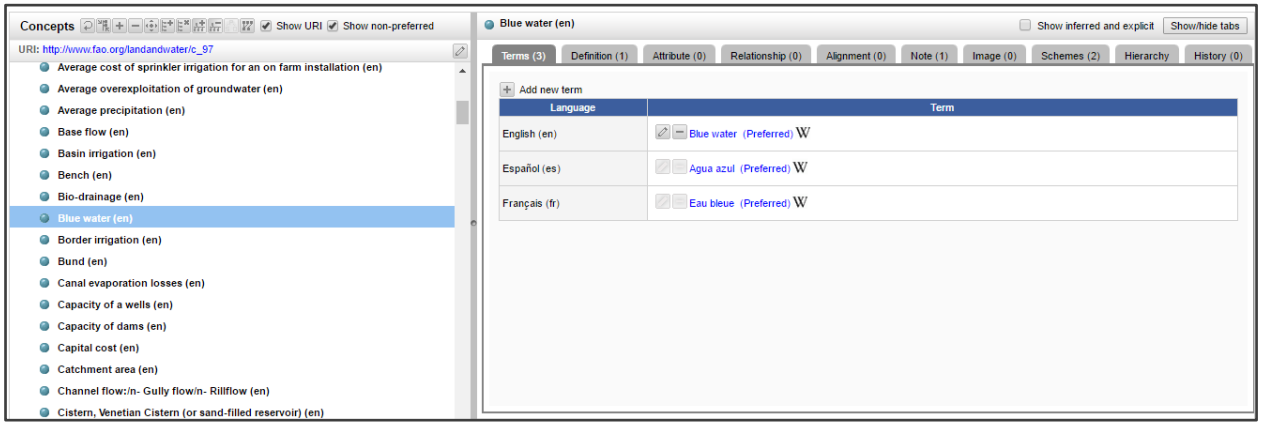

Figure 2: VocBench Concept Editor

VocBench currently supports only the handling of thesauri expressed in SKOS or SKOS-XL and of ontologies expressed in OWL 2.

Querying is performed via a SPARQL editor. There is no functionality for adding predefined queries.

[1] https://www.w3.org/wiki/Ontology_editors

[2] https://www.rd-alliance.org/ands-appraisal-thesaurus-software-tools.html

[3] http://vocbench.uniroma2.it

[4] https://bitbucket.org/art-uniroma2/vocbench2

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

[6] https://dev.mysql.com/downloads/mysql/

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

[8] http://rdf4j.org/

VocBench provides a detailed change history for each project, schema and schema entity managed within the platform. The users authorized to do so can roll back to a previous commit during the lifetime of the project, monitoring the changes applied at each commit.

For large-scale projects, VocBench recommends the usage of the Ontotext GraphDB triple store as the underlying repository. The GraphDB repository to be used is defined at a per-project basis, with different projects potentially using different GraphDB installations.

Regarding the export options available via VocBench, the only supported option at the moment is the export of the created Vocabularies in SKOS or SKOS-XL, in a variety of formats. In addition to the export of an entire project, the user can export specific schemas or even specific concepts included in the active project.

The VocBench environment supports four languages (English, Spanish, Dutch, and Thai). Regarding the creation of terms and labels, VocBench support 44 languages in total. Each user, upon her creation, is associated with their languages of expertise. They are consequently able to create content solely on those languages.

VocBench relies on the Groups/Users structure to organize its users and their permissions. There are multiple predefined user groups as presented in the following figure.

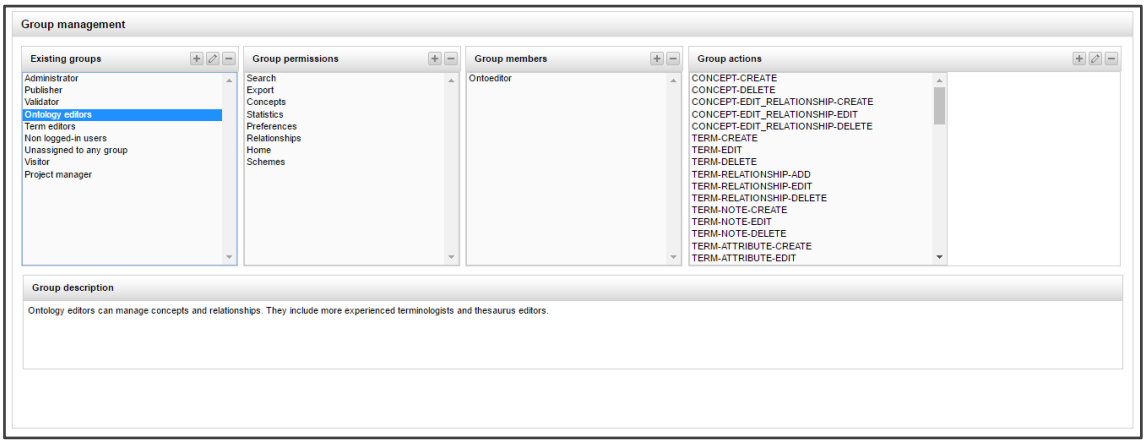

Figure 3: VocBench Group Management

The rights of each user group are defined explicitly by the administrator, and the users are subsequently and mandatorily assigned to one or more groups and one or more projects.

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).

WebProtégé[9] is a web-based ontology development environment based on the popular Protégé editor. As with VocBench, its web nature allows the provision of fairly extensive collaboration features. It also incorporates tracking and revision history functionality and multilinguality features. WebProtégé supports most of the functionalities of the classic desktop Protégé (and is also cross compatible with the latter), which culminates to a very powerful and complete system. However, it is primarily targeted to expert users with extensive knowledge of the OWL 2 specification and the overall RDF ecosystem.

[9] https://protegewiki.stanford.edu/wiki/WebProtege

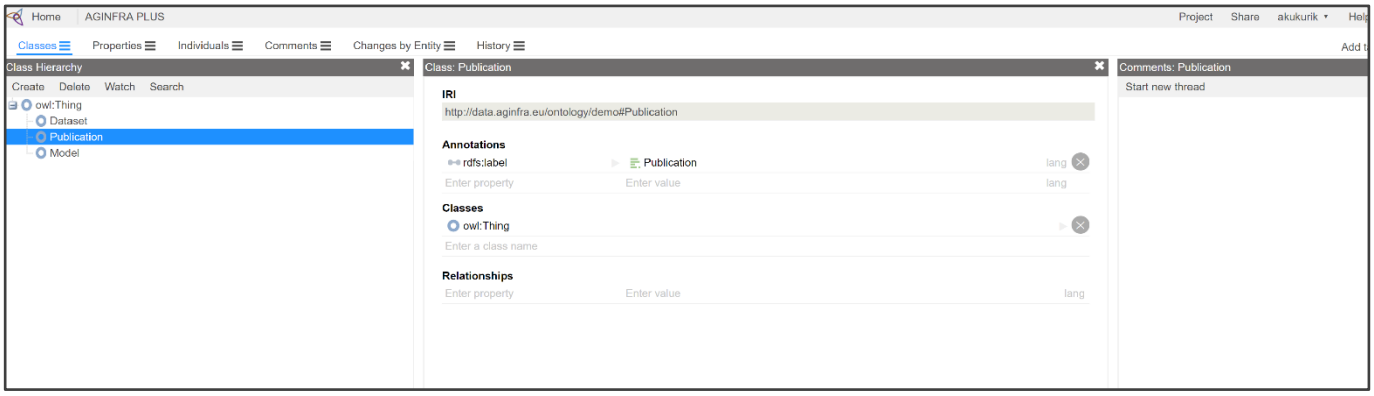

Figure 4: WebProtégé Ontology Creation

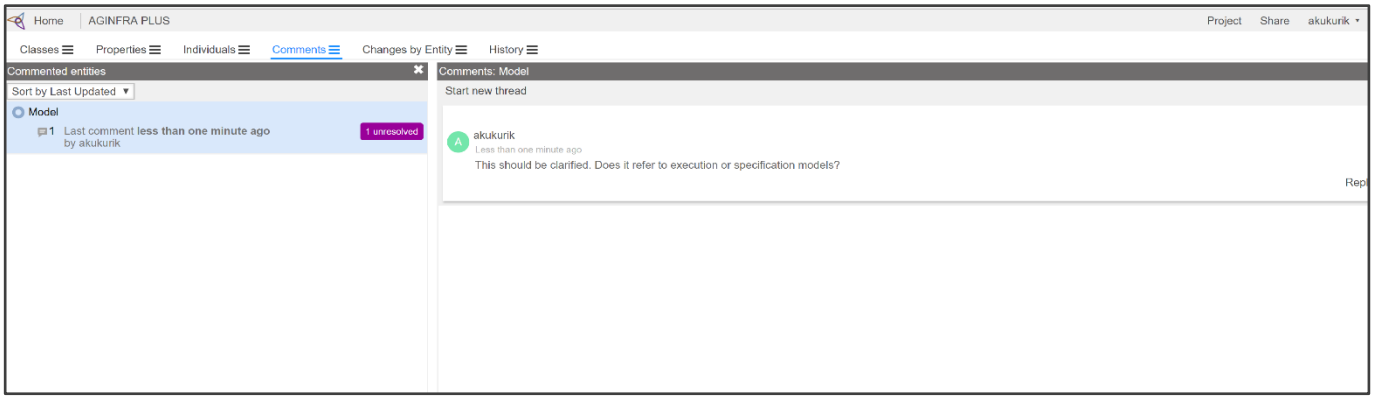

Figure 5: WebProtégé Comment Section

As WebProtégé adheres to the Semantic Web interpretation of what constitutes an ontology – as opposed to the knowledge modelling and philosophical views – it also allows the introduction of instances within the authoring environment, thus allowing the direct and explicit incorporation of data under the designed conceptualization.
From the technical point of view, WebProtégé also relies on some third-party frameworks for deployment and operation. Namely, it also runs on top of an Apache Tomcat container (also supporting alternative containers like GlassFish) and uses MongoDB[10] for preserving some configuration and execution parameters separately for its actual data store.

TemaTres[11] is an open-source web application for managing and publishing controlled vocabularies. As opposed to VocBench and WebProtégé, TemaTres does not natively rely on semantic web meta-models, but instead used a term-based approach for the representation of the vocabularies. It however support import and export functionalities in different formats, including SKOS. Regarding authorisation and access rights management, TemaTres supports pre-defined user roles (administrator, editor, guest, etc.).

4.2 DATA LINKING AND MAPPING

As described in section 2, the use cases examined in the context of AGINFRA+ pose a need for incorporating linking mechanisms at both the instantiation (specific data assets) and the conceptualisation level (schemas and ontologies used for describing the data assets).
For the first case, while there are multiple data linking frameworks proposed in the literature, they are mostly research prototypes with inadequate stability for incorporation in an operational framework, and consequent lack of support and maintenance.

[10] https://www.mongodb.com/

[11] http://www.vocabularyserver.com

The additional requirements on flexibility and relative ease of configuration and execution constitute the ***Silk*** framework[12] (Volz et al, 2009) a good candidate solution to serve as the baseline platform for building the AGINFRA PLUS data linking components.

As most frameworks, Silk support the generation of *owl:sameAs* links but allows the declaration of arbitrary RDF predicates to be used as the link relation between connected entities. Furthermore, its implementation in Scala and therefore the ability to be executed over a standard JVM suits itself to incorporation in the overall VRE infrastructure adopted by AGINFRA PLUS. The framework also provides a REST API for incorporating its functionalities within the overall workflows that will be designed to run over the VRE.

To determine link relations and link conditions, Silk uses its own XML-based formalism, the *Link Specification Language* (Silk-LSL). A Silk-LSL record defines access parameters for the involved dataset descriptions (that must be accessible via SPARQL endpoints), execution parameters for the Silk framework, link identification metrics to be activated and aggregation functions for combining these metrics. Regarding the metrics used by Silk, they are predominantly focused on string matching methodologies, however, the framework allows the incorporation of external modules implementing different metrics for comparing the entities within the compared datasets.

For the latter case, ontology and more broadly schema matching, there is extensive literature and a large number of systems of adequate maturity that produce good results on benchmarks and realistic applications. The majority of modern systems combine different methods for estimating the similarity between two concepts or properties belonging in different ontological specifications. Some notable systems that are available as open-source libraries are presented below.

One of the first examples of a composite matching system is ***COMA++*** (Do and Rahm, 2007), where matching is performed as a two-stage process: At the first stage, specific parts of the source schema are determined and compared in order to specify the most similar partitions from the other schema. This is essentially a light-weight matching process, e.g. based on the lexical similarity of the root elements of the partitions. At the second stage, the partition pairs deemed highly similar go through the whole matcher workflow and the complete alignment is ultimately obtained by merging the results of the process for each partition pair.

Falcon-AO (Hu and Qu, 2008) and its evolution (Jauro et al., 2014) perform a structure-based partitioning process to build ontology blocks and identifies similar blocks using pre-computed anchor elements, i.e. highly similar elements as determined by a string comparison method. The derived block pairs are subsequently fed to an alignment framework comprising two sequential matchers.

Anchor-Flood (Seddiqui and Aono, 2009) also relies on anchors, in this case pairs of concepts with the exact same name. It then proceeds to analyse the neighbourhood (i.e. subconcepts, superconcepts and siblings of the concepts belonging in the anchor) of each anchor for matching pairs, hence resulting in dynamically constructed segments of the compared ontologies as opposed to predetermined partitions.

The Lily system (Wang et al., 2011) operates on ontology subgraphs and defines reduction anchors, i.e. the system avoids comparisons in the neighbourhood of two classes with low or no similarity.

LogMap 2.0 ((Jimenez-Ruiz et al., 2011) (Jiménez-Ruiz et al., 2012)) uses an indexing and anchoring mechanism to discover a set of initial mappings before computing modules that reflect the meaning and context of the entities in these mappings. It then proceeds to the iterative mapping discovery and repair

[12] http://silkframework.org/

methodology based on propositional logic, as described in the initial version of the system (Jiménez-Ruiz and Grau, 2011).

AgreementMaker (Cruz et al., 2009) proposes a layered organisation of the available matchers, from lexical to structural similarity and combining the results produced by these layers at the final stage, taking into account restrictions and user preferences for the mappings ultimately produced.

YAM++ (Ngo et al., 2016) uses two distinct matchers to discover mappings between the input ontologies; an element-level and a structural-level matcher. The discovered candidate mappings are then evaluated and revised by a semantic matcher, responsible for the removal of inconsistencies and maintaining the consistency of the proposed alignment.

4.3 ONTOLOGY VISUALISATION AND EXPLORATION

In principle, the visualisation of ontologies falls under the graph visualisation field, since RDF representation are in essence directed graphs. There are multiple frameworks and libraries dedicated to graph visualisation. However, tools dedicated to the visual representation of RDF-based resources are more powerful in the sense that they are able to retrieve the semantics defined by the semantic resource and present this information to the end-user in conjunction with the visualisation of the graph corresponding to the ontology. Some such tools are the following.

SKOS Play[13] is a rendering and visualisation application for thesauri, taxonomies and controlled vocabularies expressed in SKOS and SKOS-XL.
SKOS Play exports the input SKOS data model as HTML or PDF documents and visualize them in graphical representations.

Skosmos[14] is an open source web based SKOS vocabulary browser that uses a SPARQL endpoint as its back-end. Skosmos provides a multilingual user interface for browsing and searching the data and for visualizing concept hierarchies. A REST API is also available providing access for using vocabularies in other applications such as annotation systems

NavigOwl[15] is a visualization tool which is specially designed to explore semantic networks. The Tool is enriched with graph layouts that can be applied over the semantic network in order to understand the structure of ontologies easily and it facilitates the user to build mental map in more clear and consistent view of ontology graph. The tool supports the visualisation ofRDF and OWLresources.

WebVOWL[16] is a web application for the interactive visualization of ontologies. It implements the Visual Notation for OWL Ontologies (VOWL) (Lohmann et al, 2016) by providing graphical depictions for elements of the Web Ontology Language (OWL) that are combined to a force-directed graph layout representing the ontology. Interaction techniques allow to explore the ontology and to customize the visualization. The VOWL visualizations are automatically generated from JSON files into which the ontologies need to be converted. A Java-based OWL2VOWL converter is provided along with WebVOWL and can be deployed with the visualisation library.

[13] http://labs.sparna.fr/skos-play/

[14] http://skosmos.org

[15] http://home.deib.polimi.it/hussain/navigowl/index.html

[16] http://vowl.visualdataweb.org/webvowl.html

4.4 SEMANTIC TRANSFORMATION / RDFIZATION

Given the abundance of conceptualisations and – more commonly –metadata descriptions that have been created in different formats like spreadsheets, XML files, CSV, etc., the provision of a robust transformation service for expressing such content in a Semantic Web language can facilitate and accelerate the transition to a semantically rich data space. Currently, there are multiple solutions for the transformation problem, with each usually dedicated to a specific input format[17]. Some exemplary tools and frameworks for converting traditional formats in RDF-based descriptions are the following.

The aforementioned ***SKOS Play*** suite offers a tool for converting Excel or Google Sheet documents into SKOS RDF. The documents to be converted have to follow a specific structure and column definitions.

Sheet2RDF[18]is a platform for acquisition and transformation of datasheets into RDF, developed by the ART Research Group at the University of Rome Tor Vergata.
Sheet2RDF is built on top of the CODA suite developed by the research group and provides functionality for the generation of RDF content modelled under any target RDF vocabulary.
Sheet2RDF comes both as command line tool and as an extension for the Semantic Turkey application server, providing a user interface for loading datasheets, generating a triplification rules file, projecting the datasheet content as RDF triples and finally adding them to the project loaded in the hosting application.

For the declaration of triplification rules, Sheet2RDF uses the PEARL[19] language, which enables users to describe matches over sets of annotations produced by UIMA Analysis Engines over streams of unstructured information, and to specify how the matched annotations will be transformed into RDF triples. PEARL combines the mechanism of UIMA features paths (to extract relevant information from UIMA annotations) with a subset of the SPARQL syntax to describe patterns for generating RDF triples.
The platform supports Microsoft Excel, Apache OpenOffice and LibreOffice spreadsheets, as well as CSV, TSV and other delimited formats.

XLWrap[20]is a spreadsheet-to-RDF wrapper which is capable of transforming spreadsheets to arbitrary RDF graphs based on a mapping specification. It supports Microsoft Excel and OpenDocument spreadsheets such as comma- (and tab-) separated value (CSV) files and it can load local files or download remote files via HTTP.

XLWrap is based on the application of mapping specifications, i.e. templates for parsing and subsequently converting cell values or combinations of cell values to RDF predicates. The templates are repeatedly applied to the input work sheet (or references to other work sheets) towards producing the target graph. The mappings are declared using a controlled vocabulary, the XLWrap mapping vocabulary[21].

OpenRefine[22] is a more generic and extensive web application, as it is not limited to RDF conversion. OpenRefine allows mass-editing and cleaning of tabular data, as well as, applying formatting and changing values via the declaration of the appropriate rules. It is supported by various

[17]https://www.w3.org/wiki/ConverterToRdf

[18]http://art.uniroma2.it/sheet2rdf/

[19]http://art.uniroma2.it/coda/documentation/pearl.jsf

[20]http://xlwrap.sourceforge.net

[21]http://purl.org/NET/xlwrap

[22]http://openrefine.org

extensions/plugins, including an RDFization extension. The latter allows the generation of RDF graphs via the creation of an “RDF Skeleton”, i.e. a rule base for transforming column names or cell values to the appropriate RDF predicates, using arbitrary schemas.

5 CONCLUSIONS AND NEXT STEPS

The present document reports on the functional and non-functional requirements posed by the specification of the AGINFRA+ use cases, as the latter are detailed in the deliverables of the relevant work packages. Furthermore, a set of non-functional requirements has been reported, based on the observations of the user communities' representatives.

Based on the elicited requirements, the specification for the data and semantic layer components and, additionally, the layer's architecture and its positioning within the overall AGINFRA+ environment has been designed. The specification aims to cover the requirements by using widely used open standards and components, respecting the open science and open knowledge principles and ensuring that the resulting platform will be maintainable and extensible in the long term.

While the present version of the data and semantics layer architecture is complete, it will be naturally refined and extended as work on the project progresses and further details and intricacies of each use case become apparent.

To this end, the specification report is to be treated as a live document, with updates applied whenever a major shift on the requirements is identified.

The report is the driver for evaluating and deploying the relevant technologies and tools under the specified abstract architecture and building the required communication mechanisms with specific VRE components and the VRE infrastructure in general.

As the tools are being used by the relevant communities, any extensions and modifications will be applied and reported in the respective WP2 deliverables (D2.2, D2.3).

REFERENCES

Cruz, I.F., Antonelli, F.P., Stroe, C., "AgreementMaker", Proceedings of the VLDB Endowment 2, 1586–1589,2009.

Do, H.-H., Rahm, E., "Matching large schemas: Approaches and evaluation", Information Systems 32, 857–885,2007.

Hu, W., Qu, Y., "Falcon-AO: A practical ontology matching system", Web Semantics: Science Services and Agents on the World Wide Web 6, 237–239, 2008

Hu,W., Chen, J., and Qu, Y., "A self-training approach for resolving object coreference on the semantic web," in 20th International World Wide Web Conference (WWW 2011), Hyderabad, India, 2011.

Jauro, F., Junaidu, S.B., Abdullahi, S.E., "Falcon-AO++: An Improved Ontology Alignment System", International Journal of Computer Applications 94, 1–7, 2014.

Jimenez-Ruiz, E., E., Grau, B.C., Zhou, Y. "LogMap 2.0: towards logic-based scalable and interactive ontology matching", Nature Proceedings, 2011.

Jiménez-Ruiz, E., Grau, B.C., "LogMap: Logic-Based and Scalable Ontology Matching" In: The Semantic Web ISWC 2011. Springer Berlin Heidelberg, pp. 273–288, 2011.

Jiménez-Ruiz, E., Grau, B.C., Zhou, Y., "LogMap 2.0", In: Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences - SWAT4LS 11. ACM Press, 2012.

Lohmann, S., Negru, S., Haag F., Ertl, T.,"Visualizing Ontologies with VOWL". Semantic Web 7(4), pp. 399-419, 2016

Ngo, D., Bellahsene, Z.,"Overview of YAM++ - (not) Yet Another Matcher for ontology alignment task", Journal of Web Semantics (JWS), Volume 41, pp. 30-49, December 2016

Ngomo, A.-C.N. and Auer, S., "LIMES - a time-efficient approach for large-scale link discovery on the web of data," in International Joint Conference on Artificial Intelligence, Barcelona, Spain, 2011.

Niu, X., Rong, S., Zhang Y., and Wang, H., "Zhishi.links results for OAEI 2011," in Ontology Matching Workshop at ISWC 2011, Bonn, Germany, 2011.

Raimond, Y., Sutton, C., and Sandler, M., "Automatic interlinking of music datasets on the Semantic Web," in Linking Data on the Web Workshop at WWW 2008 (LDOW 08), Beijing, China, 2008.

Seddiqui, M.H., Aono, M., "An efficient and scalable algorithm for segmented alignment of ontologies of arbitrary size" Web Semantics: Science Services and Agents on the World Wide Web 7, 344–356, 2009.

Volz, J., Bizer, C., Gaedke, M. and Kobilarov, G., "Discovering and maintaining links on the web of data," in 8th International Semantic Web Conference (ISWC 2009), Washington DC, USA, 2009

Wang, P., Zhou, Y., Xu, B., "Matching Large Ontologies Based on Reduction Anchors", In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Three, IJCAI'11. AAAI Press, pp. 2343–2348, 2011.