COMPETITIVENESS AND INNOVATION FRAMEWORK PROGRAMME

CIP-ICT-PSP-2013-7 Pilot Type B

WP3 – Service platform integration and deployment in cloud infrastructure

D3.3.3: Sematic tagging and open data publication tools

Deliverable Lead: PSNC

Deliverable due date: 28/02/2017

Actual submission date: 28/02/2017

Version: 2.0

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Commission under grant agreement no. 621074

Document Control Page

Title	D3.3.3 Semantic tagging and open data publication tools
Creator	Marcin Krystek (PSNC)
Description	This document introduces the current prototype of the semantic annotation service
Publisher	FOODIE Consortium
Contributors	Raul Palma (PSNC), Emilio Rubiera (CTIC), Małgorzata Wolniewicz (PSNC)
Creation date	23/09/2015
Type	Text
Language	en-GB
Rights	copyright “FOODIE Consortium”
Audience	☐ internal ☒ public ☐ restricted
Review status	☐ Draft ☒ WP leader accepted ☒ Technical Manager accepted ☒ Coordinator accepted
Action requested	☐ to be revised by Partners ☐ for approval by the WP leader ☐ for approval by the Technical Committee ☐ for approval by the Project Coordinator
Requested deadline	

Revision History

Version	Date	Modified by	Comments
0.1	01/02/2016	Raul Palma	Document creation
0.2	10/02/2016	Marcin Krystek	Update document structure
0.3	15/02/2016	Marcin Krystek	Added content to Sections 1-3
0.4	19/02/2016	Marcin Krystek	Added content to Sections 4-6
0.5	23/02/2016	Emilio Rubiera	Input to Section 2.3
1.0	25/02/2016	Raul Palma	Review and updates
1.2	26/02/2016	Miguel Ángel Esbrí (ATOS)	Q&A
1.3	01/02/2017	Raul Palma	Document creation
1.4	10/02/2017	Małgorzata Wolniewicz	Update content to Sections 2-3
1.5	14/02/2017	Raul Palma	Update content to Sections 1-3 and internal review
1.6	21/02/2017	Małgorzata Wolniewicz	Update content to Sections 2-4
1.7	23/02/2017	Raul Palma	Review and updates
1.8	25/02/2017	Raul Palma	Update content to Section 5 and 6
2.0	26/02/2017	Miguel Ángel Esbrí (ATOS)	Q&A

Note

This deliverable is subject to final acceptance by the European Commission.

Disclaimer

The views represented in this document only reflect the views of the authors and not the views of the European Union. The European Union is not liable for any use that may be made of the information contained in this document.

Furthermore, the information is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user of the information uses it at its sole risk and liability.

Table of Contents

Index of Figures

Index of Tables

Glossary

The glossary of terms used in this deliverable can be found in the public document “FOODIE_Glossary.pdf” available at: http://www.foodie-project.eu

Abbreviations and Acronyms

Abbreviation / Acronym	Description
API	Application Programming Interface
CPU	Central Processing Unit
DBA	Database Administrator
DBaaS	Database as a Service
DDD	Domain-Driven Design
DNS	Domain Name System
GTM	Global Transaction Manager
HA	High Availability
HDD	Hard Disk Drive
HTTP	Hypertext Transfer Protocol
IaaS	Infrastructure as a Service
IDE	Integrated Development Environment
JAR	Java ARchive
MPP	Massive Parallel Processing
OS	Operating System
POM	Project Object Model
RAM	Random Access Memory
RDBMS	Relational Database Management System
RDF	Resource Description Framework
SLA	Service Level Agreement
SQL	Structured Query Language
TCP	Transmission Control Protocol
TDD	Test – Driven Development
VM	Virtual Machine

Table 1: Abbreviations and Acronyms

Executive Summary

This document introduces the final version of the sematic tagging and open data publication system comprising: the semantic annotation service, the data semantization service, the data access and publishing Components, and the stor-age Components.

This is a technical document providing an overview of the components and their functionalities, the underlying models, as well as their implementation and deployment details.

Note that D3.3.3 is a self-contained document that supersedes previous versions (D3.3.1 and D3.3.2). Accordingly, the content from the previous documents has been reused and updated to reflect the latest developments related to this task.

1 Introduction

Semantic tagging and open data publication tools is a system of services which can be used to extract additional knowledge from unstructured data such as plain text or text files in different formats, transform (semi-) structured data into semantic format, and publish the generated data according the Linked Data principles.

In particular, semantic annotations enrich the source data with a context that is linked to some structured knowledge of a domain or application, which can then be exploited by different ap-plications and services. For instance, this knowledge is used to enrich the description of the re-sources presented to the user by the Market-place platform, as described in D3.5.2 .

Since the newly discovered knowledge is described by standard ontologies, stored in machine-readable format and accessible through standard APIs and protocols, it can also be used for further machine processing allowing better integration with existing knowledge bases and their publication in the Linked Open Data Cloud , discovering and understanding relations and de-pendencies between resources (e.g., Marketplace products), as well as the implementation of all other kind of user scenarios.

Similarly, semantic data generated through the transformation of (semi-)structured data sources, such as tabular data or XML, can then be exploited to discover additional knowledge through the integration with other data elements via semantic links.

2 Components Description

Semantic tagging and open data publication tools compose together a system in which the following components can be distinguished:

2.1 Semantic Annotation Service

The semantic annotation service is a core component of the system. It provides a simple REST API designed for creating, updating and retrieving semantic annotations. The service orchestrates other components (existing annotation tools described below) in order to control and fully perform data analysis process, creates semantic form of the generated data and persists semantic data using semantic store.

2.1.1 Annotation Tools

These are external tools used as libraries or services that allows keywords extraction, sentence disambiguation and text meaning identification. The results generated by these tools are then used to create semantic representation of the annotation. The current implementation of the semantic annotation service uses the following annotation tools:

2.1.1.1 AgroTagger

AgroTagger is a keyword extractor that uses the ARGOVOC thesaurus as its set of allowable keywords. It also identifies concepts from AGROVOC vocabulary in terms of Linked Open Data. AGROVOC is a controlled vocabulary covering all areas of interest of the Food and Agriculture Organization (FAO) of the United Nations, including food, nutrition, agriculture, fisheries, forestry, environment etc. It is published by FAO and edited by a community of experts.

2.1.1.2 Babelfy

Babelfy is a unified, multilingual, graph-based approach to Entity Linking and Word Sense Disambiguation based on a loose identification of candidate meanings coupled with a densest subgraph heuristic, which selects high-coherence semantic interpretations.

2.2 Data Semantization Service

The data semantization service is a tool implemented to carry out chiefly conversions from a heterogeneous range of data into RDF.

It is currently focused on the transformation of tabular data in different formats but in the future, it will be extended to transform other formats, such as XML, into RDF.

The component is an open source tool with a simple Web interface where users upload the file to be transformed and thanks to the mediation of a small lightweight piece of code that identifies the structure of the added data, it is capable of generating an output file fully compliant with the RDF family language.

The current version offers only a manual tool with no guidelines or best practices for the end user; however, in the next releases, it is planned that the conversion become more user-friendly and even show semi-automatic capabilities. Besides, the number of input format sup-ported will be increased.

Thanks to this tool, users will be able to transform their unstructured or semi-structured data into more friendly formats and pave the way to share this data in an easier way or simply get that data in a format fully compliant with the linked open data guidelines.

2.3 Data Access and Publishing Components

Semantic annotations and other semantic data generated and stored by the semantic services are accessible using the following frameworks and API's:

2.3.1 Silk

Silk is a linked data integration framework for integrating heterogeneous data sources. It al-lows generating links between related data items within different Linked Data sources and applying data transformations to structured data sources. Linked Data publishers can use Silk to set RDF links from their data sources to other data sources on the Web. Accordingly, in FOODIE, Silk is used to discover (and store) links between the semantic data in the se-mantic store and other key knowledge bases in the Linked Open Data Cloud.

2.3.2 Sparql

Sparql is as standard query language for RDF data. FOODIE provides access to RDF data through a SPARQL endpoint that is exposed by the semantic store.

2.3.3 Faced Search and Browser

Faced search and browser is based on Virtuoso Faceted Browser, which allows simple text search and results browsing. In FOODIE, we use this interface mainly to enable simple navigation of RDF data through their links.

2.3.4 Semantic Annotation Service API

Semantic Annotation Service provides dedicated API which allows annotation retrieval in the simplified JSON form. Currently this API is being used by the Marketplace portal, which showcases its usage through external applications.

2.4 Storage Components

The semantic annotation service processes and produces various types of data. In order to meet specificity of the data type as well as different requirements and constraints regarding data processing procedure, three different storage components are used.

2.4.1 Semantic Store

For storing semantic data, the OpenLink Virtuoso server is used. Virtuoso has built in support for SPARQL query language and inference engine. It also provides good performance, geo-spatial support and capabilities to generate linked data from different data sources types via Sponger. Sponger is also transparently integrated into Virtuoso's SPARQL Query Processor delivering URI de-referencing within SPARQL query patterns and is directly accessible via REST interfaces.

2.4.2 RDBMS

For storing simple relational data such as mappings, resource processing status and statistics the relational database management system is used. The current implementation of the semantic annotation service uses sqlite implementation as simple, lite and providing satisfying performance solution. Since sqlite provides standard SQL interface it can be easily replaced with other, more advanced solution if required.

2.4.3 File system

The semantic annotation service allows annotating unstructured files. All files send to the service are stored temporally in the file system. They are available only for the annotation procedure and will be removed after procedure successfully finishes.

3 System Design

Semantic tagging and open data publication tools compose together a system for which three main layers can be identified:

- **API layer** – implemented as set of REST services. Each service provides different functionality for the client application or final user. API's denotes overall system functionality.
- **Service layer** – it is set of different web services or libraries directly used by Semantic Annotation Service or operating on data generated by Semantic Annotation Service
- **Storage layer** – provides persistence for all data processed or generated by the Semantic Annotation Service.

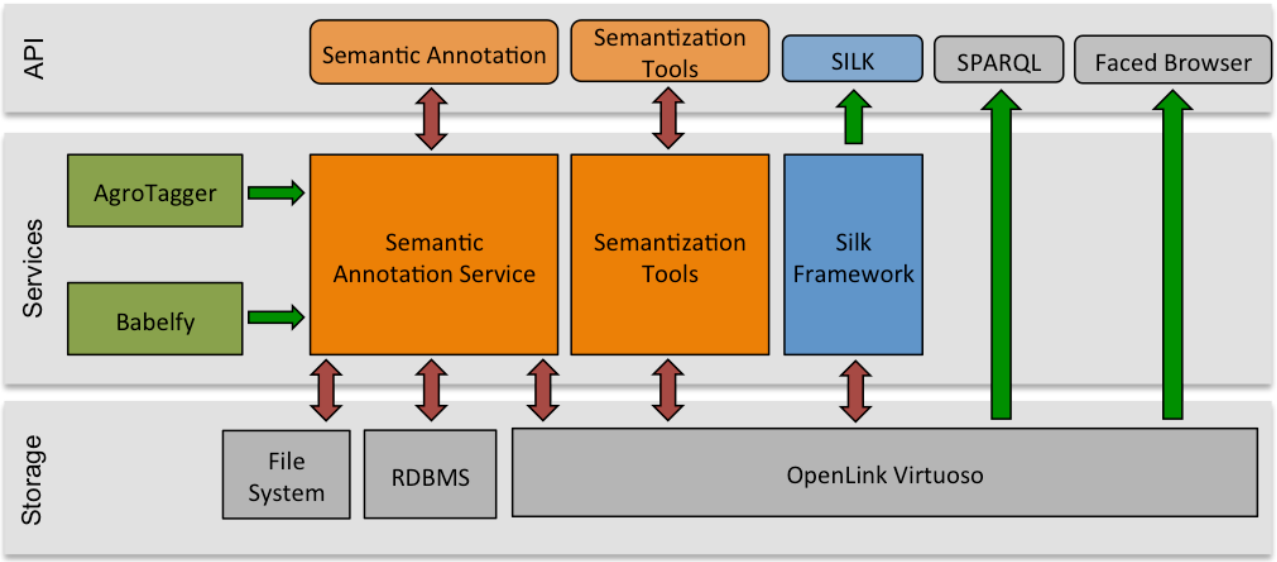

Figure 1 Sematic tagging and open data publication tools system design

Arrows on the Figure 1 denotes direction of the data flow direction.

3.1 Data Model for Annotations

Annotations created by the semantic annotation service are modelled using the Modular Unified Tagging Ontology (MUTO) which is designed specifically for tagging and folk-sonomies. MUTO allows representing public and private tagging, simple and auto generated tags and others. It is also easily extensible since all concepts defined in MUTO ontology inherits from other more general ontologies like SKOS , SIOC or vocabularies as RDFS .

The general structure of the MUTO ontology and main concepts are presented in the following diagram.

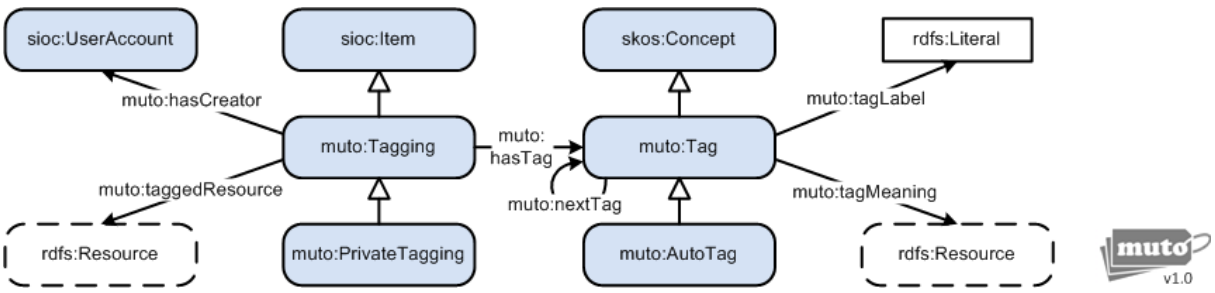

Figure 2 MUTO ontology general design

The central concept of the MUTO ontology is muto:Tagging. muto:Tagging instance is created as response for create

annotation request. For each tagging a tagged resource is also created. Tagged resource instance represents an object that is actually annotated. It can be anything (file, set of files, text description) that can be addressed or identified by the client. For each tagging instance there is a set of muto:Tag instances created. Each muto:Tag instance represents a single concept identified by the annotation tools. muto:Tag instance has its default label defined as muto:tagLabel. If the identified concept has its representation in the Linked Open Data Cloud, an additional skos:Concept instance is created and related with tag using muto:tagMeaning. This object can be also enriched with labels in different languages if such are available in LOD Cloud.

Tags identified by automatic annotation tools are created as muto:AutoTag instances. Tags de-fined by user manually are stored as muto:Tag instances. muto:PrivateTagging and sioc:UserAccount are currently not used.

4 Implementation Details

4.1 Semantic Annotation Service

4.1.1 REST API

The semantic annotation service REST API provides the following functionality:

Method	Description
POST /tagging	Creates new tagging for requested list of files and text description. Tagging will consist of set of tags (annotations) identified by an-notation tools. It also creates tagged resource instance. If resource instance URI is provided in the request, the method updates existing tagging.
GET /tagging	Retrieves tag labels for the resource and language specified in the request.
POST /tag	Adds new tag to the existing tagging
GET /statistics/meaning	Retrieves statistics showing which concepts are most commonly identified in the annotated resources.
GET /resources	Gets resources tagged by tags with specified meaning.

More technical description of the semantic annotation service REST API can be found in D3.2.3. Open API specification.

The tagging creation request may contain plain text description or set of files. The current implementation of the semantic annotation service supports processing of the following file types: PDF, DOC, XLS, HTML, TXT

4.1.2 Development and frameworks

- **Programming language**: The semantic annotation service is a Java web application. It requires at least Java 1.7 for development and for runtime environment.
- **Application framework**: The semantic annotation service uses Spring v.4.1.4 as main application framework, taking advantage of its bean management and event processing functionalities as well as support for REST API implementation.
- **Building tools**: The semantic annotation service is built using Apache Maven v.3.1.3. All project dependencies and deployment process are defined and managed by Maven.
- **Web container:** The building process creates a war file that can be deployed on any web container. Currently Apache Tomcat v.7.0 is used for the development and production deployment.
- **Source code location:** The source code of the application is available on git repository: https://git.man.poznan.pl/stash/projects/FOOD/repos/semantic-annotation/browse/SemanticAnnotationRestApi
- **Artifactory:** The dependencies that cannot be resolved using public repositories are defined in the following artifactory: http://maven.man.poznan.pl/repository/simple/foodie-releases-public

4.2 Data Semantization Service

The data semantization service comprises two components: Tabels Core and Tabels Web.

4.2.1 Tabels Core

Tabels Core has the capability to map from supported tabular formats to RDF documents; the mapping between them should be specified by a transformation program by the means of a DSL (domain specific language) developed with the only purpose of these mappings.

Tabels core was developed in Scala language and using sbt to support build definitions. The source code is available at: https://bitbucket.org/fundacionctic/tabels

4.2.2 Tabels Web

Tabels Web is a demonstrator for the interoperability capabilities of a group of technologies related to Linked Data. It integrates online tabels core, tapinos-ws, tapinos-js, undermaps and a bunch of linked data exploitation tools.

Tabels Web was developed in Grails. The source code is available at: https://bitbucket.org/fundacionctic/tabels

4.3 Silk Framework

Silk framework includes a user interface known as Silk workbench. The workbench is a web application that supports the user through the process of interlinking different data sources. It offers the following features:

- Enables the user to manage different sets of data sources, linking tasks and transfor-mation tasks.
- Offers a graphical editor that enables the user to create and edit linking tasks and trans-formation tasks.
- Supports the user to evaluate the links generated by the current link specification
- Allows users to create and edit a set of reference links used to evaluate the current link specification.

More information available at: http://silkframework.org/

5 Deployment

5.1 Service Location

- The semantic annotation service is available at: https://www.foodie-cloud.org/semanticAnnotation/
- The data semantization service is available at: http://prepro.fundacionctic.org/foodie/ (not yet deployed in FOODIE cloud)
- The SPARQL endpoint is available at: https://www.foodie-cloud.org/sparql
- Faced Browser is available at: http://www.foodie-cloud.org/fct/
- SILK user interface is available at: http://silk.foodie-cloud.org/workspace

5.2 Semantic Annotation Demo Client

A demo client for the semantic annotation service is available through FOODIE Swagger in-stance (where other FOODIE APIs are also available). Through Swagger interface, the user is able to introduce text descriptions and select files that he wants to annotate. The form values are then sent to the semantic annotation service, which conducts the tagging process. As a response, the user will receive an URI of the created resource. The user can use this resource URI to retrieve or update the existing tagging. The user can also retrieve information about the most popular tags, or all the resources tagged by a tag with specified meaning.

The Swagger interface for the semantic annotation service API is available at:

http://www.foodie-cloud.org/swagger_ui/?url=http://foodie-cloud-org/swagger_api/Semantic_Annotation_API.json

5.3 Silk Framework

SILK was used to discover links between FOODIE dataset and other common datasets in the linked open data (LOD) cloud, such as dbpedia, Eurostat, Eurovoc and others. The results for each alignment has been uploaded to the FOODIE instance of Virtuoso triple store database, into a separate graph for easy inspection. The following table summarises all the vocabularies to which FOODIE dataset (annotations of marketplace products) have been aligned, including statistics and location[1]:

Vocabulary/ Dataset	Target Entity	Number of target entities	Number of Links	Virtuoso graph
Eurostat	skos:concept (skos:prefLabel)	78086	373	http://marketplace.foodie-cloud.org/foodie-eurostat
dbpedia	dbo:Country (rdfs:label)	3361	19	http://marketplace.foodie-cloud.org/foodie-dbpedia
Eurovoc	skos:Concept (skos:prefLabel)	7104	108	http://marketplace.foodie-cloud.org/foodie-eurovoc
Emergel	skos:Concept (rdfs:label)	4380	8	http://marketplace.foodie-cloud.org/foodie-emergel

[1] The followig namespaces are used:
skos: http://www.w3.org/2004/02/skos/core#
dbo: http://dbpedia.org/ontology/
rdfs: http://www.w3.org/2000/01/rdf-schema#
geop: http://aims.fao.org/aos/geopolitical.owl#

Vocabulary/ Dataset	Target Entity	Number of target entities	Number of Links	Virtuoso graph
NALT	skos:Concept (skos:prefLabel)	67294	200	http://marketplace.foodie-cloud.org/foodie-nalt
Geopolitical ontology	geop:self_governing(geop: nameListEN)	207	14	http://marketplace.foodie-cloud.org/foodie-geo-political

6 Conclusions

This deliverable provided an overview of the final set of services and tools comprising the sematic tagging and open data publication system. In particular, this system comprises the following components:

- Semantic annotation service, which enables to create, update and retrieve semantic annotations associated to textual resources.
- Data semantization service, which enables to make transformation from heterogeneous range of data, particularly in tabular format, into RDF.
- Data access and publishing components, which enables to access the semantic data and pub-lish it according to linked data principles that includes the discovery of links with different datasets in the linked open data cloud.
- the Storage Components, which provides the underlying storage mechanisms for the semantic data, and other related data necessary for the execution of the different services.
- The semantic annotation service has been integrated in the marketplace (see D3.5.2) in order to generate automatically semantic annotations for the marketplace products. This allows the classification of products based on their content and description.

The data semantization service has been used to generate rdf data from tabular datasets, such as emergel and faostat.

The silk framework from the data access and publishing components has been used to discover links between FOODIE dataset (annotations of marketplace products) and other relevant vocabularies and datasets in the LOD cloud.

Virtuoso, which realises FOODIE semantic data store, stores and provides access to all the generated semantic data. It is accessible via a sparql endpoint and via a faceted search interface.

Overall, the set of tools and services developed and/or deployed as part of this task, allowed us to achieve the envisioned goals of the project regarding semantic tagging and open data publication.

References

[1] D3.5.2 Marketplace; Maciej Łabędzki, Patryk Promiński, Adam Rybicki

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

[3] https://github.com/fcproj/agrotagger

[4] http://babelfy.org/about

[5] http://silkframework.org/

[6] https://www.w3.org/TR/rdf-sparql-query/

[7] http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VirtFacetBrowserInstallConfig

[8] https://www.foodie-cloud.org/

[9] http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/

[10] https://www.sqlite.org/

[11] http://muto.socialtagging.org/core/v1.html

[12] https://www.w3.org/2004/02/skos/

[13] http://rdfs.org/sioc/spec/

[14] https://www.w3.org/TR/rdf-schema/