COMPETITIVINESS AND INNOVATION FRAMEWORK PROGRAMME

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

WP3 – Service platform integration and deployment in cloud infrastructure

D3.3.1: Sematic tagging and open data publication tools

Deliverable Lead: PSNC

Deliverable due date: 28/02/2015

Actual submission date: 08/04/2015

Version: 1.0

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	Deployment and integration report
Creator	Raul Palma (PSNC)
Description	This document introduces the first prototype of the semantic annotation service
Publisher	FOODIE Consortium
Contributors	Marek Jeszka (PSNC)
Creation date	28/02/2015
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	

About the project

FOODIE project aims at creating a platform hub on the cloud where spatial and non-spatial data related to agricultural sector is available for agri-food stakeholders groups and interoperable. It will offer: an infrastructure for the building of an interacting and collaborative network; the integration of existing open datasets related to agriculture; data publication and data linking of external agriculture data sources, providing specific and high-value applications and services for the support of planning and decision-making processes.

FOODIE project is addressed to four basic groups of users: a) stakeholders from the agriculture sector as end-users of final applications, b) public sector for communication with farmers about taxation, subsidies and regulation, c) researchers for large scale experimentation on real data and d) ICT companies for the development of new applications for agriculture and food sector, mainly using implemented tools

FOODIE specifically works on three pilots:

- Pilot 1: Precision Viticulture (Spain) will focus on the appropriate management of the inherent variability of crops,
- Pilot 2: Open Data for Strategic and Tactical Planning (Czech Republic) will focus on improving future management of agricultural companies (farms) by introducing new tools and management methods,
- Pilot 3: Technology allows integration of logistics via service providers and farm management including traceability (Germany).

Contact information

Miguel Angel Esbrí

Project Coordinator

Atos Spain, Madrid, Spain

E-mail: miguel.esbri@atos.net

URL: http://www.foodie-project.eu

Twitter: https://twitter.com/FOODIE_Project

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
API	Application Programming Interface

Table 1: Abbreviations and Acronyms

Executive Summary

This document introduces the first prototype of the semantic annotation service.

Components and related services overview

The semantic annotation service enables the (semi-) automatic annotation, or tagging, of un-structured or semi-structured data, enriching it with a context linked to the structured knowledge of our domain, specified in this first stage by AGROVOC vocabulary.

This process can be described as the production of semantic metadata for text and other sources (e.g., documents), by identifying mentions of known concepts and instances from AGROVOC vocabulary. Such process in turn enables a (semi-) automatic population of FOODIE semantic repository with the instance data extracted from source documents.

Hence the implementation of the semantic annotation service involves two components: the semantic repository and the annotation service itself.

Semantic repository

The repository is based on OpenLink Virtuoso (open source edition of Virtuoso Universal Server) and it is deployed as part of the DbaaS component. We have selected Virtuoso as base technology because of its scalability, performance, geospatial support and capabilities to generate linked data from different data sources types via Sponger. Also, in addition to support SPARQL query language, Virtuoso implements a RESTful API for posting to its RDF triplestore, something that may be also useful for developers. 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.

Virtuoso is installed and configured in a single node, particularly given that open source edition does not support clustering capabilities. In the future, we may consider moving to the commercial edition with cluster capabilities if deemed necessary.

Virtuoso SPARQL endpoint is available via http://foodie-vm3.man.poznan.pl/sparql

Implementation

The annotation service implements a REST API for creation of semantic annotations. Initially annotations are created with use of AgroTagger[1]. GATE[2] or others will be utilized in future.

The semantic annotation API provides:

- possibility to annotate text and files
- creation of annotation in RDF format
- storage of RDF triples
- storage of annotated files
- access to created annotations and stored files

The component is deployed on Apache Tomcat server and is available at: http://foodie-vm3.man.poznan.pl/semanticAnnotation/

The component is built into a WAR package so it should be possible to deploy it on a different server.

API details

The API details are available in D3.2.1. Open API specification [3].

File storage details

Currently the input files (that are annotated) are stored locally, on the node where the Tomcat server is installed (/home/semantics/files). In future this can be easily changed to utilize different storage option (e.g. some cloud solution like OpenStack Object Storage).

It has to be reconsidered how to store files. At this stage files are saved by their original names, but this can cause problems, when more than one file with the same name will be uploaded. For example id of the file could be used.

RDF storage details

Virtuoso is used for the storage of RDF data generated. There is a SPARQL endpoint that can be used to query inserted data. Query can be used to check what kind of triples are stored by the component:

WITH <http://foodie-vm3.man.poznan.pl/semanticAnnotation> select ?s ?p ?o WHERE {?s ?p ?o }

Access to Virtuoso is restricted by mean of a username and password. Before starting application it has to be set as Java properties (virtuoso.username, virtuoso.password). For Tomcat application server it can be easily added to startup.sh file.

AgroTagger

Currently the semantic annotation component is using AgroTagger. Consequently there is need to provide some files, that have to be stored in the main folder of application server. Files are available to download from AgroTagger webpage. Component requires this type of structure:

```
-- data
  |-- models
  |   `-- fao30
  |-- stopwords
  |   `-- stopwords_en.txt
  `-- vocabularies
     |-- agrovoc_en.rdf
     |-- agrovoc_en.rdf.gz
     `-- agrovocURILabelMappings.txt
```

AgroTagger uses vocabulary for searching and matching annotations. At this moment only English items are provided (agrovoc_en.rdf). These entities are coming from AGROVOC vocabulary, but only items of type skos:concept are used.

File parsing

Currently component is working with these types of files:

- PDF
- DOC
- XLS

- HTML

Development process

The component is built using Maven. Building and deploying is simply performed with: mvn tomcat7:redeploy (in case Tomcat is used as a application server, different command would have to be used in case of different application server e.g. JBoss). File pom.xml has been properly set to deploy application to foodie-vm3 node with use of Maven Tomcat plugin.

Source code can be found on Stash (https://git.man.poznan.pl/stash/projects/FOOD/repos/semantic-annotation/browse/SemanticAnnotationRestApi).

Dependencies required by AgroTagger and Virtuoso are stored on Artifactory (http://maven.man.poznan.pl/repository/simple/foodie-releases-public/).

Logs of application are stored in logs/SemanticsLog.log (currently it is /home/semantics/apache-tomcat/logs/SemanticsLog.log).

Demo client

A demo client web page application is available at: http://foodie-vm3.man.poznan.pl/semanticAnnotation/index.html.

Figure 1 provides a screenshot of this demo application. In particular, it enables users to input text directly on the web page or to select the file they want to process (in the supported formats). Either case, after the user hits process, the semantic annotation service identifies mentions of concepts and instances of AGROVOC vocabulary in the source, and presents them in JSON format back to the user (see Figure 2). In the case that the user uploaded a file, this process also creates the corresponding triples in Virtuoso and returns the file id. Also from this simple interface, the user can get the details of a particular source file (filename, associated annotations) providing the id parameter, and they can download the original file.

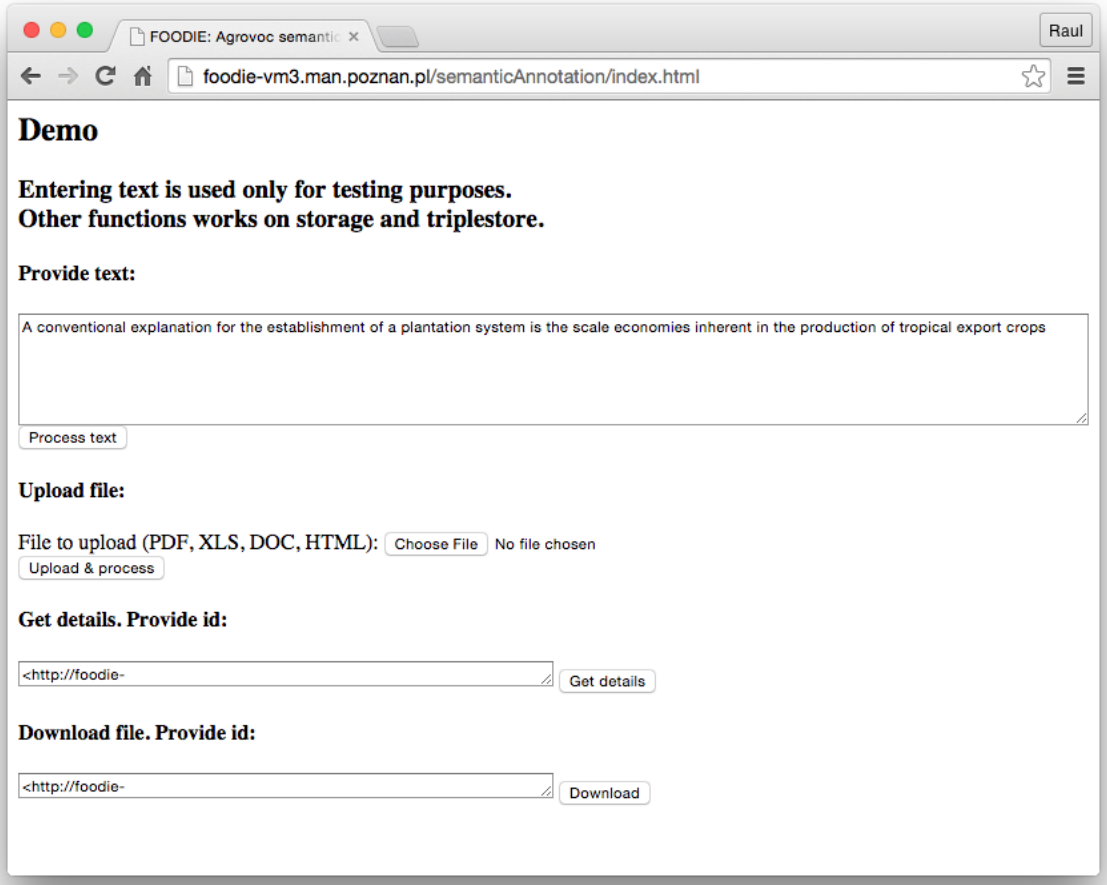

Figure 1 Demo client for Semantic Annotation Service

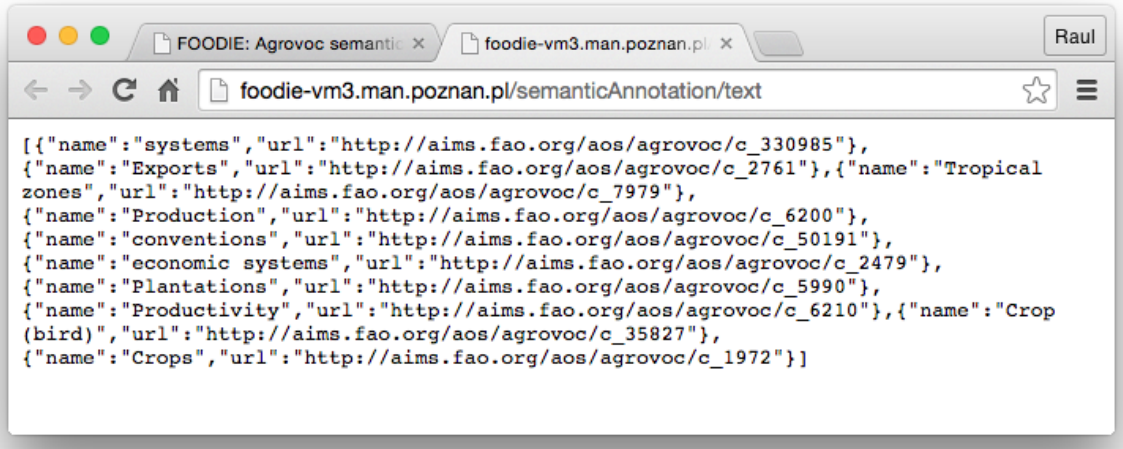

Figure 2 Identified vocabulary terms in the source

References

	References
01	https://github.com/agrisfao/agrotagger
02	https://gate.ac.uk/
03	Campos A., et al "Open and Lightweight APIs". D3.2.1 FOODIE Deliverable. March 2015

Table 2: References