Ecological Informatics 26 (2015) 119–126

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Semantic annotation of the CEREALAB database by the AGROVOC linked dataset☆

Domenico Beneventano, Sonia Bergamaschi, Serena Sorrentino, Maurizio Vincini, Fabio Benedetti

Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Vignolese 905, 41125 Modena, Italy

ARTICLE INFO

Article history:
Received 3 January 2014
Received in revised form 11 July 2014
Accepted 15 July 2014
Available online 28 July 2014

Keywords:
Linked Open Data
Semantic annotation
Semantic mapping discovery
Cereals genotypic and phenotypic data
Agricultural and plant science thesaurus

ABSTRACT

Nowadays, there has been an increment of open data government initiatives promoting the idea that particular data should be freely published. However, the great majority of these resources is published in an unstructured format and is typically accessed only by closed communities. Starting from these considerations, in a previous work related to a dataset on young workers on non permanent contracts, we proposed an experimental and preliminary methodology for facilitating resource providers in publishing public data into the Linked Open Data (LOD) cloud, and for helping consumers (companies and citizens) in efficiently accessing and querying them.
Linked Open Data play a central role for accessing and analyzing the rapidly growing pool of life science data and, as discussed in recent meetings, it is important for data source providers themselves making their resources available as Linked Open Data.
In this paper we extend and apply our methodology to the agricultural domain, i.e. to the CEREALAB database, created to store both genotypic and phenotypic data and specifically designed for plant breeding, in order to provide its publication into the LOD cloud.

1. Introduction

Nowadays, the availability of freely accessible information on the Web is constantly growing and numerous open data sources are available on public organization's web sites. However, the great majority of these resources is published in an unstructured format and is typically accessed only by closed communities; indeed, even if freely available on the Web, there are no connections among them and their structural and semantic heterogeneity makes it difficult to perform automatic or semi-automatic cross-data analysis, thus preventing to obtain high value information. The Linked Open Data (LOD) paradigm represents the key solution to improve and enrich the use of open data and to help consumers (citizens and companies) to access their integrated information. In the Semantic Web research area, the term linked data refers to a set of best practices for publishing and connecting structured data on the Web (Bizer et al., 2009). LOD extends the linked data paradigm by publishing data which are freely available to everyone and for any purpose.

Also in the plant science field, Semantic Web technologies play a central role for accessing and analyzing the rapidly growing pool of plant genomic and phenomic data, as pointed out in (Walls et al., 2012). As discussed in a recent meeting in the life sciences (Katayama, 2013), the next step in the Semantic Web evolution involves the source providers themselves making their resources available as Linked Open Data.

At the 2012 G-8 Summit, G-8 leaders committed to the New Alliance for Food Security and Nutrition, the next phase of a shared commitment to achieving global food security. As part of this commitment, they agreed to share relevant agricultural data available from G-8 countries with African partners and convene an international conference on Open Data for Agriculture, to develop options for the establishment of a global platform to make reliable agricultural and related information available to African farmers, researchers and policymakers, taking into account existing agricultural data systems. On April 29–30 2013, the G-8 International Conference on Open Data for Agriculture[1] brought together open data and agriculture experts along with U.S. Agriculture Secretary Tom Vilsack, U.S. Chief Technology Officer Todd Park, and World Bank Vice President for Sustainable Development Rachel Kyte to explore more opportunities for open data and knowledge sharing that can help farmers and governments in Africa and around the globe protect their crops from pests and extreme weather, increase their yields, monitor water supplies, and anticipate planting seasons that are shifting with climate change. US Agriculture Secretary Tom Vilsack

☆ This work is an extended and revisited version of the paper: Beneventano, D., Bergamaschi, S., and Sorrentino, S., "Semantic annotation of the CEREALAB database by the AGROVOC linked dataset", presented at the *AEIDSS 2013 1st International Workshop on Agricultural and Environmental Information and Decision Support Systems, in conjunction with the 2013 International Conference on Computational Science and its Applications (ICCSA 2013), June 24–27, 2013, in Ho Chi Minh City, Vietnam.*

E-mail addresses: domenico.beneventano@unimore.it (D. Beneventano), sonia.bergamaschi@unimore.it (S. Bergamaschi), serena.sorrentino@unimore.it (S. Sorrentino), maurizio.vincini@unimore.it (M. Vincini), fabio.benedetti@unimore.it (F. Benedetti).

[1] https://sites.google.com/site/g8opendataconference/home.

http://dx.doi.org/10.1016/j.ecoinf.2014.07.002
1574-9541/

says data is among the most important commodities in agriculture - and sharing it openly increases its value.

Nevertheless, providing a standard way to represent and query open data is not enough; even if the LOD community is constantly growing, there are still a few applications making use of its data sets. In Jain et al. (2010b), the authors argued that the LOD cloud, in its current form, is only of limited value for furthering the Semantic Web vision: in order to efficiently use LOD data sets, consumers (users and applications) need to deeply understand the semantics of source schemas. Starting from these considerations, in Sorrentino et al. (2013) we proposed an experimental and preliminary methodology for publishing, linking and semantically enriching open data by performing automatic semantic annotation of schema elements. In this paper, we extend and apply this methodology to the CEREALAB database, a public relational database created to store both genotypic and phenotypic data and specifically designed for plant breeding. More specifically, this paper presents an approach to annotate the CEREALAB database and to publish it in the Linking Open Data network. The proposed approach is shown by using the AGROVOC linked dataset, both to annotate the CEREALAB schema and to discover schema-level mappings among the CEREALAB dataset and other resources of the Linking Open Data network. Although it has been developed for the CEREALAB database needs, the principles of generality applied for its design will enable any other community interested in publishing dataset in the Linking Open Data network.

The paper is organized as follows. In the rest of this introduction we will give an overview of the CEREALAB database (Section 1.1) and a brief description of the AGROVOC linked dataset and of other biological ontologies and thesaurus (Section 1.2). The proposed methodology for Annotation and Publication of Linked Open Data is introduced in Section 2. The methodology is then applied to the CEREALAB database: Semantic Annotation of the CEREALAB database by the AGROVOC linked dataset is illustrated in Section 3 and Semantic mappings discovery is discussed in Section 4. Section 5 analyzes related work. Finally, in Section 6, we give our concluding remarks and describe future work.

1.1. The CEREALAB database

The CEREALAB database (Milc et al., 2011) is a tool realized for marker-assisted selection (MAS) of wheat, barley and rice. It helps cereal breeders for practical MAS, e.g. for choosing molecular markers associated to economically important phenotypic traits. The development of the CEREALAB database was one of the objectives of the CEREALAB projects and of the BIOGEST–SITEIA laboratory[2] funded by Emilia-Romagna regional government (Italy), aiming to increase the competitiveness of regional seed companies through the use of modern selection technologies. The CEREALAB database contains both phenotypic and genotypic data, obtained from the integration of available open source databases with the data obtained by the genotyping activity of the CEREALAB project. As far as genotypic data are concerned, the Gramene[3] database (Liang et al., 2008) was used as a data source for rice, and GrainGenes[4] (Carollo et al., 2005) for wheat and barley. Among the selected datasets of phenotypic data (the complete list is in (Milc et al., 2011)) the Germplasm Resources Information Network[5] data for wheat and barley descriptors were used.

As an example of the CEREALAB relational database, Fig. 1 shows some instances of the table `GERMPLASM`, with the following attributes:

- `GPN` (`GermPlasm Name`): the name of a variety;
- `FHB`: the resistance of the germplasm to the FHB disease (Fusarium Head Blight);

GPN	Yield	FHB	FrostDamage	Type
Eureka	12	MR	Least Susceptible	cultivar
Fortuna	11	MR	Moderately Susceptible	landrace
Mentana	20	S	Most Susceptible	line
Kenora	20	MR	Moderately Susceptible	landrace
Oasis	21	MR	Moderately Susceptible	cultivar

Fig. 1. Instance of the table `GERMPLASM` (S = susceptible, MR = moderately resistant and R = resistant).

- `Frost Damage`: the susceptibility of the germplasm to freezing injury;
- `Type`: the variety's type;
- `Yield` : the grain yield expressed in tons/hectare.

Data integration is obtained by using the MOMIS (Mediator envirOnment for Multiple Information Sources) Data Integration system (Beneventano et al., 2000, 2001). MOMIS is characterized by a classical wrapper/mediator architecture: the local data sources contain the real data, while a Global Schema (GS) provides a reconciled and integrated view of the underlying sources. The GS and the mappings between the GS and the local sources are semi-automatically defined at design time by the Integration Designer component of the system (Bergamaschi et al., 2011a). After GS creation end-users can pose queries over the GS in a transparent way w.r.t. the local sources. An open source version of the MOMIS system is delivered and maintained by the academic spin-off DataRiver.[6]

1.2. Agricultural and plant science thesaurus

In this section we briefly analyze some ontologies and thesaurus developed in the agricultural and plant science domain, focusing our attention on AGROVOC which we will use in the paper.

Ontologies are essential tools for accessing and analyzing the rapidly growing pool of plant genomic and phenomic data, since they offer a flexible framework for comparative plant biology, based on common botanical understanding (Walls et al., 2012); in particular, the authors highlighted that as genomic and phenomic data become available for more species, the annotation of data with ontology terms will become less centralized, while at the same time, the need for cross-species queries will become more common, causing more researchers in plant science to turn to ontologies. As stated in Walls et al. (2012), one of the most relevant ontologies for the plant sciences (excluding specialized ontologies used by crop breeders and agronomists) is the plant ontology,[7] which covers gross plant anatomy and morphology at the level of the cell and higher, as well as plant development stages.

AGROVOC[8] is a multilingual structured thesaurus created by FAO and the Commission of the European Communities since 1980 covering the fields of food, agriculture, forestry, fisheries, and other related domains. The traditional AGROVOC thesaurus consists of words or expressions (terms) in multiple languages and organized using equivalence (USE/UF), broader term (BT), narrower term (NT), and related term (RT) relationships. Fig. 2 shows an excerpt of the AGROVOC semantic network, for instance, for the term “Lodging” we can easily derive that it is a kind of “plant damage” (as they are connected by a BT relationship) and that it is related to other terms such as “Lodging resistance” and “Rain”. Moreover, AGROVOC provides term definitions in different languages. This property is particularly relevant as, starting from the annotations, we can automatically enrich the CEREALAB schema with the natural language descriptions of its elements thus helping even not skilled users to understand the meaning of specific agricultural terms.

AGROVOC Concept Scheme (Rajbhandari and Keizer, 2012) is a semantically richer version of AGROVOC, which in addition to “terms”

[2] www.biogest-siteia.unimore.it.
[3] http://www.gramene.org.
[4] http://www.graingenes.org.
[5] http://www.ars-grin.gov.
[6] http://www.datariver.it.
[7] http://www.plantontology.org/.
[8] http://aims.fao.org/standards/agrovoc/.