Preliminary Work Towards Publishing Vocabularies for Germplasm and Soil Data as Linked Data

Valeria Pesce[1], Guntram Geser[2], Caterina Caracciolo[1], Johannes Keizer[1], and Giovanni L'Abate[3]

[1] Food and Agriculture Organization of the United Nations, Rome, Italy
{Valeria.Pesce, Caterina.Caracciolo, Johannes.Keizer}@fao.org
[2] Salzburg Research, Salzburg, Austria
guntram.geser@salzburgresearch.at
[3]Consiglio per la Ricerca e la sperimentazione in Agricoltura, Centro di Ricerca per l'AgroBiologia e la Pedologia (CRA-ABP), Firenze, Italy
giovanni.labate@entecra.it

Abstract. The agINFRA project focuses on the production of interoperable data in agriculture, starting from the vocabularies and Knowledge Organization Systems (KOSs) used to describe and classify them. In this paper we report on our first steps in the direction of publishing agricultural Linked Open Data (LOD), focusing in particular on germplasm data and soil data, which are still widely missing from the LOD landscape, seemingly because information managers in this field are still not very familiar with LOD practices.

Keywords: Agriculture, germplasm, soil, Knowledge Organization Systems, metadata sets, vocabularies, RDF, Linked Data, classifications

1 Introduction

agINFRA (www.aginfra.eu) is a co-funded FP7 programme aiming to provide tools and methodologies for creating large networks of agricultural data using grid- and cloud-based technology. One of the outputs of the project will be the publication of the data managed by project partners as Linked Open Data (LOD), to achieve full data interoperability. Project partners contribute various "data sets"[1], all of which come with some sort of metadata associated, for the purpose of correct data storage and retrieval.

Two fundamental things play a crucial role in data interoperability:

a) The metadata elements needed to describe each individual piece of information in the data sets, and

[1] There is no agreed definition of what a "data set" is. For this paper, a broad definition should be assumed: see the definition by the W3C Government Linked Data Working Group: A collection of data, published or curated by a single source, and available for access or download in one or more formats". http://www.w3.org/TR/vocab-dcat/#class--dataset

b) The sets of values for (some of) the metadata elements of above, usually called "controlled vocabularies", or "authority data".

While the former are often referred to as metadata sets, metadata element sets or vocabularies, the latter are often called controlled vocabularies, authority data, value vocabularies or Knowledge Organization Systems (KOSs). However, they both are commonly referred to as "vocabularies" (cf. [1], [2]). In this paper, we say vocabularies and KOSs, respectively.

Independently of the terminology adopted, both types of vocabularies are crucial, and necessary to understand the data. Therefore the first step planned in the agINFRA project was the identification and publication (when necessary) of the vocabularies and KOSs used by the various data owners to describe and classify their data.

The types of resources covered by agINFRA are: bibliographic resources, educational resources, germplasm data, soil data. The analysis performed within the project revealed that these types of resources have different features. On the one hand, bibliographic and educational resources are described by rather homogeneous metadata sets, and also the use of KOSs is rather consistent. On the other hand, for germplasm and soil information, there is quite an important amount of data available, but so far little work has been devoted for its inclusion in the Linked Data cloud. This is why in this paper we focus on germplasm and soil data.

2 Vocabularies and KOSs for Germplasm Data and Soil Data

By "germplasm" one means the collection of genetic resources for an organism. In the case of plants, it can be a seed or any plant part that can be turned into a whole plant. Germaplasm is collected both to develop new hybrids/varieties/cultivars, and for conservation purposes. In all cases, various pieces of information need to be stored, including the taxonomic name adopted for it and the authority for the species name, its common/commercial names when existing, identifier for the germplasm and for the institution collecting it, the geographical area of origin (e.g. latitude, longitude, altitude), and of course the date of acquisition. Often, the data set also keep track of pedigree, phenotype, chromosomal constitution, breeding institution, biological status of accession, the type of germplasm storage and so on.

As for "soil", there is a wide variety of definitions and interpretations of it, and each database on soil will store different information depending on the type of perspective adopted. For example, agronomists, environmental researchers, geologist, engineers, or water experts typically use different notions of characterizing depths, history, chemical composition and morphological aspects and classifications, as well as sampling methodologies and geographical reference systems. For this reason it is especially important that metadata standards for all these aspects are established and used, and that the possibility for their integration is carefully explored and exploited. For both germplasm and soil data, some metadata standards and KOSs already exist, but few data sets already use them.

Germplasm Data. The set of Multi-crop Passport Descriptors (MCPD) is widely used for information exchange among crop conservation and research institutions worldwide. Its first version (V.1) dates back to 2006, while V.2 was published in 2012 [3]. MCPD is also used by the national germplasm inventories in Europe to provide information to the EURISCO catalogue[2] (with six additional descriptors for the specific purposes of EURISCO). The EURISCO catalogue also includes the germplasm collections of the Italian Agricultural Research Council (CRA). The Crop Germplasm Research Information System (CGRIS)[3] of the Chinese Academy of Agricultural Sciences (CAAS) uses its own set of passport descriptors which represents the de facto standard in China and will be mapped to the MCPD.

Importantly, the MCPD does not include descriptors for Characterization and Evaluation (C&E) measurements of plant traits/scores, which is the most important information for plant researchers and breeders. An initial set of C&E descriptors [4] for the utilization of 22 crops have been developed by Bioversity International[4] together with CGIAR and other research centers. C&E measurement data determine the values of germaplasm, such as resistance to specific pathotypes, grain yield, and protein content. Therefore, they are critical for selecting relevant germplasm. However, as assessed by the EPGRIS3 project, C&E data is difficult to standardize and integrate in central databases [5]. A major recent achievement therefore is the Darwin Core extension for genebanks (DwC-germplasm) which is represented in RDF/SKOS. The extension has been derived from the MCPD standard and includes basic descriptors for C&E measurements [6] as suggested by EPGRIS3.

The traditional wealth of checklists of plant names and taxonomies is recently being further developed into the form of ontologies. See for example the Plant Ontology, explicitly referenced in the DwC-germplasm vocabulary, the Trait Ontology and the Phenotypic Quality Ontology. They all provide important controlled vocabularies for the domain at hand.

Soil Data. Most of soil-related data is still stored in databases, the description of which is often called "metadata", for example by the U.S. National Soil Information System (NASIS) [7]. The international Working Group on Soil Information Standards[5] (WG-SIS) aim to develop, promote and maintain internationally recognized and adopted standards for the exchange and collation of consistent harmonized soils data and information worldwide. Widely used metadata standards for soil are ISO 19115 and ISO 19119, which covers geographic information and services, and it is applied to catalog and fully describe datasets, including individual geographic features and feature properties. ISO 19139 provides the XML schema implementation, including the extensions for imagery and gridded data. Users of the Content Standard for Digital

[2] http://eurisco.ecpgr.org/
[3] http://icgr.caas.net.cn/cgris_english.html
[4] http://www.bioversityinternational.org/
[5] http://www.soilinformationstandards.org/

Geospatial Metadata[6] (CSDGM) have been recommended by the U.S. Federal Geographic Data Committee (FGDC) transitioning to the ISO standards [8].

The main international KOSs for talking about soil are the Soil Taxonomy [9] and the World Reference Base for Soil Resources [10]. An important recent achievement is the Multilingual Soil Thesaurus[7] (SoilThes) that has been developed in the eContentplus project GS SOIL [11]. SoilThes was created as an extension of the General Multilingual Environmental Thesaurus (GEMET)[8] and contains the concepts of the World Reference Base, the soil vocabulary of ISO 11074[9] and additional soil-specific concepts. GEMET is the official thesaurus for the Infrastructure for Spatial Information in the European Comminity (INSPIRE) directive[10], within which draft guidelines for data specification on soil are under development [12]. Another ISO Standard related to soil data is ISO 28258: Soil quality Digital exchange of soil-related data. In relation to XML schema implementation, the Centre for Geospatial Science in the University of Nottingham has developed SoTerML [13] (Soil and Terrain Markup Language), a markup language to be used to store and exchange soil and terrain related data. SoTerML extends of GeoSciML for SOTER model compliant with ISO/TC190/SC 1 N140 "Recording and Exchange of Soil-Related Data". SoTerML development is being done within the e-SOTER Platform. GEOSS plans a global Earth Observation System and, within this framework, the e-SOTER project[11] addresses the need for a global soil and terrain database.

A recent initiative to harmonize different Soil schemas is the Soil-ML project [14], a soil equivalent of the Geoscience Markup Language (GeoSciML) Definitions for application schema "ISO 28258 Definitions".[12]

3 Vocabularies for Germplasm and Soil Data as Linked Data

For germplasm-related vocabularies, some of the most relevant work has been mentioned above: the Darwin Core extension for genebanks (DwC-germplasm) is already represented in RDF/SKOS. A lot of activity around semantic technologies is also going on around the major plant /trait /gene ontologies, the Plant Ontology (explicitly referenced in the DwC-germplasm), the Gene Ontology, the Trait Ontology [15] and the Phenotypic Quality Ontology [16]. They give an overview of the interlinking between these ontologies and their availability as OWL and as web services.

A very interesting project is the iPlant Semantic Web Program[13], focused on “next-generation” data and service integration: the program has implemented the SSWAP

---

6 http://www.fgdc.gov/metadata/csdgm/

7 https://secure.umweltbundesamt.at/soil/en/about.html

8 GEMET, http://www.eionet.europa.eu/gemet/

9 ISO 11074:2005 Soil quality - Vocabulary, http://www.iso.org/iso/catalogue_detail.htm?csnumber=38529

10 INSPIRE: http://inspire.jrc.ec.europa.eu/

11 http://www.isric.org/specification/SoTerML.xsd

12 See http://schema.isric.org/sml/4.0/UML_Model/Soil%20Overview.pdf for a graphical overview of the schema.

service[14], based on the SSWAP protocol[15]. Three major information resources (Gramene, SoyBase and the Legume Information System) use SSWAP to semantically describe selected data and web services. Moreover, the Gene Ontology and Plant Ontology will be soon incorporated into SoyBase: "This will further facilitate cross-species genetic and genomic comparisons by providing another level of semantic equivalence between taxa." [16]

As far as soil-related vocabularies are concerned, GEMET has been published as SKOS and mapped to AGROVOC; also SoilThes has been published as SKOS and is linked to GEMET. For the spatial aspect, soil data can rely on many advanced RDF standards, mainly in the framework of the EU INSPIRE Directive.

The methodology adopted by agINFRA for the publication of vocabularies as LOD aims at reusing existing resources as much as possible. According to the methodology agreed in the project, the first step consists in analyzing the datasets available and the metadata sets and KOS used (presented in this paper). The table below summarizes the germplasm and soil data sets considered so far in agINFRA, together with the metadata sets and KOS used.

Table 1. Germplasm and soil datasets in agINFRA, with adopted metadata sets and KOSs.

Type of resource	Collection name	Metadata set used	KOS used
Germplasm	**CRA Germplasm (Italy)**	Multi-crop Passport Descriptors (MCPD) V.2	(Options see discussion below)
	CGRIS (China)	Own set of germplasm descriptors	
Soil datasets and maps	**Italian Soil Information System (ISIS)**	ISO 19115/19139[16]	US Soil Taxonomy , World Reference Base for Soil Resources

Then, we can distinguish the following cases:

1) The data set already uses some standard vocabularies published as LOD. Then the LOD publication is straightforward.

2) The data set uses some local vocabularies, with the same intended meaning as some standard vocabulary. Then, if the data owners agree on replacing them with those standard vocabularies, we are back to case 1.

---

[13] iPlant: http://www.iplantcollaborative.org/discover/semantic-web

[14] SSWAP: http://sswap.info/

[15] Simple Semantic Web Architecture and Protocol (SSWAP): an OWL implementation that offers the ability to describe data and services in a semantically meaningful way.

[16] ISO 19115/19139: Geographic information – Metadata, and XML schema implementation, http://www.iso.org/iso/home/store/catalogue_tc/catalogue_tc_browse.htm?commid=54904&published=on&includesc=true

3) The data set uses some local vocabularies, with the same intended meaning as some standard vocabulary. But the data owners need to keep the local ones. Then, agINFRA will introduce a set of mapping between the local and standard vocabularies.

4) The data set uses some local vocabularies, with no overlap with any standard vocabularies. Then agINFRA will publish them as LOD under the project namespace.

4 Conclusions

The study of current germplasm and soil data management practices revealed that experts in these two areas are actually looking forward to the adoption of LOD technologies to improve the interoperability of their data. The publication of additional germplasm and soil-related vocabularies will be a big step forward and will represent one of the novel contributions that agINFRA makes to the agricultural data management community.

We foresee that publishing both types of vocabularies as Linked Data will amplify their power by making them machine-readable, easily re-usable and linked or potentially linkable to other vocabularies.

Acknowledgements. The research leading to these results has received funding from the European Union Seventh Framework Programme (*FP7/2007-2013*) under *grant agreement* n° *283770*. We would like to thank Fang Wei, from the Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing, China, for his advice on the germplasm and soil related information in this paper. We would also like to thanks Vassilis Protonotarios, from the University of Alcala, Spain, for his contribution to an earlier version of this work.

5 References

1. Méndez, E., Greenberg, J.: Linked Data for Open Vocabularies and HIVE's Global Framework. El profesional de la información, vol. 21, pp. 236-244 (2012), http://www.elprofesionaldelainformacion.com/contenidos/2012/mayo/03_eng.pdf
2. Isaac, A., Waites, W., Young, J., Zeng, M.: Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets. W3C Incubator Group Report (2011), http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/
3. Alercia, A., Diulgheroff, S., Mackay, M.: FAO/Bioversity Multi-Crop Passport Descriptors V.2 (MCPD V.2) (2012), http://eurisco.ecpgr.org/fileadmin/www.eurisco.org/documents/MCPD_V2_2012_Final_PDFversion.pdf
4. Bioversity International: Key access and utilization descriptors for crop genetic resources (2011), http://www.bioversityinternational.org/index.php?id=3737
5. van Hintum, T.: Inclusion of C&E data in EURISCO - analysis and options. EPGRIS-3 Proposal, Wageningen (2009), http://edepot.wur.nl/186143

6. Endresen, D., Knüpffer, H.: The Darwin Core Extension for Genebanks opens up new opportunities for sharing germplsam data sets. Biodiversity Informatics, 8, 2012, pp. 12-29, https://journals.ku.edu/index.php/jbi/article/viewFile/4095/4064
7. U.S. Department of Agriculture, NRCS: NASIS-Related Metadata, http://soils.usda.gov/technical/nasis/documents/metadata
8. Federal Geographic Standard Committee (FGDC): Geospatial Metadata Standards (2012), http://www.fgdc.gov/metadata/geospatial-metadata-standards
9. U.S. Department of Agriculture, Natural Resources Conservation Service: Soil Taxonomy - A Basic System of Soil Classification for Making and Interpreting Soil Surveys (1999), http://soils.usda.gov/technical/classification/taxonomy
10. IUSS Working Group: World reference base for soil resources, 2nd edition, 2006. World Soil Resources Reports No. 103. FAO, Rome (2006), http://www.fao.org/ag/agl/agll/wrb/doc/wrb2006final.pdf
11. GS SOIL: Establishment of a multilingual soil-specific thesaurus. Deliverable 3.5. Prepared by Herbert Schentz et al., 31 May 2012 (2012), http://www.gssoil-portal.eu/Best_Practice/GS_SOIL_D3.5_Soil_Specific_Thesaurus_final.pdf
12. INSPIRE Thematic Working Group Soil: D2.8.III.3 INSPIRE Data Specification on Soil – Draft Technical Guidelines (2013), http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_SO_v3.0rc3.pdf
13. Pourabdollah, A., Leibovici, D.G., Simms, D.M., Tempel, P., Hallett, S.H., Jackson, M.J.: Towards a standard for soil and terrain data exchange: SoTerML, Computers & Geosciences, Volume 45, August 2012, Pages 270-283, ISSN 0098-3004, http://dx.doi.org/10.1016/j.cageo.2011.11.026. http://www.sciencedirect.com/science/article/pii/S0098300411004195
14. Montanarella, L., Wilson, P., Cox, S., McBratney, A.B., Ahamed, S., McMillan, B., Jaquier, D., Fortner, F.: Developing SoilML as a Global Standard for the Collation and Transfer of Soil Data and Information. Geophysical Research Abstracts vol. 12. Proceedings EGU General Assembly 2010, Copernicus (2010).
15. Arnaud, E., Cooper, L., Shrestha, R., Menda, N., Nelson, R.T., Matteis, L., Skofic, M., Bastow, R., Jaiswal, P., Mueller, L., McLaren, G.:. Towards a Reference Plant Trait Ontology for Modeling Knowledge of Plant Traits and Phenotypes (2012), http://wiki.plantontology.org/images/6/6e/Ref_TO_KEOD_2012.pdf
16. Cooper, L., Walls, R.L., Elser, J., Gandolfo, M.A., Stevenson, D.W., Smith, B., Preece, J., Athreya, B., Mungali, C.J., Rensing, S., Hiss, M., Lang, D., Reski, R., Berardini, T.Z., Li, D., Huala, E., Schaeffer, M., Menda, N., Arnaud, E., Shrestha, R., Yamazaki, Y., Jaiswal, P.: The Plant Ontology as a Tool for Comparative Plant Anatomy and Genomic Analyses (2013), http://pcp.oxfordjournals.org/content/54/2/e1.short
17. Nelson, R. T., Avraham, S., Shoemaker, R.C., May, G.D., Ware, D., Gessler, D.D.G.: Applications and methods utilizing the Simple Semantic Web Architecture and Protocol (SSWAP) for bioinformatics resource discovery and disparate data and service integration (2010), http://www.biodatamining.org/content/3/1/3#B13