Opening Soil Data through Semantics and Linked Vocabularies

Giovanni L'Abate*,Caterina Caracciolo**, Ferdinando Villa***, Edoardo Costantini*
*Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria (CREA)
**Food and Agriculture Organization of the UN (FAO)
***Basque Centre for Climate Change (BC3); IKERBASQUE, Basque foundation for science

Abstract
The World Reference Base for Soil Resources (WRB) is the internationally recognized classification system for soils. The WRB is commonly used to classify soil profiles and generate maps of soil distribution. However, it's still not available in an open standard and machine readable format; instead, it is published in textual reports and its encoding in data and metadata is left to users. As a result, data artifacts using the WRB show a wide variety of structures and approaches to versioning, which interfere with proper exchange and integration. This problem applies to soil data as well as to other datasets and domains that relate to soil science, from natural and physical to social and climate sciences. We discuss the usage of the WRB classification while applying principles of openness, interoperability and semantic correctness, meant to support reliable data exchange in science and, ultimately, better informed policy and decision making.

Introduction
Soil, the superficial layer of the earth, is the substrate for all human activities, playing a critical role in agriculture, climate control, runoff control, and many other life-supporting processes. The availability of reliable and unambiguous data on soil is crucial to modelers and to the scientific and policy communities. In this work we discuss about soil data generated from expert analysis and classification of geographically located soil profiles, used to compile soil maps. Soil taxonomies synthetically describe soil types by assigning soil profiles to hierarchical classes system using soil data. Among them, the WRB is one of the most widely used, together with the USDA Soil Taxonomy (Soil Survey Staff, 2014).
WRB is endorsed by the International Union of Soil Sciences (IUSS) and FAO and it is the result of an international collaboration coordinated by the International Soil Reference and Information Centre, which produces regular updates (1998[1], 2006[2], 2014[3]). WRB interprets soil morphology mainly as an expression of the pedogenetic processes. At the first level, the Reference Soil Groups (RSGs), classes are differentiated according to characteristic soil features produced by primary processes. At the second level (RSGs with qualifiers), soil features resulting from secondary soil-forming processes are used to further specify the primary characteristics.
The classification system allows for the combined use of a set of principal and supplementary qualifiers (qualifiers, either with prefixes or suffixes). This mechanism allows for a very precise characterization and classification of individual soil profiles. It also allows for effective connection to national classification systems. To date, WRB is published in textual formats (electronic or printed). It is then left to users to define database schemata and data structures as they find most convenient in terms of maintenance and use. The consequence is a lack of clarity and an even worse opportunity for data interoperability. This fact also limits the ease of updating classifications when revisions are published.
The ultimate goal of our work is to promote data and service interoperability, by proposing a formalized terminology and explicit rules of assembly for WRB-based soil taxonomies, in line with principles of openness and semantic coherence. Our goal is to enable soil data management organizations to deliver services compliant with a shared and unambiguous data model. Many applications could then be conceived to access and utilize this standardized data for a variety of purposes, from scientific research to policy making and commercial purposes. The problem of local vocabularies used for describing soil data was addressed by some of the authors within the

1 http://www.fao.org/docrep/w8594e/w8594e00.HTM
2 ftp://ftp.fao.org/docrep/fao/009/a0510e/a0510e00.pdf
3 http://www.fao.org/3/a-i3794e.pdf

agINFRA project (L'Abate et al., 2015). However, that work presented a limitation, in that the vocabulary created[4] focused on an INSPIRE-compliant soil database structure[5] (INSPIRE Thematic Working Group Soil, 2015) and did not adopt a compositional/faceted approach to render WRB, which is in fact a faceted classification. In the work presented here, we aim at addressing that limitation by clarifying the semantics and laying out the compositional constraints inherent in the classification. We have produced an implementation to allow for annotation, validation and machine translation, leveraging the functionalities of the k.LAB[6] (Athanasiadis and Villa, 2013) semantic modeling platform.

Methodology and results

The assumption underlying our work is that soil groups and properties should be described separately from the rules used to combine them into the description of a given soil type. We aim at separating the conceptual and terminological part from the classificatory one, exposing the former through a vocabulary published for general reuse, and the latter through a formal grammar, for which compositional rules are formally defined and implemented in k.LAB. The WRB 2006 version was selected as starting point for our work, since a wide range of databases (Table 1), published online through OGC standards (Web Feature Service[7]), use this version. In compliancy with INSPIRE, the CREA SoilProfile database share a fully specified WRB annotation implementing Terms instead of codes.

The 32 RSGs, qualifiers, prefixes and suffixes were extracted (Table 2) from previously published agINFRA vocabulary. The compositional rules were analyzed and defined in a formal grammar. A software parser was written for the k.LAB platform that can read the specification, parse them into terms, and validate the terms according to the vocabulary. After validation, the software applies the compositional rules according to the grammar, produces the definition of a concept that is then linked to the formal ontologies used in k.LAB. The open source parser is being tested for generality and correctness; in its production release it will support the creation of ontology concepts for soil types according to arbitrary ontologies chosen by the user.

Conclusions and future work

At the time of writing, the reusable fragment of the agINFRA soil vocabularies was identified and selected, and the compositional rules for the WRB classification were defined and implemented. While the software components of the work reported is near-final state, some issues remain before complete applications can be implemented. The authors are investigating hosting alternatives for maintaining and exposing the vocabularies through stable, reliable, and public endpoints and web services. In our view, the WRB vocabulary will become an "authority list" of concepts, to be reused by a variety of third party applications. The use of WRB authorities, in conjunction the compositional grammar implemented in k.LAB will allow for the annotation of diverse data sources in an unambiguous and machine actionable way. To this end, a phase of testing is taking place within the k.LAB platform, after which the source code of the WRB parser will be advertised as an independent service for use also within projects not related to k.LAB.

Acknowledgements

We acknowledge the support of the 7th FP agINFRA project in previous work. The work described was performed with time donated by the authors.

References

Athanasiadis, N Villa, F. A roadmap to domain specific programming languages for environmental modeling: key requirements and concepts, Proc. 2013 ACM workshop on Domain-specific

4 http://vocabularies.aginfra.eu/soil#

5 http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_SO_v3.0rc3.pdf

6 http://www.integratedmodelling.org

7 https://portal.opengeospatial.org/files/?artifact_id=8339

modeling, pg. 27--32, 2013, ACM, doi:10.1145/2541928.2541934.
L'Abate, G et al. Exposing vocabularies for soil as Linked Open Data. Info Proc Agri (2015), http://dx.doi.org/10.1016/j.inpa.2015.10.002
INSPIRE Thematic Working Group Soil. D2.8.III.3 INSPIRE DataSpecificationonSoil–DraftTechnicalGuidelines.Link:http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_SO_v3.0rc3.pdf . 2015.
Soil Survey Staff, 2014. Keys to Soil Taxonomy, 12th ed. USDA-Natural Resources Conservation Service, Washington, DC.

Database	**Owner**	**Feature/Raster**
SoilProfile	CREA	Point
SoilSamples	CREA	Point
SpectralLibrary	CREA	Point
WOSIS	ISRIC	Point
Soil_Regions	CREA	Polygon
Soil type classification WRB	SBG	Polygon
SoilGrids250m	ISRIC	Raster

Table 1. Online Databases implementing Open Geospatial Consortium Web Feature Service and adopting the WRB 2006 classification.

Term	**EnglishprefLabel**	**Concept**
(Abruptic)	xl_en_5081	c_12694
(Aceric)	xl_en_5082	c_12685
(Acric)	xl_en_5083	c_12683
(Acroxic)	xl_en_5084	c_12693
(Albic)	xl_en_5085	c_12681
(Alcalic)	xl_en_5086	c_12698
(Alic)	xl_en_5087	c_12688
(Aluandic)	xl_en_5088	c_12680
(Alumic)	xl_en_5089	c_12697
(Andic)	xl_en_5090	c_12690
…		

Table 2. Example of WRB 2006 classification suffixes extracted out of the agINFRA Soil Vocabulary.