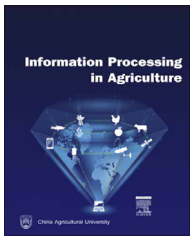
HOSTED BY

Available at www.sciencedirect.com

INFORMATION PROCESSING IN AGRICULTURE 2 (2015) 208–216

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




# Exposing vocabularies for soil as Linked Open Data

*Giovanni L'Abate* [a,*], *Caterina Caracciolo* [c], *Valeria Pesce* [b], *Guntram Geser* [e], *Vassilis Protonotarios* [d], *Edoardo A.C. Costantini* [a]

[a] *Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria (CREA), Agrobiology and Pedology Research Centre, Firenze, Italy*
[b] *Global Forum on Agricultural Research, Rome, Italy*
[c] *Food and Agriculture Organization of the United Nations, Rome, Italy*
[d] *University of Alcala, Alcala de Henares, Spain*
[e] *Salzburg Research, Salzburg, Austria*

ARTICLE INFO

*Article history:*
Received 2 January 2014
Received in revised form
9 October 2015
Accepted 14 October 2015
Available online 10 November 2015

*Keywords:*
Agriculture
Soil
Knowledge Organization Systems
Vocabularies
Resource Description Framework
Linked Open Data
INSPIRE

ABSTRACT

Standards to describe soil properties are well established, with many ISO specifications and a few international thesauri available for specific applications. Besides, in recent years, the European directive on "Infrastructure for Spatial Information in the European Community (INSPIRE)" has brought together most of the existing standards into a well defined model. However, the adoption of these standards so far has not reached the level of semantic interoperability, defined in the paper, which would facilitate the building of data services that reuse and combine data from different sources.

This paper reviews standards for describing soil data and reports on the work done within the EC funded agINFRA project to apply Linked Data technologies to existing standards and data in order to improve the interoperability of soil datasets. The main result of this work is twofold. First, an RDF vocabulary for soil concepts based on the UML INSPIRE model was published. Second, a KOS (Knowledge Organization System) for soil data was published and mapped to existing relevant KOS, based on the analysis of the SISI database of the CREA of Italy. This work also has a methodological value, in that it proposes and applies a methodology to standardize metadata used in local scientific databases, a very common situation in the scientific domain. Finally, this work aims at contributing towards a wider adoption of the INSPIRE directive, by providing an RDF version of it.



## 1. Introduction: data interoperability, metadata and the agINFRA project

In an era where data are produced at extremely high rates from a wide variety of sources and have to be made available to multiple stakeholders, from researchers and scientist to the general learners, the need for quickly identifying relevant data and linking or somehow combining data coming from heterogeneous data sources is strongly felt. The term normally used to define the set of features that data or metadata need to have in order to allow for this linking and combining of heterogeneous data is "data interoperability". "Data

* Corresponding author at: Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria (CREA), Agrobiology and Pedology Research Centre, Piazza M. D'Azeglio 30, 50121 Firenze, Italy. Tel.: +39 0552491239; fax: +39 055241485.
E-mail address: giovanni.labate@entecra.it (G. L'Abate).
Peer review under the responsibility of China Agricultural University.
http://dx.doi.org/10.1016/j.inpa.2015.10.002

interoperability is a feature of datasets and of information services that give access to datasets, whereby data can easily be retrieved, processed, re-used, and re-packaged ("operated") by other systems." [1].

In recent years, advocates of data interoperability have moved away from recommending the use of homogeneous metadata and formats, and embraced instead the view that it can be reached by using semantically defined classes, properties, concepts, and by identifying them with resolvable Uniform Resource Identifiers (URIs), in order to aloe for easy reuse of them. The resulting web of interlinked things is termed "Linked Data", and the type of interoperability that results from it is usually called "semantic interoperability" [2].

A few examples exist of applications that adopt the linked data approach in agricultural domain, like AGRIS [3]. Users of AGRIS can search for bibliographic references as well as full text documents and other types of data. The backbone of the AGRIS infrastructure, what allows the various pieces of information to be linked together, is the AGROVOC [4,5] thesaurus. However, the limited availability of linked data in agriculture hampers the diffusion of initiatives like AGRIS. Here is where agINFRA comes into play.

Data interoperability depends on the way data are described and classified. Two things are necessary to describe data. First, one needs metadata elements to describe various aspects of the data, e.g., title and abstract for publications, or porosity for a soil observation. Then, values for those metadata elements are needed. These values may be either "free values" (as in the case of the title of a book, or depth in meters of a soil sample), or they may be taken from "controlled vocabularies," or "authority data", such as thesauri that provide subject heading references for the metadata element "subject", or allowed types of documents for the metadata element "document type".

The "metadata elements" used to describe a given type of data, or a dataset, are usually referred to as "metadata vocabulary", "metadata sets", "metadata element sets", or simply "vocabularies", while the "controlled vocabularies" allowed for any of the metadata elements are also often called "authority data", "value vocabularies" or "Knowledge Organization Systems (KOSs)". A common source of confusion is that the term "vocabulary" (cf. [6]) is often used as a short for both dimensions. We often use one or the other of these forms, although we prefer to keep the two notions separate and tend to talk about "metadata elements" that may be grouped together in "metadata vocabularies", and that may take their values from KOS, or controlled vocabularies.

Metadata sets and KOSs have a long history, but they have gained renewed interest in the context of use of the RDF (Resource Description Framework) triple-based data model. To ensure that the description of data by means of RDF triples (statements formed by "subject – predicate – object") is unambiguous, the predicate used in the triple must be unambiguous. The way to ensure that predicates are unambiguous is to provide them with a defined semantics and collect them in public vocabularies, described and promoted so as to become standard. Each metadata element (predicate of an RDF triple) is then given an URI, and the same is done for concept used as value of the element (the object of that triple). Metadata elements expressed as RDF vocabularies have then "machine-readable" semantics: "objects" described with RDF vocabularies can be "operated" by machines. In general, if elements in metadata vocabularies are linked together, they will be Linked Open Data (LOD) vocabularies. For instance, continuing with the terminology proper of RDF triples, consider the link between properties as in the case property "themes" in the W3C Data Catalog Vocabulary (DCAT), defined as "subproperty" of "subject" in the Dublin Core metadata vocabulary [7], or the links between objects defined in KOSs, like "soil density" from the AGROVOC thesaurus and "soil density" from the NAL Thesaurus. In general, we say that data described with any linked vocabulary qualify as Linked Open Data.

Following the line of reasoning described above, agINFRA first focussed on identifying and recommending existing RDF vocabularies or publishing new ones if necessary. agINFRA [8] is a project (2011–2015) co-funded by the European Commission, within the FP7 Research and Innovation funding programme. agINFRA aims to facilitate the accessibility of agricultural data by providing the workflows and necessary grid and cloud based infrastructures required for the development of large agricultural data pools, which will be available to all stakeholders. In this direction, agINFRA aims to provide the tools and methodology to be used for the publication of the data managed by project partners as Linked Open Data (LOD). This is expected to significantly facilitate the interoperability between heterogeneous data sources, not previously linked in any way. The first step of the agINFRA consortium towards the publication of vocabularies as linked data was the identification of the metadata sets and KOSs used by the agINFRA data providers in their data sources [9], and their publication as LOD if these were not already published. agINFRA deals with data (and metadata) pertaining to different areas, namely bibliography, education, germplasm, and soil. This paper reports on the work done in particular on soil data.

The paper is organized as follow: Section 2 is a review on standards for soil data. Section 3 describes the work done within agINFRA project and reports on the obtained results. Conclusions follow in Section 4.

## 2. An overview of metadata vocabularies and KOSs for soil data

### 2.1. Soil metadata vocabularies

Several disciplines look at the soil in different ways (e.g. Engineering, Biology [10,11], Soil cartography [12,13]) and therefore typically use different references for characterizing soil features, like depth, history, chemical composition, morphology, and classification, as well as sampling and laboratory methodologies, and geographical reference systems.

For soil data, different metadata standards already exist [11,14,15]. They are formalized in various ways, from database structures to ISO standards [16] to XML implementations [17,18] to, in a few cases, RDF [19,20].

The international Working Group on Soil Information Standards (WG-SIS) [14], an initiative within the International Union of Soil Science, aims to develop, promote and maintain internationally recognized and adopted standards for the exchange and collation of consistent harmonized soils data