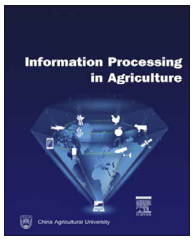
HOSTED BY

ELSEVIER

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

and information. Geographical aspects of soil data are covered by well established standards, namely the ISO standards ISO 19115 "Geographical Information – Metadata" [21] and ISO 19119 "Geographic Information – Services" [22]. They cover geographic information and services, respectively. Taken together, they fully describe datasets, including individual geographic features and feature properties. Moreover, ISO 19139 "Geographic Information – Metadata – XML Schema implementation" [23] defines the Geographic MetaData XML (gmd) encoding, and XML Schema implementation for ISO 19115, including the extensions for imagery and gridded data. The adoption of these ISO standards is growing. For example, the U.S. Federal Geographic Data Committee (FGDC) recently recommended that users of the Content Standard for Digital Geospatial Metadata (CSDGM) [24,25] adopt the ISO standard.

For soil data and terrain attributes, the World SOil and TERrain (SOTER) Digital Database is a major initiative, started by the International Society of Soil Science (ISSS) in 1986. The Geoscience Markup Language (GeoSciML) was developed for the SOTER model, and then extended by SoTerML [26,27], the XML language "Soil and Terrain and Markup Language", developed by the Centre for Geospatial Science in the University of Nottingham, compliant with another ISO standard, the ISO/TC190/SC 1 N140 "Recording and exchange of soil-related data".

A recent initiative to harmonize different soil schemas is the Soil-ML project [28], a soil equivalent of the GeoSciML [18] providing Definitions for application schema "ISO 28258 Definitions" [29].

2.2. The INSPIRE directive

In 2007, the European Union established the directive on Infrastructure for Spatial Information in Europe (INSPIRE) [30] to provide an infrastructure for spatial information in Europe and support Community environmental policies, as well as policies or activities which may have an impact on the environment. Within that framework, also a comprehensive standard covering both geographic and scientific aspects of soil data was created. One of the goals of INSPIRE is to harmonize different national norms. The section on standards for soil has recently been completed.

INSPIRE defines a data model, described in the INSPIRE Implementing Rules on interoperability of spatial data sets and services [31,32], and the data specification guidance documents specific for soil [33].

The INSPIRE data models developed by the INSPIRE Thematic Working Groups, are graphically represented according to the UML (Unified Modelling Language) notation, and are based on the INSPIRE XML schema. INSPIRE was designed according to the ISO standard ISO/TS 19103:2005 "Geographic information – Conceptual schema language" [34]. Application schemas are specified in UML notation, version 2.1 according to ISO 19109 "Geographic Information – Rules for application schema" [35] and the Generic Conceptual Model.

INSPIRE thus provides standards both for the metadata and the data model, and for the controlled values to be used (KOSs).

Figs. 1 and 2 show a high level view of the INSPIRE model related to soil, represented according to UML notation. Fig. 1 shows the properties related to soil profile, while Fig. 2 shows the properties related to soil coverage.

Table 1 gives a schematic representation of the XML specification for Soil Profile according to INSPIRE.

From Fig. 1 one may see that the core entities in the model are SoilProfile, ProfileElement, and SoilDerivedObject. From the same figure, one may notice that the entity "SoilProfile" can be instantiated either as an "ObservedSoilProfile" or as a "DerivedSoilProfile", which is non-georeferenced and can be derived (e.g. averaged) from one or more observed profiles. Both types of soil profiles can be described in terms of Profile Elements: Horizons or Layers, which describe the vertical section of the soil profile. Each of these entities has several properties (observed, measured or derived), many of which represent properties or parameters expressed through specific units of measurement or classes and are divided in several major groups, such as Chemical, Physical, and Biological parameters. One or more DerivedSoilProfiles are associated to "SoilBody" object. The Soil Body concept represents an association of soils that are found together in a spatially delineated area. A soil related property, e.g. organic carbon content, can be derived from Soil Body features and represented as a polygon ("SoilDerivedObject").

A soil map, called "SoilThemeCoverage" (Fig. 2), is a spatial object type associated to a set of SoilDerivedObjects, which holds values for a property based on one or more soil and possibly non-soil parameters within its spatial, temporal or spatiotemporal domain.

The aspects of INSPIRE that relate to controlled vocabularies are described in the next section.

2.3. KOSs to describe soil

The main international classifications of soil types are the World Reference Base for Soil Resources [12] developed by the IUSS and FAO and published in 1998, and the USDA Soil Taxonomy [13] first published in 1975. An important recent achievement is the Multilingual Soil Thesaurus (SoilThes), an extension of the General Multilingual Environmental Thesaurus (GEMET) [36] developed in the eContentplus project GS SOIL [37]. SoilThes contains the concepts of the World Reference Base (WRB), the soil vocabulary of ISO 11074 and additional soil specific concepts. GEMET is the official thesaurus for the Infrastructure for Spatial Information in the European Community (INSPIRE) directive, within which draft Technical guidelines for data specification on soil has been recently published [33].

As described in the previous section, INSPIRE also includes controlled vocabularies. The INSPIRE infrastructure also includes a "Registry" [38], a public reference directory of all the published "registers" (in INSPIRE terminology) assigning published identifiers to specific controlled values. The INSPIRE registry defines the code list Extensibility (Table 2) that indicates how a code list (classes or ranges of values) may be extended with additional values defined by data providers (Table 3). Such additional values should be published in a register and should not replace or redefine any value already specified in the register.

The examples in the tables above clarify the importance of KOSs in providing correct values for many of the properties

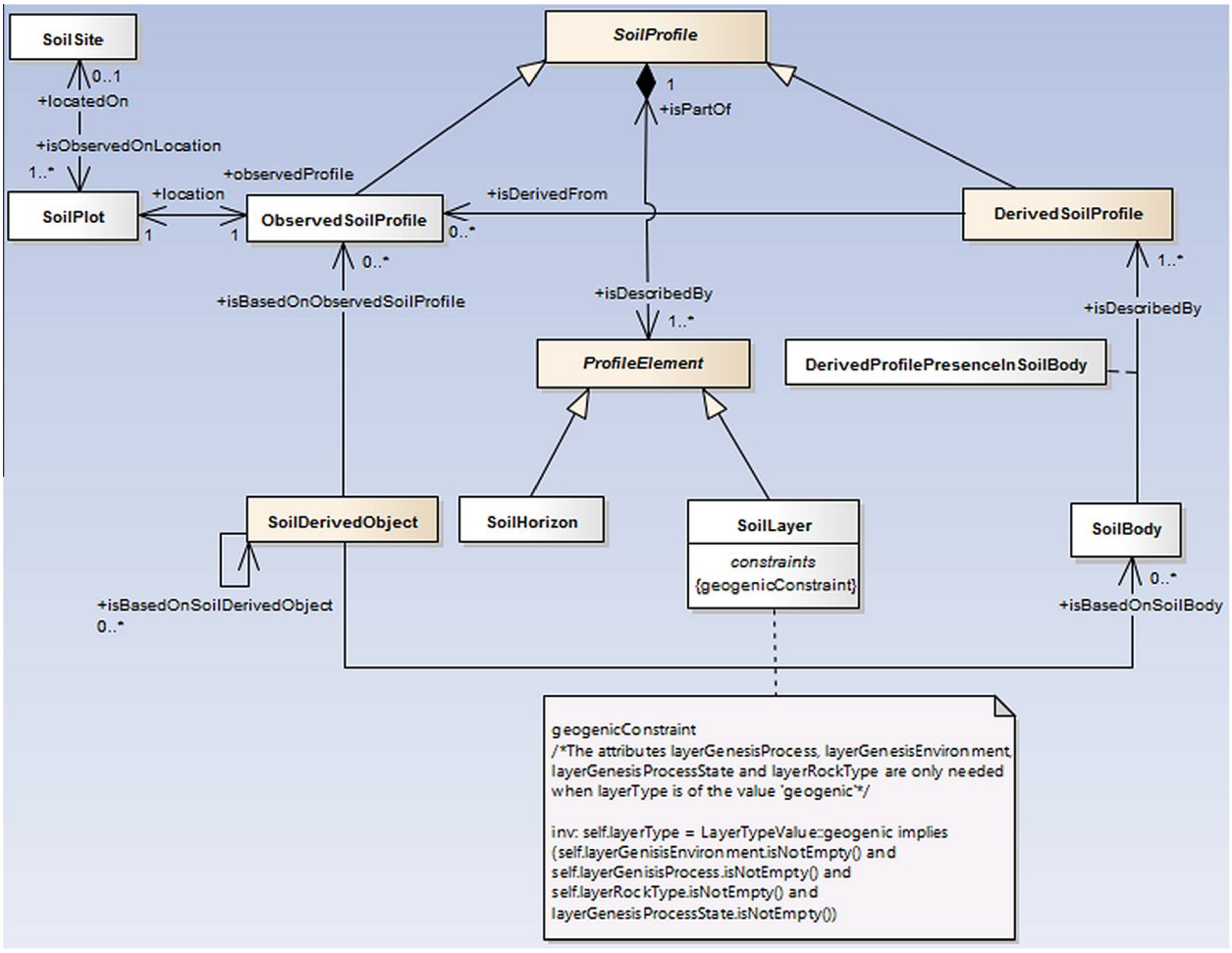


Fig. 1 – A fragment of the INSPIRE model, covering "soil profile". Simple arrows stand for Association (Knows); full white arrows are used for Inheritance (Is a); full black rhombus means Composition (Has).

(metadata elements) specified in the INSPIRE model. The ranges of these values can come from large classification systems or thesauri, as well as from small local lists of values. While a few comprehensive thesauri and classifications have been published as LOD, a wide abundance of smaller local lists of values (as conceived in the INSPIRE local registries) exist, that are seldom even published as KOSs, maybe because they tend to be used in local databases only. Therefore, the rendering of these small lists of values as KOSs is an interesting work for agINFRA. Publishing these small internal lists in a format compatible with the rest of the RDF-based infrastructure, i.e., by using SKOS (Simple Knowledge Organization System) allows us to map the values to external URIs and therefore make them LOD, thus making the data and applications using them more interoperable.

The INSPIRE infrastructure involves a number of items which require clear descriptions and the possibility of referencing through unique identifiers. Examples for such items include INSPIRE themes, code lists, application schemas or discovery services. Registers provide a means to assign identifiers to items and their labels, definitions and descriptions (in different languages). The content of these registers are based on the INSPIRE Directive, Implementing Rules and Technical Guidelines.

Allowed ranges for controlled vocabularies in INSPIRE are "free values" (see Table 3), or sets of discrete values from controlled values (i.e., values from a KOS). These second types of values may also be further constrained so that only certain values from a given KOS (e.g. the INSPIRE Registry or the thesaurus of the US National Library (NALT) [19] may be accepted (see Table 4).

The existing metadata standards indicate the ranges and accepted values for many of the properties: some values can be taken from published classifications (provided that a reference to the classification is made), while the other lists are either ISO standards or just local lists of values that have not been published elsewhere (Extensible Code list according to INSPIRE) [39].

One task of the agINFRA project is to publish most of these lists as LOD.

3. Publishing LOD vocabularies for soil data within agINFRA

In the work of achieving data interoperability within agINFRA, we focussed on the Soil Information System of Italy (SISI) database [40,41], the main agINFRA soil data base, maintained by the Italian CREA (Consiglio per la ricerca in agricoltura e

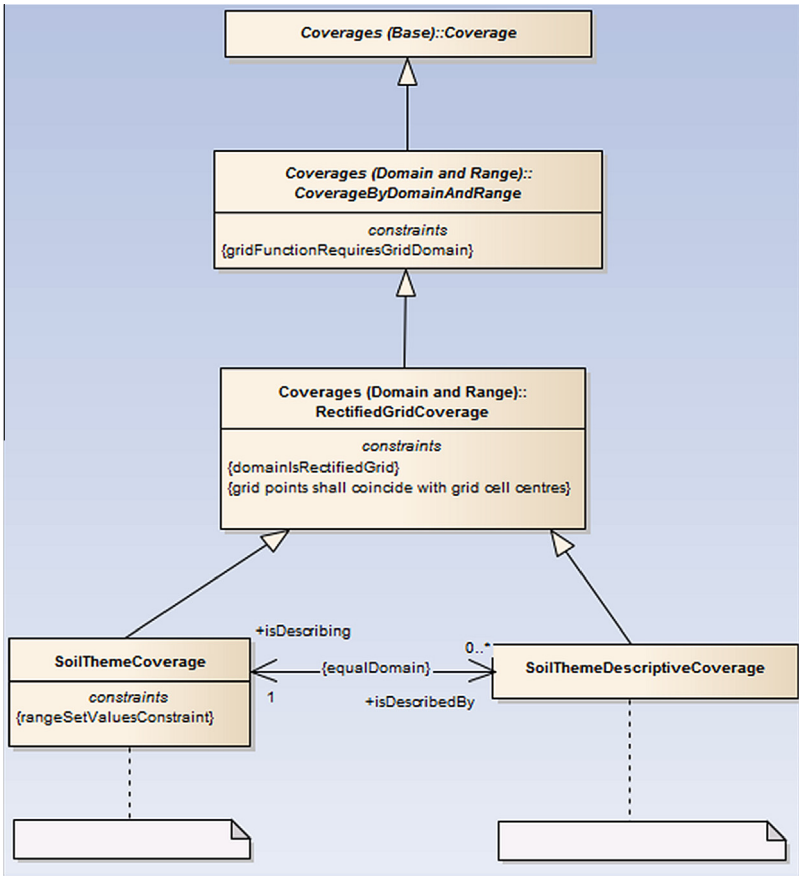


Fig. 2 – A fragment of the INSPIRE model, covering "soil coverage". Simple arrows are used for Association (Knows); full white arrows are used for Inheritance (Is a).

l'analisi dell'economia agraria). Metadata elements used in SISI mainly come from INSPIRE, while various standards are relevant for their values, like for example the USDA Soil Taxonomy and WRB.

In Europe, INSPIRE is the reference standard for soil data organization and dissemination. Since agINFRA is an EC funded project, INSPIRE was adopted as the starting point for a LOD metadata vocabulary for soil data. INSPIRE is a good starting point for both the definition of an RDF metadata vocabulary, since it defines entities, attributes and relation, and for the identification of KOSs that need to be published, as it defines "registers" of values.

3.1. An RDF soil metadata vocabulary

As briefly described above (Section 2.2), INSPIRE has an UML representation and an XML representation. The analysis of the UML diagrams revealed that a representation of the model in RDF would not be difficult to achieve. Briefly, what is considered a "class" in UML is not very different from what is considered a "class" in RDF, while UML attributes are conceptually very close to RDF "properties" (see [42] for a discussion on the correspondences between UML and RDF). Therefore, we transformed the main INSPIRE classes into RDF classes. The diagrams in Fig. 3 also indicate the subclass relationships and their main attributes and relationships into RDF properties. INSPIRE also has constraints and rules, which we decided not to formalize in a complex ontology.

In the agINFRA approach, vocabularies are used to exchange data and not to support applications. The vocabularies created in agINFRA are not tightly coupled to applications, since this should not rely on rules and constraints being enforced by the vocabulary. However, the INSPIRE recommendations can accompany the vocabulary and the soil community may decide to enforce them in practice.

An RDF vocabulary was then developed to express the core classes of the INSPIRE model. The diagram in Fig. 3 represents the structure of such a vocabulary.

To create and maintain such an RDF vocabulary, the tool Neologism [43] was used. The vocabulary was published under the namespace http://vocabularies.aginfra.eu/soil (as a work in progress).

3.2. New LOD KOSs for soil data

A few KOSs to describe soil are available as LOD. GEMET is available as an RDF/SKOS linked dataset, mapped to AGROVOC. SoilThes is also available as an RDF/SKOS dataset and linked to GEMET; the USDA Soil Taxonomy is part of the National Agricultural Library Thesaurus (NALT) [19], which

Table 1 – A schematic view of the INSPIRE XML representation of "soil profile".

≪feature type≫
SoilProfile
+ inspireId: (to be generated) + WRBSoilName: + WRBSoilNameType + WRBQualifierGroup: WRBQualifierGroupType + qualifierPlace: prefix + qualifierPosition: 1 + WRBqualifier: Haplic + WRBspecifier: – + WRBQualifierGroup: WRBQualifierGroupType + qualifierPlace: suffix + qualifierPosition: 1 + WRBqualifier: Calcaric + WRBspecifier: – + WRBReferenceSoilGroup: Arenosol + isOriginalClassification: true + otherSoilName: + soilName: Typic Xeropsamment + soilClassificationScheme: DocumentCitation + Name: Carta suoli Sicilia: convenzione con la Regione Sicilia per la realizzazione della Carta dei Suoli a scala 1:250,000 nell'ambito del programma interregionale "Agricoltura e Qualità" + shortName: Carta suoli Sicilia + date: 2011 + link: http://www.sias.regione.sicilia.it/ + isOriginalClassification: true + localidentifier: DSP 59.9ARCA1.1 + soilProfileParameter: – + validFrom: 2008 + validTo: 2009 + beginLifespanVersion: 2010 + endLifespanVersion: –

Table 2 – INSPIRE: code list extensibility.

Label	Definition
Empty code list	No values are specified for this code list in this register, i.e. it is allowed values to comprise any values defined by data providers
Extensible with narrower values	The code list can only be extended with narrower values, i.e. it is allowed values to comprise the values specified in this register and narrower values defined by data providers
Extensible with values at any level	The code list can be extended with additional values at any level, i.e. it is allowed values to comprise the values specified in this register and additional values at any level defined by data providers
Not extensible	The code list cannot be extended, i.e. it is allowed values to comprise only the values specified in this register

is published as LOD. For the spatial aspect, soil data can rely on many advanced RDF standards, mainly in the framework of the EU INSPIRE Directive.

However, none of the above KOSs as LOD was exploited in the SISI database, or not in their full potential. For example, values from the above KOSs were used in the database only as strings and not as URIs, with only indirect indication of the source authority from which the values come. This is not unusual in research databases: researchers know the standards and use the correct values when submitting the data. The source authority is implicit for them and they know their colleagues will understand the meaning of the string. In some cases, they even use the local language version of a value. This all works well as long as data are used by human beings, but it is problematic for data sharing and in general for the use of the data within applications (i.e. machines), for example to match (link) them with data coming from other sources.

One case that is particularly difficult is the case of values for which reference to a published thesaurus is recommended, but only a specific subset of terms is valid for a specific property. Actually, thesauri are rarely structured around "facets" (or the various properties of entities that can be described by the terms in the thesaurus): they usually

Table 3 – An example of code list extensibility for "slope steepness".

Slope class	Slope steepness factor (%)
Level	(<0.2)
Nearly level	(0.2–2)
Nearly level	(3–5)
Gently sloping	(6–13)
Strongly sloping	(14–20)
Moderately steep	(21–35)
Steep	(36–60)
Very steep	(61–90)
Extremely steep	(>90)

have an internal logic that reflects the domain they represent. However, in some cases it is still possible to extract specific subsets of terms. For instance, in the USDA Soil Taxonomy, concepts related to soil are organized in a hierarchical structure, and in some cases is possible to identify the specific "branch" or "sub-branch" that can be used as the range of controlled values for a specific property, such as soil type. This enables the possibility of prescribing a range of controlled values to be used (e.g., narrower concepts of a given concept) and map them to other values.

We decided that values in the SISI database should be first published as *new* KOS, them mapped to existing larger KOS. An URI would be assigned to each value found in the database and then mapped, whenever possible, to the URI of the corresponding value in an already published KOS (e.g. the Thesaurus by the National Agricultural Library (NAL) of the U.S. A or AGROVOC). The most relevant controlled values that we identified in the CREA database were the INSPIRE code list and many values corresponding to terms in the NAL Thesaurus.

Table 4 – Example of XML elements in the INSPIRE model that allows only a subset of values from a KOS.

XML element:	WRBReferenceSoilGroupValue
Label:	WRB reference soil group (RSG)
Definition:	A code list of possible reference soil groups (i.e. first level of classification of the World Reference Base for Soil Resources).
Description:	Reference Soil Groups are distinguished by the presence (or absence) of specific diagnostic horizons, properties and/or materials. NOTE The WRB soil classification system comprises 32 different RSGs. SOURCE World reference base for soil resources 2006, first update 2007, World Soil Resources Reports No. 103, Food and Agriculture Organization of the United Nations, Rome, 2007.
Extensibility:	None
Identifier:	<http://inspire.ec.europa.eu/codeList/WRBReferenceSoilGroupValue>

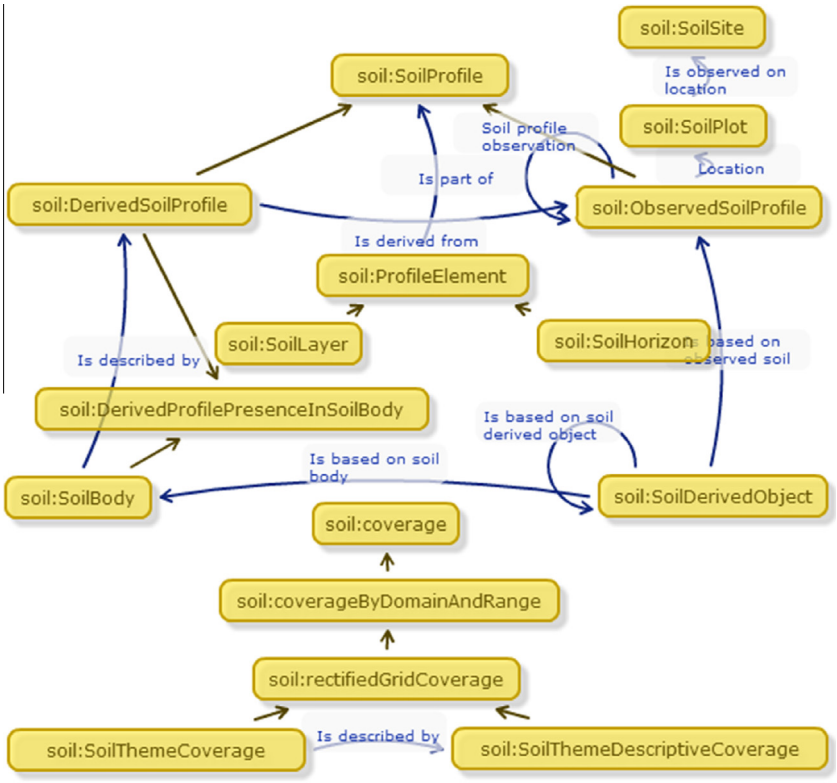


Fig. 3 – Graphical representation of the RDF vocabulary created from the conversion of the INSPIRE model.

In order to automatically import used Concepts and URIs into VocBench, a table was prepared that contained the fields listed below. Rows in the table represent a concept:

- ID (= a unique identifier for the concept)
- prefLabelEn (= preferred label for the concept in English)
- altLabelEn (= alternative label for the concept in English)
- prefLabelIt (= preferred label for the concept in Italian)
- broader (= broader concept)
- exactMatch (= URI of the concept from external KOS considered to be the same)
- narrowMatch (= URI of the concept from external KOS considered to be narrower)

The original SISI relational database was mapped to different microthesauri (28 tables in total) and their concepts (880 fields in total). All 28 microthesauri were considered as part of one big KOS. For every concept (field), a natural language descriptor in English was provided. If available among existing online vocabularies [19,20], the reference URI was also provided. Beside the appositely implemented "mapping" table with 880 records, two of the existing tables from the original SISI relational database were used to extract the resulting about 5,900 concepts. Mapping URIs of each concept were searched in the INSPIRE Registry, the NAL Thesaurus and the Linked Thesaurus Framework for Environment (LusTRE) [20]. When an exact match was found, the descriptor from INSPIRE or the Agricultural Thesaurus was adopted as an alternative English label of the concept, and the exact match was used as preferred label. Broader matches were defined according to the INSPIRE Implementing Rules and the SISI application schema. URIs to exactMatch and narrowMatch were stored in the above mentioned table. Preferred Alternative Labels in Italian were also created.

The main results of this work were the publication of an RDF Soil Vocabulary based on the INSPIRE model and the SISI web application, and the publication of an RDF Soil Concepts KOS. The editing tool used to maintain the Soil Concept KOS was VocBench.

4. Conclusions

In this paper, we provided an overview of the current state of metadata for soil data, distinguishing between metadata elements sets, or simply vocabularies, and KOS, i.e., controlled vocabularies that provide values for metadata elements. We have in particular described the INSPIRE initiative for what concern the modeling of soil information. The goal of our work was to enhance the level of interoperability between data sets, for which we believe metadata standardization is a fundamental ingredient. The work reported in this paper took the SISI database of the CREA as a starting point for a twofold action. On the side of metadata elements concerning soil, we provided an RDF representation of the UML concepts of the INSPIRE model concerning soil. On the side of the KOS used for metadata elements on soil, we published a mapping of the values used in the SISI database of the CREA to existing controlled vocabularies. When no corresponding values in known controlled vocabularies could be found, no mapping was provided.

As a result of this work, the SISI database is now more LOD oriented than before and therefore more suitable for data interoperability. Moreover, our work also resulted in a methodological achievement, in that it proposed and applied a methodology to deal with the conversion of small sized KOS, used in local databases, into linked vocabularies. Finally, we believe our work can contribute to a wider adoption of the INSPIRE initiative, by providing the community of soil researchers with an RDF version of it.

At the time of revision of this paper, we learned about an initiative of the Joint Research Centre (JRC) aimed at defining and promoting an RDF version of INSPIRE [44]. In its first phase of work, such an initiative focussed on defining methodologies for the creation of RDF vocabularies based on UML schemes. Soil was not considered into their work.

Given the current scarcity of linked data on soil, we believe that our work will improve research in agriculture since every missing node in the chain of linking weakens, if not breaks, the linking mechanism. We can imagine a number of applications that will be possible thanks to the work we conducted in agINFRA, and that we plan on continuing. For example, one could search for data or maps based on the geographical extent and/or the parameter if interested, for instance, soil type, pH, organic carbon, bulk density.

The study of current soil data management practices revealed that experts in this area are actually looking forward to the adoption of LOD technologies to improve the interoperability of their data. The publication of additional soil related INSPIRE compliant vocabularies will be a big step forward and will represent one of the novel contributions that agINFRA makes to the agricultural data management community. In this context, the agINFRA project aims to provide the tools and methodologies for enabling the soil data related vocabularies as linked data, enhancing the interoperability not only between soil data sources but also between soil data sources and sources of other types of data, like bibliographic and educational.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 283770.

REFERENCES

[1] Chinese Academy of Agricultural Sciences, Global Forum on Agricultural Research, Food and Agriculture Organization of the United Nations. In: Interim proceedings of international expert consultation on "building the CIARD framework for data and information sharing", 20–23 June 2011, Beijing. Link: http://www.ciard.info/sites/default/files/IECProceedings-main-doc_0.pdf. 2015.

[2] Heath T, Bizer C. Linked Data (Synthesis Lectures on the Semantic Web: Theory and Technology). USA: Morgan & Claypool Publishers; 2011.
[3] Food and Agriculture Organization of the United Nations. AGRIS: international information system for the agricultural science and technology. Link: http://agris.fao.org. 2015.
[4] Food and Agriculture Organization of the United Nations. AGROVOC multilingual agricultural thesaurus. Link: http://aims.fao.org/agrovoc/. 2015.
[5] Caracciolo C, Stellato A, Morshed A, Johannsen G, Rajbhandari S, Jaques Y, et al. The AGROVOC linked dataset. Seman Web 2014;4(3):341–8.
[6] 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.
[7] Dublin Core Metadata Initiative. DCMI metadata terms. Link: http://dublincore.org/documents/dcmi-type-vocabulary/. 2015.
[8] agINFRA project. Link: http://www.aginfra.eu/. 2015.
[9] Besemer H, Edge P, Subirats I, Tsiflidou E, Protonotarios V. Review of existing agricultural data management practices, lifecycles and workflows. Link: http://www.fao.org/docrep/017/aq190e/aq190e.pdf. 2012/2015.
[10] Cole JR, Myrold DD, Nakatsu CH, Owens PR, Kowalchuk G, Tebbe C, et al. Development of soil metadata standards for international DNA sequence databases. In: Proc. 19th world congress of soil science, soil solutions for a changing world. Brisbane, Australia; 2010. p. 5–8.
[11] Terragenome. International soil metagenome sequencing consortium. Metadata standards. Link: http://www.terragenome.org/metadata-for-soil-metagenomics/. 2015.
[12] Food and Agriculture Organization of the United Nations. World reference base for soil resources 2014. International soil classification system for naming soils and creating legends for soil maps. WORLD SOIL RESOURCES REPORTS. Link: http://www.fao.org/3/a-i3794e.pdf. 2015/2015.
[13] United States Department of Agriculture. Keys to soil taxonomy. 12th ed. Washington, DC: USDA-Natural Resources Conservation Service; 2014.
[14] International Working Group on Soil Information Standards. Link: http://www.soilinformationstandards.org/. 2015.
[15] World Soil Information Metadata Service. Link: http://meta2.isric.org/geonetwork/srv/en/main.home. 2015.
[16] ISO 11074:2005. Soil quality – Vocabulary. International Organization for Standardization. Link: https://www.iso.org/obp/ui/#iso:std:iso:11074:ed-1:v1:en. 2015.
[17] ISRIC. World Soil Information. XML Schemas. Link: http://schema.isric.org/. 2015.
[18] Computational Geoscience Group. Geoscience vocabularies for linked data. Link: http://resource.geosciml.org. 2015.
[19] National Agricultural Library Thesaurus. Link: http://agclass.nal.usda.gov/. 2015.
[20] LusTRE: Linked Thesaurus fRameowrk for Environment. Link: http://linkeddata.ge.imati.cnr.it:2020/. 2015.
[21] ISO 19115:2014(en) Geographic information — Metadata. International Organization for Standardization. Link: https://www.iso.org/obp/ui/#iso:std:iso:19115:-1:ed-1:v1:en. 2015.
[22] ISO 19119:2005(en) Geographic information — Services. International Organization for Standardization. Link: https://www.iso.org/obp/ui/#iso:std:iso:19119:ed-1:v1:en. 2015.
[23] ISO/TS 19139:2007(en) Geographic information — Metadata — XML schema implementation. International Organization for Standardization. Link: https://www.iso.org/obp/ui/#iso:std:iso:ts:19139:ed-1:v1:en. 2015.
[24] Federal Geographic Standard Committee (FGDC). Geospatial Metadata Standards, Link: http://www.fgdc.gov/metadata/geospatial-metadata-standards. 2012/2015.
[25] Federal Geographic Standard Committee (FGDC). Content Standard for Digital Geospatial Metadata. Link: http://www.fgdc.gov/metadata/csdgm/. 2015.
[26] Pourabdollah A, Leibovici DG, Simms DM, Tempel P, Hallett SH, Jackson MJ. Towards a standard for soil and terrain data exchange: SoTerML. Comput. Geosci. 2012;45:270–83.
[27] SoTerML Schema Specification. Link: http://www.isric.org/specification/SoTerML.xsd. 2015.
[28] Montanarella L, Wilson P, Cox S, McBratney AB, Ahamed S, McMillan B, et al. Developing SoilML as a global standard for the collation and transfer of soil data and information. Link: http://eusoils.jrc.ec.europa.eu/esdb_archive/eusoils_docs/Poster/montanarella_EGU2010_XML.pdf. 2010/2015.
[29] ISO 28258:2013(en) soil quality — digital exchange of soil-related data. International Organization for Standardization. Link: https://www.iso.org/obp/ui/#iso:std:iso:28258:ed-1:v1:en. 2015.
[30] INSPIRE. Infrastructure for Spatial Information in the European Community. Link: http://inspire.ec.europa.eu/. 2015.
[31] D007474/02. Implementing Directive 2007/2/EC of the European Parliament and of the Council as regards interoperability of spatial data sets and services. INSPIRE Committee. Link: http://inspire.ec.europa.eu/index.cfm/newsid/4204. 2010/2015.
[32] Regulation (EU) No 1253/2013. Annex II+III amendment to Implementing Rules on the interoperability of spatial data sets and services published. INSPIRE. Link: http://inspire.ec.europa.eu/index.cfm/newsid/11303. 2013/2015.
[33] INSPIRE Thematic Working Group Soil. D2.8.III.3 INSPIRE Data Specification on Soil – Draft Technical Guidelines. Link: http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_SO_v3.0rc3.pdf. 2015.
[34] ISO/TS 19103:2005(en) Geographic information — Conceptual schema language. International Organization for Standardization. Link: https://www.iso.org/obp/ui/#iso:std:iso:ts:19103:ed-1:v1:en. 2015.
[35] ISO 19109:2005(en) Geographic information — Rules for application schema. International Organization for Standardization. Link: https://www.iso.org/obp/ui/#iso:std:iso:19109:ed-1:v1:en. 2015.
[36] Eionet. GEMET Thesaurus. Link: http://www.eionet.europa.eu/gemet. 2015.
[37] GS Soil project. Link: http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=ECP-2008-GEO-318004. 2015.
[38] INSPIRE Registry. Inspire archive. Link: http://inspire.ec.europa.eu/Registry. 2015.
[39] INSPIRE Registry. Registry of INSPIRE Code Lists. Link: http://inspire.ec.europa.eu/codelist/. 2015.
[40] L'Abate G, Allegri G, Barbera R, Bruno R, Fargetta M, Costantini E A C. The SISI webgisapplication for online Italian soil data consultation. EFITA-WCCA-CIGR Conference "Sustainable Agriculture through ICT Innovation", Turin, Italy; 2013. Link: https://www.academia.edu/4295019/The_SISI_web-GIS_application_for_online_Italian_soil_data_consultation.
[41] Costantini EAC (ed.). Linee guida dei metodi di rilevamento e informatizzazione dei dati pedologici. Italia: CREA_ABP, Firenze; 2008.
[42] Chang WW. A Discussion of the Relationship between RDF-Schema and UML. Link: http://www.w3.org/TR/NOTE-rdf-uml/. 1998/2015.
[43] National University of Ireland. Neologism. Link: http://neologism.deri.ie/. 2015.
[44] INSPIRE MIG permanent technical sub-group (MIG-T). ARE3NA study on "RDF & PIDs for INSPIRE". Link: https://ies-svn.jrc.ec.europa.eu/projects/rdf-pids/wiki/ARE3NA_RDF_+_PIDs_study. 2015.