21st International Congress on Modelling and Simulation, Gold Coast, Australia, 29 Nov to 4 Dec 2015
www.mssanz.org.au/modsim2015

Towards an Ontology-based Soil Information System

Yanfeng Shu [a] and Qing Liu [a]

[a] *Private Bag 12, Hobart, TAS 7001, Australia*
Email: Yanfeng.Shu@csiro.au

Abstract: Environmental information is critical to the sustainable use and management of the world's resources. Soils are a fundamental part of the environmental information requirement, and appropriate soil data and information are crucial to support evidence-based policy, planning and resource management decisions.

For data to be useful, one basic requirement is that they be interpretable. Sufficient information should be provided to allow data to be unambiguously interpreted and used. Examples of such information include the location at which the soil was sampled, the property that was measured, the unit of measure, and quality assurance and quality control information. Furthermore, data should be easily integrated with other data sources, which is required in many modelling applications. For example, simulation of crop production may require, besides soil data, also weather, crop and fertilizer data.

To meet these requirements, we have developed a soil ontology for modelling soil information. In this paper, we focus on the design of the ontology and its potential applications. We describe the use of the ontology to facilitate data access by mapping data to the ontology and making them available as Linked Data. We also discuss applications of the ontology for data integration, data classification and data validation.

Keywords: *Soil information, ontology, Linked Data, data integration, data classification*

1 INTRODUCTION

Environmental information is critical to the sustainable use and management of the world's resources. Soils are a fundamental part of the environmental information requirement. They underpin agricultural sustainability and food production, climate change and carbon sequestration, biodiversity conservation and the provision of ecosystem services, human health and infrastructure development (DAFF, 2011). Appropriate soil data and information are crucial to support evidence-based policy, planning and resource management decisions (Wilson, 2012).

For data to be useful, one basic requirement is that they be interpretable. Sufficient information should be provided to allow data to be unambiguously interpreted and used. Examples of such information include the location at which the soil was sampled, the property that was measured, the unit of measure, and quality assurance and quality control information. Furthermore, data should be easily integrated with other data sources, which is required in many modelling applications. For example, simulation of crop production may require, besides soil data, also weather, crop and fertilizer data.

To meet these requirements, we have developed a soil ontology for modelling soil information. We define the ontology by capturing the essential semantics of soil data in terms of concepts and relationships. We express the ontology using the Web Ontology Language (OWL)[1]. OWL is the language recommended by the World Wide Web consortium (W3C) for representing ontologies. The logical foundation of OWL is provided by description logics (DLs) (Baader et al., 2003). Being formal logic-based, OWL enables precise representations of knowledge. Moreover, one can reason about representations in OWL and check whether they are consistent or not. As such, OWL is an excellent candidate for a lingua franca in the soil domain, and able to inject the necessary automatization for soil data interpretation and integration.

In this paper, we focus on the design of the soil ontology and its potential applications. We describe the use of the ontology to facilitate data access by transforming data into ontology individuals and making them available as Linked Data (Bizer et al., 2009). We also discuss applications of the ontology for data integration, data classification and data validation.

The rest of the paper is organised as follows. Section 2 gives details of the soil ontology. Section 3 describes the use of the ontology for transforming soil data into Linked Data. Section 4 outlines several other applications of the ontology. And finally, Section 5 concludes the paper and points out future work.

2 THE SOIL ONTOLOGY

We construct the ontology by leveraging existing modelling efforts, including

- the Open Geospatial Consortium (OGC)'s O&M model (Cox, 2007a,b): a conceptual model for observations, measurements and sampling features.
- ANZSoilML[2]: a GML-based soil information model developed by CSIRO and Landcare Research (NZ).

During the construction, we also reuse existing ontologies for describing locations and time:

- the wgs84_pos vocabulary[3]: modelling anything in space and its geo-location.
- the W3C Time ontology[4]: modelling temporal units, temporal entities, instants, intervals, etc.

Figure 1 shows part of the ontology. A key concept of the ontology is `SoilFeature`, which represents a conceptual feature that is hypothesized to exist coherently in the world. We define `SoilProfileElement` as a type of `SoilFeature` and `SoilLayer` as a type of `SoilProfileElement`.

Related to `SoilFeature` are concepts such as `Metadata` which describes some contextual information of a `SoilFeature`, e.g. the data source information; `SoilClassification` which specifies the soil classification terms from an appropriate soil classification scheme (e.g. the Australian Soil Classification);

[1] `http://www.w3.org/TR/owl-ref/`
[2] `http://www.clw.csiro.au/aclep/ANZSoilML/trunk/docs/html/index.html`
[3] `http://www.w3.org/2003/01/geo/wgs84_pos`
[4] `http://www.w3.org/TR/owl-time/`

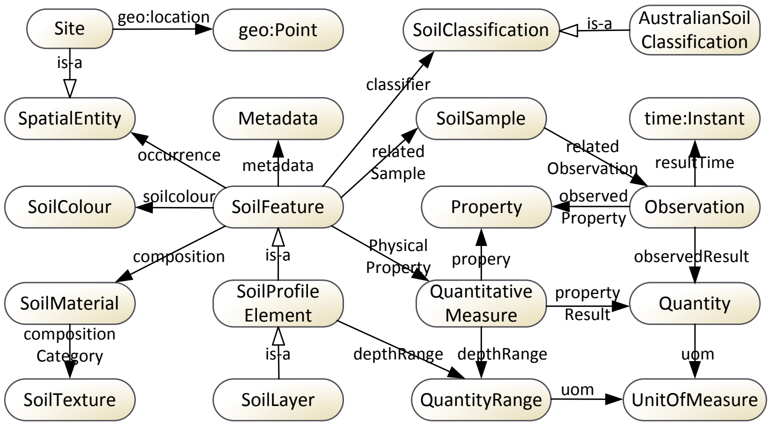

Figure 1. Portion of the soil ontology.

SoilColour which describes the soil colour; and SpatialEntity which describes the spatial information of a SoilFeature. We define Site as a subclass of SpatialEntity, and model its location information by geo:Point defined by wgs84_pos.

A SoilFeature may have associated observations and measurements. We capture this information using the concept QuantitativeMeasure, and associate a QuantitativeMeasure with a depth range QuantityRange, an observed Property and an observed result Quantity. Examples of soil properties include Bulk Density (BD), drained upper limit (DUL), saturation (SAT), pH and EC.

A SoilFeature may also have associated SoilSamples. Given a SoilSample, there may exist an Observation, which is associated with an observed Property, an observed result Quantity, and an observation time time:Instant defined by the W3C Time ontology.

3 MAKING SOIL DATA AVAILABLE AS LINKED DATA

In this section, we show how to facilitate soil data access by using the ontology described earlier. To facilitate data access and use, we would like to have data available as Linked Data. Linked Data refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets (Bizer et al., 2009).

We use APSoil as an example to illustrate our approach. APSoil[5] is a database of soil water characteristics enabling estimation of Plant Available Water Capacity (PAWC) for individual soils and crops. It covers many cropping regions of Australia and is designed for use in simulation modelling and agronomic practice. Figure 2 shows the interface of APSoil data.

To publish APSoil data as Linked Data, we map data to the ontology and translate data into a set of ontology individuals. During translation, we follow the Linked Data principles[6], i.e.

- Use URIs as names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL).
- Include links to other URIs. so that they can discover more things.

In particular, we employ the following URI scheme for the generated individuals:

- The base URI for all generated individuals is http://www.csiro.au/soil/resource/.

[5]https://www.apsim.info/Products/APSoil.aspx
[6]http://www.w3.org/DesignIssues/LinkedData.html

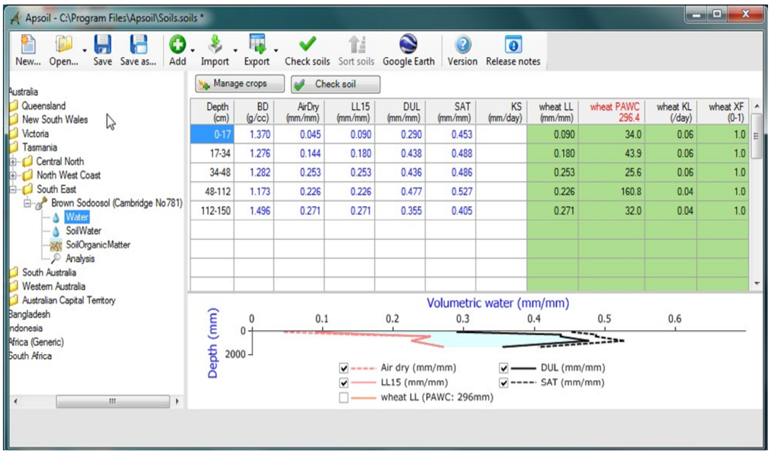

Figure 2. APSoil data.

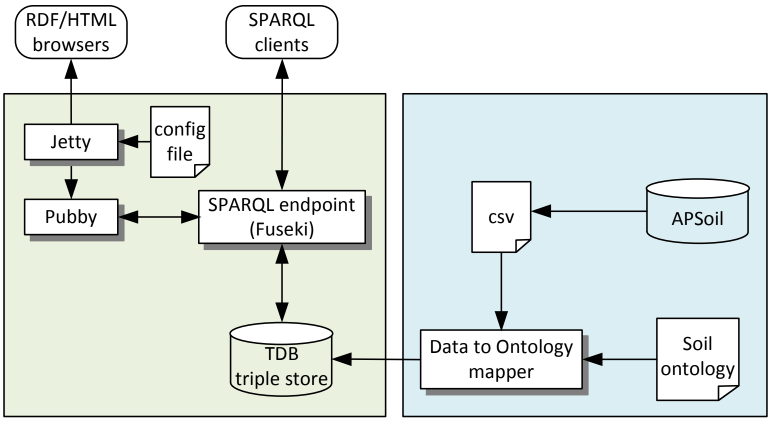

Figure 3. The system architecture for publishing APSoil data as Linked Data.

- The local ID of the URI of an individual is formed by concatenating the concept name and the unique APSoil number of each site, with a possible extension of the soil layer number or the observed property name if required.

We use patterned local IDs. The patterns used include *<concept name>_<APSoil number>*, *<concept name>_<APSoil number>_<soil layer number>*, and *<concept name>*_at_*<APSoil number>_<soil layer number>*_on_*<observed property name>*. Suppose a `SoilProfileElement` whose corresponding APSoil number is 781 and soil layer number is 1. Then we can identify it by `http://www.csiro.au/soil/resource/SoilProfileElement_781_1`. Suppose again a BD measurement was made on it. Then we can identify the measurement by `http://www.csiro.au/soil/resource/QuantitativeMeasure_at_781_1_on_BD`.

In addition, we link site information (i.e. country and state) to DBPedia[7]. We then store the generated RDF in a TDB[8] triple store. Finally, we expose data through a Fuseki[9]-facilitated SPARQL endpoint and integrate the endpoint with Pubby[10]. In this way, we publish data as Linked Data and make them available for humans and computers. Figure 3 shows the implemented system architecture.

There are two ways to access and explore data. One is to send SPARQL[11] queries to the endpoint. For example,

[7] `http://dbpedia.org/`
[8] `https://jena.apache.org/documentation/tdb/`
[9] `https://jena.apache.org/documentation/fuseki2/`
[10] `http://www.wiwiss.fu-berlin.de/pubby/`
[11] `http://www.w3.org/TR/sparql11-query/`

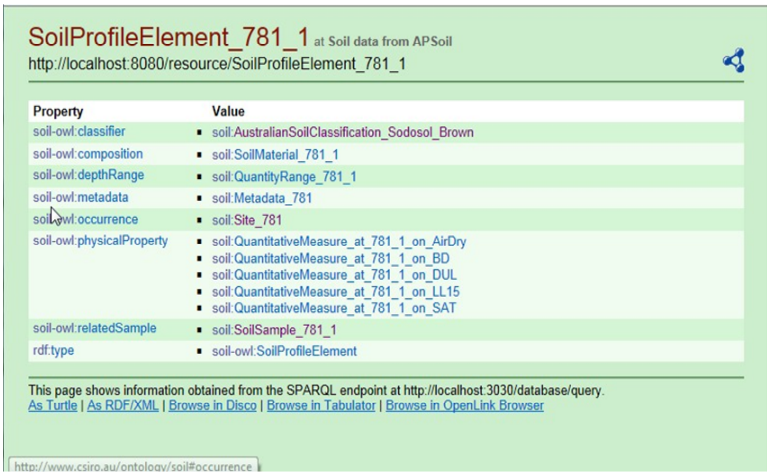

Figure 4. The Linked Data browser facilitated by Pubby.

to retrieve the latitude and longitude information of all Tasmanian sites, we can submit the following query:

```
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX soil:  <http://www.csiro.au/soil/resource/>
PREFIX soil-owl:  <http://www.csiro.au/ontology/soil#>
PREFIX geo:  <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX dbpedia:  <http://dbpedia.org/resource/>
SELECT DISTINCT ?point ?lat ?long
WHERE{
   ?x rdf:type soil-owl:SoilProfileElement.
   ?x soil-owl:occurrence ?site.
   ?site soil-owl:state dbpedia:Tasmania.
   ?site geo:location ?point.
   ?point geo:lat ?lat.  ?point geo:long ?long.
} ORDER BY ?point
```

Alternatively, one can access data through a RDF/HTML browser facilitated by Pubby which adds a Linked Data interface to the endpoint as shown in Figure 4.

4 OTHER POTENTIAL APPLICATIONS OF THE SOIL ONTOLOGY

Besides data access, we can also use the ontology for data integration, data classification and data validation. In this section, we discuss applications of the ontology in these aspects.

4.1 Data Integration

One challenge in integrating multiple data sources is resolving heterogeneity issues. Data from different sources typically have different formats or structures. Also, it is common for different sources to use different terminologies and value representations. To address these issues, we can use the ontology as the global model and map data to the ontology. In doing so, we move from different structures, terminologies and value representations of data to *structurally flat*, *precise* and *consistent* representations of data.

Suppose that in one data source, BD was measured in g/cc, while in another source BD was measured in kg/cc. In such a case, we can use the ontology to specify the standard unit of measure for BD, and meanwhile model the relationship between g/cc and kg/cc, including how one can be converted into another. In this way, we can support automatic unit conversion during data translation or query processing and thus have consistent value representations for BD.

4.2 Data Classification

The basic reasoning tasks for OWL are checking whether a concept is subsumed by another concept, or whether an individual is an instance of a concept. We can automate these reasoning tasks using standard OWL reasoners such as FACT++[12] and Pellet[13] (Sirin et al., 2007).

Suppose we want to classify soils into 3 groups: acidic (pH < 5.5), neutral ($5.5 \leq$ pH ≤ 8), alkaline (pH > 8). To do this, we introduce 3 concepts to the ontology, i.e. `AcidicSoil`, `NeutralSoil` and `AlkalineSoil`, and define the following SWRL[14] rules:

(a) `SoilFeature(?s)` ∧ `QuantitativeMeasure(?p)` ∧ `physicalProperty(?s, ?p)` ∧ `propertyName(?p, pH)` ∧ `propertyResult(?p ?r)` ∧ `Quantity(?r)` ∧ `value(?r, ?v)` ∧ `swrlb:lessThan(?v, 5.5)` → `AcidicSoil(?x)`

(b) `SoilFeature(?s)` ∧ `QuantitativeMeasure(?p)` ∧ `physicalProperty(?s, ?p)` ∧ `propertyName(?p, pH)` ∧ `propertyResult(?p ?r)` ∧ `Quantity(?r)` ∧ `value(?r, ?v)` ∧ `swrlb:greaterThanOrEqual(?v, 5.5)` ∧ `swrlb:lessThanOrEqual(?v, 8)` → `NeutralSoil(?x)`

(c) `SoilFeature(?s)` ∧ `QuantitativeMeasure(?p)` ∧ `physicalProperty(?s, ?p)` ∧ `propertyName(?p, pH)` ∧ `propertyResult(?p ?r)` ∧ `Quantity(?r)` ∧ `value(?r, ?v)` ∧ `swrlb:greaterThan(?v, 5.5)` → `AlkalineSoil(?x)`

Then an OWL reasoner (with the SWRL support) will automatically classify soil features into right groups.

4.3 Data Validation

Data validation is an important part of data integration and analysis tasks. In a data validation application, complete knowledge is typically assumed for some or all parts of the application domain. This closed-world assumption, however, contradicts to the open-world semantics of OWL which has the following characteristics (Tao et al., 2010):

- the presence of the Open World Assumption(OWA): a statement cannot be inferred to be false on the basis of failures to prove it.
- the absence of the Unique Name Assumption(UNA): two different name may refer to the same object.

With such a semantics, conditions that trigger constraint violations in closed world systems would generate new inferences in standard OWL systems. To use OWL for the data validation purpose, one way is to employ a closed-world semantics for OWL, and interpret OWL axioms based only on the information explicitly present in the ontology. As such, we can reduce constraint checking to SPARQL query answering by translating each constraint represented as an OWL axiom into a SPARQL ASK query such that when the constraint is violated, the query is entailed, i.e. a non-empty result is returned (Tao et al., 2010).

Suppose that we want to ensure that every soil feature has at least one classifier. We can express this constraint by the axiom: $\texttt{SoilFeature} \sqsubseteq \geq 1\ \texttt{classifier.SoilClassification}$, and then check satisfaction of the constraint by translating it into the following SPAQRL query:

```
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX soil-owl:  <http://www.csiro.au/ontology/soil#>
ASK WHERE {
   ?x rdf:type soil-owl:SoilFeature.
   FILTER NOT EXISTS {?x soil-owl:classifier ?y.
                      ?y rdf:type soil-owl:SoilClassification.}
}
```

When the query is executed, a non-empty result indicates that the constraint is violated, i.e. there exists a soil feature without any classifier.

[12] `http://owl.man.ac.uk/factplusplus/`
[13] `http://pellet.owldl.com/`
[14] `http://www.w3.org/Submission/SWRL/`

5 CONCLUSIONS

In this paper, we presented a soil ontology and discussed its potential applications. In particular, we described the use of the ontology to facilitate data access by transforming data into Linked Data. Our ongoing and future work includes further development of the ontology (e.g. extending it with plant ontology), using it to link more datasets, and applying it in agricultural applications.

REFERENCES

Baader, F., D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider (2003). The Description Logic Handbook: Theory, Implementation, and Applications. In *Cambridge University Press*.

Bizer, C., T. Heath, and T. Berners-Lee (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems 5*(3), 1–22.

Cox, S. (2007a). Observations and Measurements - Part 1 - Observation schema Version 1.0 OGC. In *OGC document 07-022r1*.

Cox, S. (2007b). Observations and Measurements - Part 2 - Sampling Features Version 1.0 OGC. In *OGC document 06-188r1*.

DAFF (2011). A Stocktake of Australias Current Investment in Soils Research, Development and Extension: A Snapshot for 2010-11. Department of Agriculture, Fisheries and Forestry.

Sirin, E., B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz (2007). Pellet: A Practical OWL-DL Reasoner. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web 5*(2), 51–53.

Tao, J., E. Sirin, J. Bao, and D. L. McGuinness (2010). Integrity Constraints in OWL. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI'10)*.

Wilson, P. (2012). Improving Australian Soil Data and Information Governance. *CSIRO Technical Report*.