* Integration of Plot-based Ecological Data: A Semantic Approach

Siddeswara Guru[1[0000-0002-3903-254X]] Simon Cox[2[0000-0002-3884-3420]] Edmond Chuc[1[0000-0002-6407-9864]] Tina Schroeder[1[0000-0002-8481-9069]] Mosheh Eliyahu[1] Yi Sun[1] Jenny Mahuika[1] and Alvin Sebastian[1]

[1] Terrestrial Ecosystem Research Network (TERN), University of Queensland, St Lucia, 4072 Australia
[2] Environmental Informatics, CSIRO, Clayton, 3168, Australia
s.guru@uq.edu.au

Abstract. There is a large amount of plot-based ecological data collected by different agencies and at different jurisdictions. Often, data are collected using varying survey methods and procedures, even though observed properties are similar. Typically, use of these data for analysis is confined to a jurisdiction from where the data was collected. However, integration of these datasets would enable their use at a larger scale for analysis and synthesis. In this paper, we will describe a Semantic Web approach to integrate plot-based ecological data from different agencies in Australia. We will also discuss some of the initial implementation progress.

Keywords: Plot-based Ecological Data, Ontology, Data Integration, Semantic Web, Linked Data, Interoperability, SKOS, Controlled Vocabularies.

1 Introduction

There is a significant amount of data collected to monitor the environment by measuring biodiversity and ecological processes at a certain point in time and space. Plot-based monitoring is used to survey soil properties, vegetation, animal populations and ecosystem processes by using repeatable methods and procedures. Recurring measurements of the same observed properties would enable us to study the long-term impact of on-going environmental and resource management practices. However, even if the measurement is one-off, the observations will act as an inventory for flora and fauna species and ecological processes at a site. Generally, all these data collections are project-based, collected for a specific purpose, and use comparable monitoring methodologies at different plots, covering several geographical locations. These datasets become much more useful once they are integrated with other similar projects or programs at a larger scale.

In Australia, every state and territory collects and publishes plot-based vegetation and ecological data to meet the legislative requirements, including reporting to the Environmental Protection and Biodiversity Conservation (EPBC) Act and the State of the Environment (SoE) every five years. The integrated access of these datasets is useful for analysis at a national scale and will be a gateway to access harmonized plot-based ecological data from multiple agencies and data providers.
In this paper, we will discuss approaches we have taken for the integration of plot-based ecological data to enable unified search and access to data from different projects/programs, jurisdiction, observation themes, observable properties, survey methods, temporal scales, and taxonomies.

2 Solution Approach

We have proposed a Semantic Web data integration approach to address the challenges to integrate plot-based ecological data. This will allow us to organize all the data using the Resource Description Framework (RDF) [1] irrespective of the structure of the underlying source data. The proposed data integration solution uses a hybrid approach where domain-related terms in each of the data sources are mapped to a shared ontology and data source-centric terminologies and methodologies are built as controlled vocabularies using the Simple Knowledge Organization System (SKOS).

2.1 TERN-Plot Ontology

The TERN-Plot ontology is derived from the Observations and Measurements (O&M) and the Semantic Sensor Network (SSN) ontology. The core structure of the TERN-Plot ontology consists of classes and properties to describe plots, sampling activities that happen within a plot, and an observation or collection of observations which have procedures to produce results. The TERN-Plot ontology has dependencies on RDF [2], RDFS [3], Dublin core [4], SSN/SOSA [5], Darwin Core [6], and GeoSPARQL [7]. The official documentation of the TERN-Plot ontology is available at http://www.linked.data.gov.au/def/plot/.
The TERN-Plot ontology uses O&M to capture the observation elements; and eight new classes to describe the ecology-plot domain. The ontology supports linkages between a *feature of interest* to observations. All of the scientific details, and most of the bio-physical descriptions are captured as the results of *observations* on the features. Each observation relates to one *observable property*. The set of *Observable Properties* is maintained as a controlled vocabulary. This complements the feature-based domain model to complete the TERN-Plot ontology.

2.2 Controlled Vocabularies

A small number of controlled vocabularies are required to provide values for properties in the TERN-Plot ontology. Eventually, these will be managed as a whole-of-project set of vocabularies, and applied to the data from all data providers. Following are some

of the vocabularies that will be captured: Observable Properties – range value of *sosa:observedProperty* on Observations; Observation Group classifiers – for *dct:type* on ObservationCollections; Procedure – range value of *sosa:usedprocedure* on methods, and Interim Biogeographic Regionalization for Australia (IBRAs) – range value of *sosa:isSampleOf* on Sites (unless a more refined value is available). For instance, for observable property Erosion Severity, the value space (i.e. permitted values) of the observation result is provided by another controlled vocabulary managed as categorical variables Erosion Severity Values. The procedure used for each observation will also be managed as a controlled vocabulary (e.g. Methods). Vocabulary terms are aligned with international vocabulary EnvThes (Environmental Thesaurus).
Vocabularies are defined for each data provider, except for universal terms like unit of measure, place, jurisdictions, etc. The controlled vocabularies are identified initially from 'authority tables' in the various providers databases. Contents are converted to an RDF representation using the appropriate type through an ETL pipeline developed in Python.

2.3 Architecture and Early Implementation

Figure 1 provides a high-level architecture of the overall platform. The implementation of the platform is under progress. However, a prototype of data mapping and transformation has been implemented for CORVEG dataset, a database containing vegetation and soil observations from the Queensland state government as a proof-of-concept. Typically, data is received as a collection of CSV files with each file associated to a relational database table from government enterprise databases. The source data is mapped to the TERN-Plot ontology using an ETL process which uses Apache Spark, Python, RDFLib, and a host of other technologies to efficiently transform plot datasets from its tabular form into RDF. The RDF data is validated using SHACL to ensure its compliance with the TERN-Plot ontology. Qualitative value range for validation includes observed properties, geospatial coordinates, taxonomy names, dates and units of measure. Quantitative data are also validated. Once the transformed data passes all validation, they are ingested into a GraphDB triple-store instance. The transformation process is documented in https://ternaus.atlassian.net/wiki/spaces/CDI/overview.

3 Conclusion

In this paper, we have discussed challenges in the integration of plot-based ecology data and proposed semantic web solution. We have developed ontology based on O&M and SSN ontology to conceptualize ecology plots and use controlled vocabularies to represent database centric terms. The proposed approach provides flexibility to map diverse data terms. We have built a prototype dashboard for users to interact and extract data. In our future work, we are focusing on ingesting more databases and improve ETL as an automated process. We are keen to provide users with a flexible interactive dashboard to query and access any observations with all contextual information.

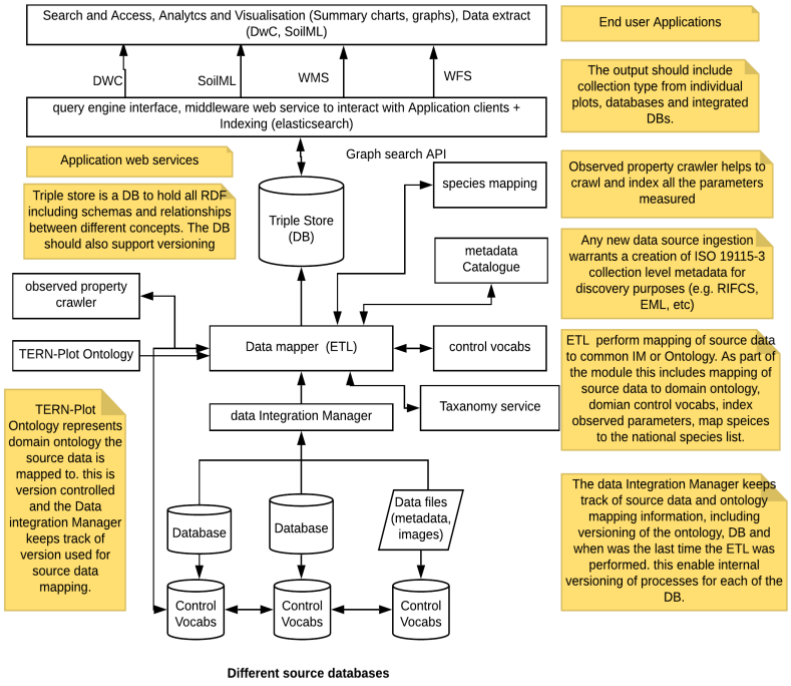

Fig. 1. Plot-based ecological data integration platform, showing source data at the bottom with source-level controlled vocabularies, data mapper performing ETL to a common TERN-Plot ontology and eventually storing the data into the triple-store with Elastic search driving the business logic for the data access and visualization.

References

1. Cyganiak, R., Wood, D. & Lanthaler, M. RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation (2014)
2. Schreiber, G. & Raimond, Y. RDF 1.1 Primer. *W3C Recommendation* (2014).
3. Brickley, D. & Guha, R. V. RDF Schema 1.1. (2014).
4. DCMI Usage Board. DCMI Metadata Terms. (2012). Available at: http://www.dublincore.org/specifications/dublin-core/dcmi-terms/. (Accessed: 19th June 2019)
5. Haller, A. *et al.* Semantic Sensor Network Ontology. *W3C Recommendation* (2017).
6. Wieczorek, J. *et al.* Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS One* **7**, e29715 (2012).
7. Perry, M. & Herring, J. OGC GeoSPARQL – a geographic query language for RDF data. *OGC 11-052r4* (2012).