Establishing a Linked Data Infrastructure for the OGC Body of Knowledge

Gobe Hobona[1][0000-0002-8733-4702] , Rob Atkinson[1][0000-0002-7878-2693], Greg Buehler[1][0000-0003-1386-69], Scott Simmons[1][0000-0002-9085-010X], and Ingo Simonis[1][0000-0001-5304-5868]

[1] Open Geospatial Consortium, Wayland, MA, USA
ghobona@ogc.org

Abstract. The OGC Body of Knowledge is a structured collection of concepts and related resources that can be found in the set of documents published by the Open Geospatial Consortium (OGC). An explicit view of this knowledge is available from the OGC Virtual Knowledge Store and related components such as the OGC Definitions Server and the OGC Glossary of Terms. The OGC Body of Knowledge is intended to provide a reference for users and developers of geospatial software and services. This paper describes the approach taken to develop the OGC Body of Knowledge and presents the results of the approach. It is intended to encourage and facilitate discussion within the OGC membership and wider geospatial community.

Keywords: linked data, body of knowledge, knowledge management

1 Introduction

The Open Geospatial Consortium (OGC) is an international consortium of more than 520 businesses, government agencies, research organizations, and universities driven to make geospatial (location) information and services findable, accessible, interoperable, and reusable (FAIR). The OGC Body of Knowledge is a structured collection of concepts and related resources that can be found in the OGC Library [1]. It is, in effect, a view of explicit knowledge available from the OGC Virtual Knowledge Store and related components such as the OGC Definitions Server (http://www.opengis.net/def/) and the OGC Glossary of Terms (https://www.ogc.org/ogc/glossary). A Body of Knowledge exists beyond the boundaries of knowledge assets because of the tacit practices, skills, experiences, products, processes, and interdisciplinary knowledge that define the field that are incorporated into that body of knowledge [2]. This paper describes the approach taken to develop the OGC Body of Knowledge.

2 Methodology

2.1 Development Process

To develop the OGC Body of Knowledge, the OGC designed an approach that involved extraction of knowledge from a prior-existing OGC knowledge base, transformation of the knowledge into formal statements and then loading of the statements into persistent storage. The formal statements are represented as triples and persisted in a triple store that was named the Virtual Knowledge Store (VKS).

The content stored in the VKS can be queried at any time by applications by running searches on the SPARQL interface. The SPARQL interface is implemented through a deployment of the RDF4J workbench (https://rdf4j.org). For human users, a document is generated from the triples that have been stored in the VKS. The document is serialized as asciidoc and then compiled using asciidoctor software (https://asciidoctor.org). The main presentation approach for the VKS is a Linked Data approach supporting HTTP content-negotiation, with HTML and SKOS (RDF-encoded) the key formats supported. Flexibility in terms of providing multiple alternative views using different data models is supported through "Content Negotiation by Profile" [3].

2.2 The Information Model

The triples representing the OGC Body of Knowledge in the VKS implement the Simple Knowledge Organization System (SKOS) standard of the World Wide Web Consortium (W3C) (http://www.w3.org/TR/skos-reference). SKOS is designed to enable concepts to be composed and published on the World Wide Web, linked with data on the Web and integrated into concept schemes. The data model defined by SKOS specifies Concept, ConceptScheme and Collection types that together make it possible to represent knowledge organization systems such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other types of vocabularies.

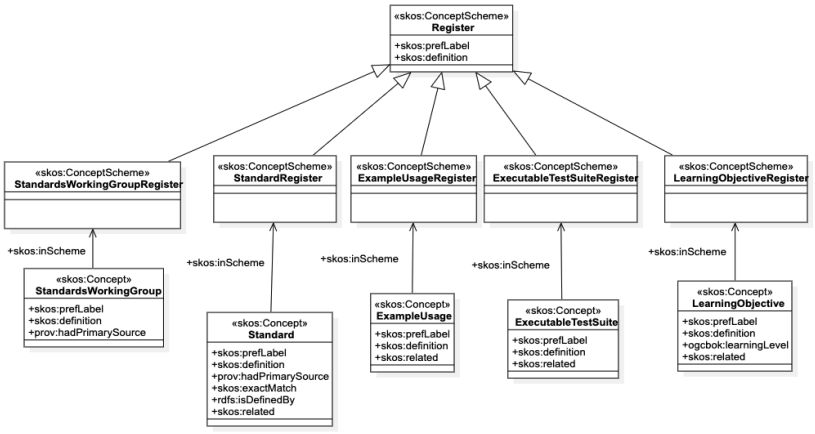

Fig. 1. UML Class model of the SKOS representation of element types in the OGC Body of Knowledge.

SKOS is encoded as an application profile of the Resource Description Framework (RDF). As illustrated in Figure 1 above, the information model made use of the **skos:Concept** class which can represent a notion of a thing, **skos:ConceptScheme** class which represents an aggregation of concepts, and SKOS semantic relations which are links between SKOS concepts where the link is inherent in the meaning of the linked concepts. For example, **skos:related** was used to indicate the relationship between a Standard and Learning Objectives, Executable Test Suites, Example Usage and a Standards Working Group.

2.3 Context within OGC Infrastructure

The OGC Body of Knowledge sits within the Member Support part of OGC infrastructure, as shown in Figure 2. The blue-filled boxes are those that have been implemented to date, and the grey-filled boxes are those that are under development. The OGC Body of Knowledge is represented in the diagram with a black-filled box. Some of the boxes are labelled as TBA (To be added) to indicate that more content is yet to be added to the VKS. The figure also highlights that the body of knowledge is one of a series of tools intended to support OGC Members.

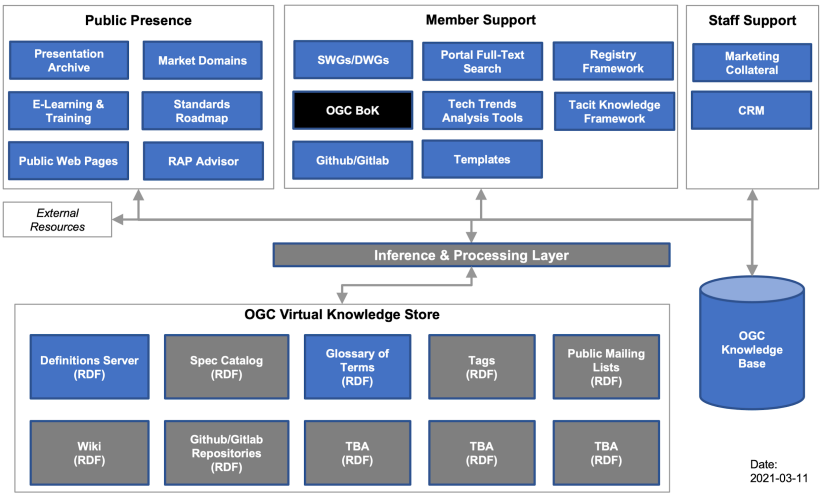

Fig. 2. A high level overview of the infrastructure within which the OGC Body of Knowledge (BoK) sits.

The Inference and Processing Layer shown in Figure 2 represents a series of tools that apply a semi-automated approach to processing of content (i.e., involving a combination of automated and manual steps) in order to build the body of knowledge. The steps can be summarized as follows:

Step 1. Configure extraction rules for the types of documents that are to be processed (manual)

Step 2. Configure rules for cross-referencing extracted content to third party content (manual)
Step 3. Extract the content from the documents (automated)
Step 4. Transform the extracted content into triples (automated)
Step 5. Review the triples to verify that they are correct (manual)
Step 6. Load the triples into the triple store (automated)
Step 7. Optionally generate an asciidoc document from the triples (automated)

One of the areas where manual processing became unavoidable was the cross-referencing of content from the OGC Library with content from 3rd parties. For example, several OGC standards identify media types (formerly known as MIME type) that are to be used with specific encodings. As the global register of media types is maintained by the Internet Assigned Numbers Authority (IANA), it was necessary to manually review the OGC Standards and cross-reference them to the appropriate IANA resources. Once codified as triples, that information could then be queried by an application.

3 Results

3.1 Extraction

Extraction of content from recent documents (post-2013) was found to be generally easier to automate than extraction from historical documents. This is because recent documents published by the consortium were published in both HTML and PDF format, whereas prior to 2013 OGC documents were mainly published in PDF format. HTML documents offered more structure to the content than text extracted from PDF documents. Therefore, a semi-automated approach was seen as appropriate for processing of content. Some of the tools that helped with the extraction of content were:

- Apache POI (https://poi.apache.org): A toolkit for creating and maintaining content in several different Open Office related formats.
- Apache Tika (https://tika.apache.org): A toolkit for detecting and extracting metadata and text from over a thousand different file types (such as PPT, XLS, and PDF).
- Dstl Baleen (https://github.com/dstl/baleen): An extensible text processing capability that allows entity-related information to be extracted from unstructured and semi-structured data sources.

3.2 Loading

The extracted triples are imported into the VKS through the Definitions Server's Administration tool that is built on the Django framework. Django is a Python-based free and open-source web framework that enables developers to implement solutions based on the model-template-view architectural pattern that separates the concerns of an application's model from its views. The Administration tool stores a snapshot of the SKOS file and then transmits the triples to an RDF4J Workbench instance which

exposes the triples through a SPARQL interface. A screenshot of the Administration tool is shown in Figure 3.

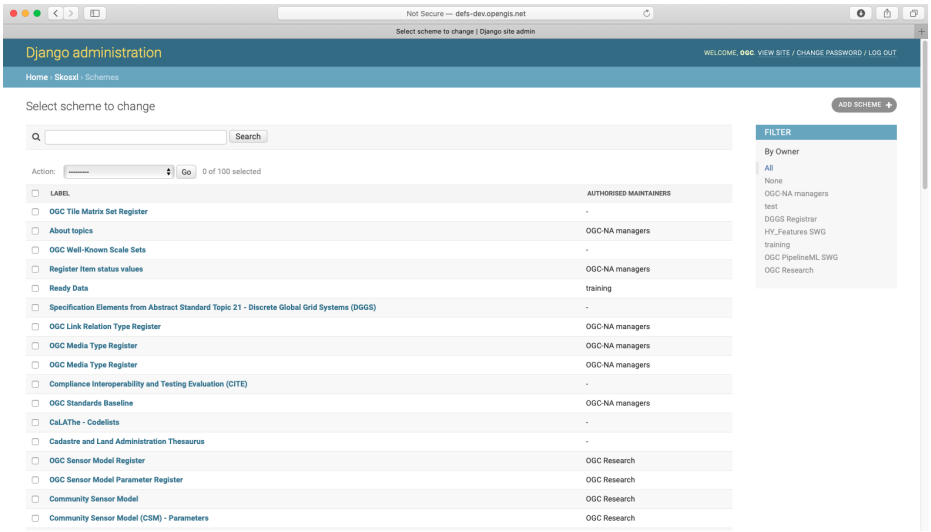

Fig. 3. A screenshot of the Definitions Server's Administration tool.

The Administration tool provides a range of functions to support consistency of the VKS:

- entailment to standardise views of SKOS data
- specialised importers for various canonical forms of content, such as XML based forms and dictionaries encoded in Geography Markup Language (GML)
- option for manual entry
- association of metadata to content sources
- assignment of governance rights
- review mode (publishing to staging repositories)
- batch loading facilities

3.3 Governance

A key part of ensuring that the body of knowledge can be maintained long term is the establishment of a process for governance of information content. OGC already had a sub-committee, called the OGC Naming Authority, that was responsible for managing content in the Definitions Server and the Glossary of Terms. Therefore, the scope of the Naming Authority was expanded to include management of the OGC Body of Knowledge. This meant that the item registration process adopted by the Naming Authority could then also be applied to the body of knowledge. An illustration of the item registration process, which is based on the one specified in the ISO 19135 standard (https://www.iso.org/standard/54721.html), is shown in Figure 4.

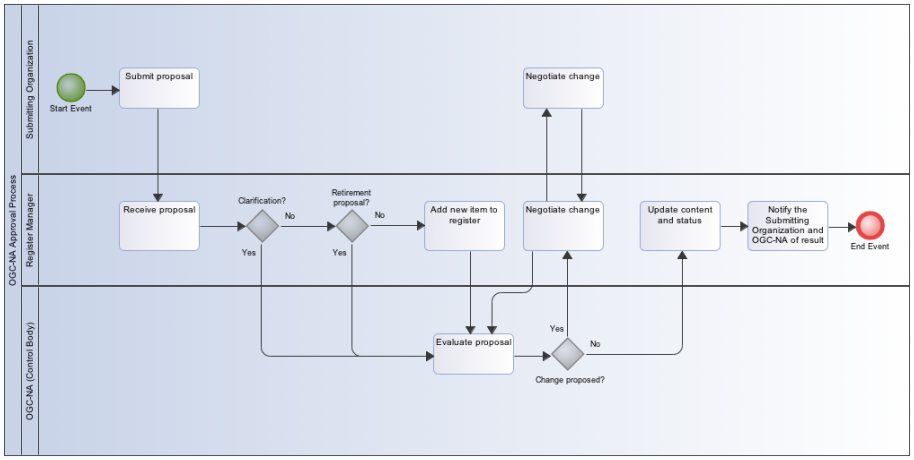

Fig. 4. Item registration process adopted by the OGC Naming Authority (adapted from ISO 19135).

4 Discussion

4.1 Separation of Content from Presentation

All of the explicit knowledge that is presented in the OGC Body of Knowledge is formally captured in the VKS as knowledge graphs consisting of concepts and their relationships. This approach makes it possible to separate content from presentation, or more specifically, it makes the explicit knowledge independent of the medium through which it is presented. The rationale for this separation of concerns between content and presentation is that the explicit knowledge that is available in the VKS can be reused by other systems to address other needs that may not have been foreseen.

Separation of the VKS into separate knowledge graphs preserves the data provenance and supports batch replacement of modules with updates. Updates may add additional metadata, container (grouping) views, links to related concepts, etc. URIs and content remain stable, with status flags used to deprecate superseded content.

Using Linked Data and Content Negotiation by Profile approaches, graphs can be accessed in various forms by setting appropriate access headers. Humans using a Web Browser can access details of individual terms, or link to access the containing graph objects via the SKOS ‘inScheme” predicate link. Applications can access the RDF-encoded SKOS content from the same URIs, using different access headers.

4.2 Coverage

The current scope of the OGC Body of Knowledge is limited to introductions to concepts in OGC standards, learning objectives related to those standards, references to compliance tests for those standards, and examples of innovation initiatives that

reference the standards. As such the OGC Body of Knowledge serves as an informative reference, and is not intended to replace any OGC standard. In fact, it is intended to help direct a reader towards relevant OGC standards where normative content can be found. Moreover, the OGC Body of Knowledge is intended to be complementary to other educational resources such as the OGC e-Learning resources, and the OGC website.

4.3 Future

The architecture and resource model for the OGC Body of Knowledge permits other web-accessible knowledge resources to be linked to or enhanced by this content. This was successfully demonstrated through the integration of the Cadastre and Land Administration Thesaurus (CaLAThe) into the OGC Definitions Server – a part of the VKS [4]. OGC and its partner organizations are working to establish such links to improve the consistency of geospatial knowledge and ensure that the most authoritative source for any particular type of knowledge can be easily discovered and reused.

Another area identified for future work is that of automation. Natural Language Processing (NLP) tools, such as OpenNLP, enable the parsing of sentences. Therefore, an ability to identify elements of sentences in natural language text could potentially support the automatic formulation of statements as triples. Another area for future work is the harmonization of vocabularies for registers. Such harmonization would have to be based on Linked Data technologies to ensure that the vocabularies can be used by the wider Semantic Web.

5 Conclusions

The combination of SKOS and asciidoctor proved to be effective at enabling the representation of the OGC Body of Knowledge in different forms to facilitate both machine and human interpretation. Whereas the development of the OGC Body of Knowledge will continue for some time to come, the approach taken thus far has shown that Linked Data technologies such as SKOS and SPARQL services can aid the development of such bodies of knowledge.

References

1. Hobona, G.: OGC Body of Knowledge - Version 0.1 - Discussion Paper. Open Geospatial Consortium, (2020) https://docs.opengeospatial.org/dp/19-077.html
2. Hart, H., Baehr, C.: Sustainable Practices for Developing a Body of Knowledge. Technical Communication 60(4), 259-266 (2013)
3. Svensson, L.G., Atkinson, R., Car, N.: Content Negotiation by Profile, W3C Working Draft, Data Exchange Working Group, 2019, https://www.w3.org/TR/dx-prof-conneg/
4. Stubkjær, E., Çağdaş, V.: Alignment of standards through semantic tools – The case of land administration. Land Use Policy. 104, 105381 (2021).