Linked Data Practice at Different Levels of Semantic Precision: The Perspective of Libraries, Archives and Museums

by Antoine Isaac and Thomas Baker

EDITOR'S SUMMARY

Libraries, archives and museums rely on structured schemas and vocabularies to indicate classes in which a resource may belong. In the context of linked data, key organizational components are the RDF data model, element schemas and value vocabularies, with simple ontologies having minimally defined classes and properties in order to facilitate reuse and interoperability. Simplicity over formal semantics is a tenet of the open-world assumption underlying ontology languages central to the Semantic Web, but the result is a lack of constraints, data quality checks and validation capacity. Inconsistent use of vocabularies and ontologies that do not follow formal semantics rules and logical concept hierarchies further complicate the use of Semantic Web technologies. The Simple Knowledge Organization System (SKOS) helps make existing value vocabularies available in the linked data environment, but it exchanges precision for simplicity. Incompatibilities between simple organized vocabularies, Resource Description Framework Schemas and OWL ontologies and even basic notions of subjects and concepts prevent smooth translations and challenge the conversion of cultural institutions' unique legacy vocabularies for linked data. Adopting the linked data vision requires accepting loose semantic interpretations. To avoid semantic inconsistencies and illogical results, cultural organizations following the linked data path must be careful to choose the level of semantics that best suits their domain and needs.

KEYWORDS

linked data	precision	archives
ontologies	SKOS	museums
semantics	RDF	museums

Linked Data and the Charm of Weak Semantics

Linked data for libraries, museums and archives (LAM) draws on and integrates a broad variety of structured schemas and vocabularies, some of which have developed over a period of decades. In the linked data environment, all such semantic artifacts are expressed in RDF (Resource Description Framework) as *data*, and in the RDF context most of these types of data may also be called *vocabularies*. From the perspective of LAM practitioners, such semantic artifacts present themselves broadly as *element schemas*, which define the basic properties and classes according to which the data is structured, and *value vocabularies*, which provide pragmatically organized structures of concepts to which the resources they describe are related.

In the LAM context, the semantic level of those artifacts may differ significantly according to specific application requirements. LAM practitioners, for example, may work with ontologies that situate a class Person in a precisely defined hierarchy of classes, but it may also be necessary to distinguish a person's identity as defined by a specific national library. Large, organically evolved schemes of broader and narrower concepts, moreover, may not lend themselves to expression as precise hierarchies of classes.

This article briefly surveys the types of semantic artifacts of relevance to LAM practice and draws a few conclusions about the possibility and desirability of expressing their semantics with precision.

The Formal Classes of RDFS and OWL Ontologies

The Resource Description Framework (RDF) data model that forms the base of Semantic Web technology and linked data relies on a formal typing

Antoine Isaac is R&D manager for Europeana and guest researcher at the Free University Amsterdam. He can be reached at antoine.isaac<at>europeana.eu
Thomas Baker, an organizer of the Dublin Core Metadata Initiative, is an associate professor at Sungkyunkwan University in Seoul, South Korea. He can be reached at tb12<at>thbaker.org

mechanism, where resources are asserted to belong to classes (such as ex:Painting) of the type defined by the RDF Schema specification (RDFS) (www.w3.org/TR/rdf-schema/):

```
@prefix ex: <http://example.org/>
@prefix rdf: <www.w3.org/1999/02/22-rdf-syntax-ns#>
ex:monalisa rdf:type ex:Painting
```

The classes and properties used to produce such RDF statements may be defined using formal axioms in the Web Ontology Language OWL (www.w3.org/TR/owl2-overview/), for example to say that every Book has at least one creator. In the RDFS and OWL model, classes are organized in subsumption hierarchies, for example:

```
@prefix rdfs: <www.w3.org/2000/01/rdf-schema#>
ex:RenaissancePainting rdfs:subClassOf ex:Artwork
ex:Painting rdfs:subClassOf ex:Artwork
```

Such hierarchies are transitive. From the two axioms above, for example, it follows that every resource with rdf:type set to ex:RenaissancePainting can be automatically classified as an ex:Artwork.

It is possible to add a great deal of detail with formal axioms in OWL ontologies. This facility brings more modeling precision and the ability to perform more complex automatic reasoning. However, this precision comes at the price of hard work and has hindered the adoption of formal ontology languages. Too often, would-be technology adopters have felt they needed to capture all semantics of their domain, which requires a significant amount of time and raises the bar for potential data re-users who have to handle data with more complex data semantics.

One of the reasons for the success of the linked data movement has been its focus on lighter-weight ontologies, where classes and properties are defined using minimal axioms. This minimalism makes it easier to re-use vocabularies across applications within a domain or even across domains, which makes the data that uses them more interoperable. In the cultural sector, the Dublin Core vocabulary (http://dublincore.org/documents/dcmi-terms/) remains very popular for expressing metadata as RDF. Together with other relatively simple ontologies, the Dublin Core elements provide the basis for the European Data Model (www.europeana.eu/schemas/edm/) used by Europeana (http://europeana.eu), the Digital Public Library of America (http://dp.la) and the German Digital Library (www.deutsche-digitale-bibliothek.de/) for gathering and publishing metadata for thousands of digitized cultural collections. Even more sophisticated models in the cultural sector may avoid coining too many formal axioms. The CIDOC Conceptual Reference Model (www.cidoc-crm.org/), for example, is provided as a simple RDFS ontology.

Another example is Schema.org (http://schema.org), a broadly scoped but semantically lightweight vocabulary developed by Google, Bing, Yahoo and Yandex to harvest data about the content of websites. The use of Schema.org is being explored for cultural data too, as in recent experiments led by OCLC [1].

The creation and use of formal ontologies is also unexpectedly hindered by the open-world assumption (http://en.wikipedia.org/wiki/Open-world_assumption) that underpins the formal semantics of current ontology languages. According to this assumption, if an OWL ontology states that every painting is expected to have one creator, a dataset where a specific painting does not have any creator will not be considered wrong. There could in principle be data elsewhere that provides the missing information. A reasoning engine interpreting this data will infer that the painting has a creator even if that creator is unknown.

The open-world assumption is a very powerful principle for web data. Yet it is counter-intuitive for many data modelers accustomed to more traditional, data validation systems such as the ones based on XML schema or used in relational databases. It also implies that existing Semantic Web technology does not provide a mechanism to close the world and perform the sort of data checking that would be useful for ensuring a minimal quality for data being exchanged. Europeana, for instance, had to fall back on basic XML schema validation rules to specify how the Europeana Data Model (EDM) data should be validated. Now that the community has had a taste of web- and graph-based data representation, this necessity has felt like quite a step backwards.

Such requirements for data checking have recently come to the fore, and efforts have started recently at W3C to develop technology for making and sharing specifications (data shapes) that would allow such checking (www.w3.org/2014/data-shapes/). A task group of the Dublin Core Metadata Initiative has also been created to identify the requirements for RDF Application Profiles in the Dublin Core community and make sure these requirements are taken on board in the W3C efforts (http://wiki.dublincore.org/index.php/RDF_Application_Profiles).

Vocabularies from Formal OWL Ontologies Down to SKOS Concept Schemes

Another hurdle for the practical deployment of Semantic Web technology in the cultural domain has been the use of the words *ontology* and *vocabulary* to describe quite different kinds of artifacts. For example, recent W3C documents (www.w3.org/standards/semanticweb/ontology) define *vocabulary* to encompass the formal sort of RDFS and OWL ontologies described above. However, the notion of vocabulary also covers a wide range of knowledge organization systems (KOS) used in libraries, archives and museums, such as thesauri, classification systems and name authority files. Such value vocabularies can also have structure. Concepts from thesauri, for example, are often organized using hierarchical and associative links as defined in standards such as ISO 25964 [2]. These relationships, however, do not bear precise formal semantics in the sense of OWL. Indeed, some hierarchies appear quite wrong when interpreted according to a formal approach, especially if hierarchical links are considered transitive [3].

The lack of a way to express less formal semantics hindered many early projects that tried to apply Semantic Web technology in the cultural sector by massaging existing knowledge organization systems into formal ontologies. Given the scope of the artifacts considered, this effort required considerable ontological debugging that was ultimately of dubious value. Indeed, most information retrieval scenarios using KOS for searching or browsing collections do not require more than the information that one concept is broader than another.

RDFS and OWL ontologies are used to structure data by providing classes to which the resources described in data can be stated to belong along with their properties, while value vocabularies merely provide resources to be used as values in statements about those resources. A concept from a thesaurus, such as *architecture*, will, for example, be used as the object of a dc:subject statement describing a book, while the property dc:subject itself is defined in an ontology of properties and classes for describing books. In a recent report, the W3C Library Linked Data group has called the latter *value vocabularies* [4], acknowledging the difference with RDFS and OWL ontologies from a knowledge engineering perspective.

To port the large body of existing value vocabularies to the linked data cloud, simpler data models have been developed, notably the W3C Simple Knowledge Organization System (www.w3.org/TR/skos-reference/). SKOS is the result of a careful weighting of requirements of data publishers and consumers and minimizing the ontological commitment all parties should make [5]. It focuses on concepts (skos:Concept), a small number of semantic relations between them (skos:broader, skos:narrower, skos:related), labels and a few documentation properties. The formal axioms of SKOS remain very simple. For example, SKOS features a property skos:broaderTransitive that offers a transitive interpretation of the skos:broader links between concepts that are in a direct parent/child relationship. However, this property is intended as a mere generalization over the semantics of the original skos:broader property; it means nothing more than the fact that it links a concept to an ancestor concept in the hierarchy defined in the KOS.

To solve some shortcomings of SKOS when capturing data in more complex KOS, proposals have been made to refine and extend the core model, for example, by adding specializations that reflect certain types of parent/child relationships (for instance iso:broaderPartitive in the ISO25964 data model [6]) or introducing formal axioms that tackle the composition of different types of links in thesauri [3].

Still, even more complex KOS remain at the level defined by the notion of value vocabularies. Contrary to the RDFS and OWL ontology languages, KOS do not provide the ability to create the full apparatus of classes and properties required to govern the creation and use of RDF statements.

Traditional efforts for defining ontologies have often tried to present a continuum between simple KOS and RDFS and OWL ontologies (http://tw.rpi.edu/weblog/tag/ontology/), but the roles they play with regard to the making of RDF data are essentially different. Consider the following statements:

@prefix schema: < http://schema.org/>
@prefix dc: < http://purl.org/dc/terms/>
ex:monalisa rdf:type schema:Painting
ex:monalisa dc:type <http://id.loc.gov/authorities/subjects/sh85105182>

Schema.org defines Painting as an RDFS class, fit for use with the rdf:type predicate and additionally specifies ontological information, such as the set of properties with which one expects the instances of the class to be described. The Library of Congress Subject Headings, in contrast, provide the notion of Portraits (http://id.loc.gov/authorities/subjects/sh85105182) as a SKOS concept, adding only informal information such as labels and links to other concepts. Note that the Dublin Core property dc:type, whose informal semantics cover the notion of genre, does not require an RDFS or OWL class as its object, contrary to rdf:type, which strictly indicates an instance-class link in the sense of RDFS and OWL.

The systematic use of SKOS for representing legacy KOS data can be seen as conflicting with the Semantic Web ambition of describing entities in terms of appropriate classes. Lots of KOS consist of subjects to which the SKOS notion of concept fits well. But many KOS describe persons, places and other types of entities for which there also exist dedicated ontologies, such as the Friend-of-a-Friend (FOAF) (http://foaf.org) ontology for persons or the generic Schema.org ontology.

The linked data vision implies a paradigm shift compared to the approach that relied on specific publication places and formats to define reference (authority) data. As highlighted in the Library Linked Data group's report on available vocabularies and datasets [7], data not expressed in SKOS or OWL can still play a reference role for creating other data, just as KOS and authority files did in traditional library, archive or museum environments. The most typical example in linked data is the DBpedia dataset (http://dbpedia.org), where resources extracted from Wikipedia are described using a formalized ontology that captures more than just labels or conceptual abstractions.

It is often not trivial to select a specific real-world entity as the target for expressing a KOS as linked data. Converting to ontologies of real-world entities may require more work than is practical. In practice, many thesauri and other existing KOS provide a mixture of entities (persons, places, concepts) with no easy way to distinguish these various components. Disentangling such entities may require far more effort than KOS publishers can afford. Moreover, much KOS data has historically been devoted to the management of information at semantic or lexical levels, such as synonyms or vague associative links, which may not fit perfectly into domain-specific ontologies.

As a result, models for cultural data tend to remain liberal with respect to the form of data they can accept. The EDM, for example, allows person-related data expressed using either SKOS concepts or real-world entities such as FOAF persons. The Virtual International Authority File (http://viaf.org) publishes data following the two approaches in parallel. The central entities of VIAF are of type schema:Person. But next to this real person one can find a SKOS representation for every corresponding name authority from the libraries contributing to the VIAF dataset, as in the following extract:

<http://viaf.org/viaf/24604287> rdf:type schema:Person
<http://viaf.org/viaf/sourceID/DNB%7C118640445#skos:Concept>
  rdf:type skos:Concept
<http://viaf.org/viaf/sourceID/DNB%7C118640445#skos:Concept>
  skos:prefLabel "Leonardo, da Vinci, 1452-1519"
<http://viaf.org/viaf/sourceID/DNB%7C118640445#skos:Concept>
  foaf:focus
  <http://viaf.org/viaf/24604287>
<http://viaf.org/viaf/sourceID/BNF%7C11912491#skos:Concept>
  rdf:type skos:Concept
<http://viaf.org/viaf/sourceID/BNF%7C11912491#skos:Concept>
  skos:prefLabel "Léonard, de Vinci, 1452-1519"

<http://viaf.org/viaf/sourceID/BNF%7C11912491#skos:Concept>
foaf:focus
<http://viaf.org/viaf/24604287>

where the first and second concepts correspond to the name authorities from the national libraries of Germany and France, respectively.

VIAF uses the foaf:focus property to relate a concept to "the underlying or 'focal' entity" with which it is associated. This enables data consumers to traverse the published RDF graphs to find the data that best suits their application needs.

The foaf:focus pattern reflects in a faithful way the provenance of the data, leading to better trust. In fact, keeping concepts from different origins separate and distinct avoids the risk of inconsistencies that can occur if data is merged too quickly around a single entity. Across different data sources, different names could be used for a same person, or even different dates of birth and deaths.

The danger of such discrepancies is a key aspect of another problem: that of representing the result of entity reconciliation processes. Many datasets refer to the same persons, places or concepts. This situation hinders consumption of data on the web, as relevant data about the same entities can be distributed over different sets. Various entity-linking processes (manual or automatic) have thus been devised to palliate the issue – an area that is out of scope for this paper. To represent the fact that two resources correspond to one same thing, the OWL language provides a property called owl:sameAs. But this property comes with strong semantics; all the statements subjected to two resources in an owl:sameAs relationship could be swapped from one resource to the other, leading to a complete merge. As this can raise some of the inconsistencies mentioned above, other vocabularies have offered alternative links with weaker semantics, such as skos:exactMatch and skos:closeMatch in SKOS. The former reflects a stronger semantic similarity than the latter, but it does not suppose a complete merge of the two concepts it relates. Note that at present, however, many ontologies have tried to solve the problem at their own level; there is not yet consensus on how to tackle the representation of weaker similarity links in a more general way, leaving data publishers a vast array of options [8].

Choosing the Right Level of Semantics

RDF and linked data allow a certain latitude in choosing the type of resource described by some data. In many cases, a FOAF Person can be used instead of a SKOS Concept and the other way around, without breaking the data. Even more interesting, it is possible to provide an OWL class that is defined with some SKOS properties (www.w3.org/TR/skos-primer/) or to use an existing SKOS concept in the position of an OWL class or even to give it a dual type of SKOS Concept and OWL Property, as in the Library of Congress MARC Relators (http://id.loc.gov/vocabulary/relators). In the same vein, one may use the word vocabulary to refer alternatively to RDFS and OWL ontologies or SKOS concept schemes.

Reference datasets based on a specific ontology, KOS resources expressed as SKOS concept schemes and formalized ontologies expressed in RDFS or OWL are most often designed to meet specific goals. While the recommendation may seem obvious, data modelers and publishers should be careful about selecting the level of semantics that best fits their particular domains and tasks [9]. Some application scenarios will require highly formalized ontologies to establish a sufficiently precise meaning for their vocabulary and to perform complex tasks such as reasoning. But over-formalizing can be as dangerous as the lack of semantic axioms, as it can hinder the understanding and re-use of an ontology or dataset while making it more difficult to produce in the first place. ■

Resources on next page

Resources Mentioned in the Article

[1] Godby, C. J., & Denenberg, R. (2015). *Common ground: Exploring compatibilities between the linked data models of the Library of Congress and OCLC*. Dublin, Ohio: Library of Congress and OCLC Research. Retrieved from www.oclc.org/content/dam/research/publications/2015/oclcresearch-loc-linked-data-2015.pdf

[2] ISO TC46/SC9/WG8 working group for the ISO 25964 standard about thesauri. (2011). *ISO 25964-1:2011 Information and documentation – Thesauri and interoperability with other vocabularies – Part 1: Thesauri for information retrieval* [Technical report]. International Organization for Standardization.

[3] Alexiev, V., Cobb, J., Garcia, G., & Harpring, P. (2014). Getty vocabularies: Linked open data: Semantic representation. Retrieved from http://vocab.getty.edu/doc/

[4] Baker, T., Bermès, E., Coyle, K., Dunsire, G., Isaac, A., Murray, P., … Zeng, M. (October 25, 2011). *Library Linked Data Incubator Group final report*. [W3C Incubator Group Report October 25, 2011]. Retrieved from www.w3.org/2005/Incubator/lld/XGR-lld/

[5] Baker, T., Bechhofer, S., Isaac, A., Miles, A., Schreiber, G., & Summers, E. (2013). Key choices in the design of Simple Knowledge Organization System (SKOS). *Journal of Web Semantics, 20*. 35-49.

[6] ISO TC46/SC9/WG8 Working Group for the ISO 25964 standard about thesauri. & Isaac, A. (2012). *Correspondence between ISO 25964 and SKOS/SKOS-XL Models* [Technical report]. National Information Standards Organization, Retrieved from www.niso.org/schemas/iso25964/correspondencesSKOS/

[7] Isaac, A., Waites, W., Young, J., & Zeng, M. (Eds.) (October 25, 2011). *Library Linked Data Incubator Group: Datasets, value vocabularies, and metadata element sets* [W3C Incubator Group Report, October 25, 2011]. Retrieved from www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset/

[8] Bergman, M. (2010). Bridging the gaps: Adaptive approaches to data interoperability [slides]. Keynote talk, DCMI International Conference on Dublin Core and Metadata Applications, Pittsburgh, October 20-22, 2010. Retrieved from www.slideshare.net/mkbergman/dcmi-20101022

[9] Farias Lóscio, B., Burle C., & Calegari, N. (Eds.) (February 24, 2015). Data on the web best practices: W3C First Public Working Draft 24 February 2015. Retrieved from www.w3.org/TR/2015/WD-dwbp-20150224/