On the use of existing upper ontologies as a metadata integration mechanism

Miguel-Angel Sicilia
University of Alcalá, Ctra. Barcelona, km. 33.6 – 28771 Alcalá de Henares (Madrid), SPAIN
Mail:msicilia@uah.es

Abstract:

Metadata scheme harmonization and the development of 'generic' metadata standards have been proposed and studied as techniques to alleviate the problem of metadata integration. However, the use of existing upper onotologies as base models for metadata schemas provides an alternative that allows for a significant reuse of knowledge representations that have been evaluated and refined in the last years. In that direction, this paper explores the linking of *Dublin Core* terms to the *OpenCyc* knowledge base. The results of the integration point out that these upper ontologies are actually prepared for the representation of at least the base elements of current metadata schemas, and they could consequently be used as a core model. This approach would enable higher levels of metadata coherence and richer semantics, as long as the upper ontologies used are reasonably stable.

Keywords:

upper ontology, metadata, metadata interoperability, Dublin Core.

1 Introduction

The growing interest in metadata-related research has resulted in the proliferation of metadata schemas, both general-purpose and specific to some domain, sector or community. According to Greenberg (2003) metadata is "structured data about an object that supports functions associated with the designated object. [...]. The last component of this metadata definition refers to the functions associated with the designated object. The emphasis here is on the ability of metadata to support the activities and behaviors of an object". This orientation to enabling concrete functions naturally results in metadata schemas that use different terms for the definition of similar concepts or relations, coming from different biases oriented to support specific functions for concrete applications. This disparity in the representation of the same elements is at the hearth of the problem of "metadata interoperability", which is in essence a problem of aligning semantics coming from disparate intellectual processes. Metadata schema "harmonization" has been proposed elsewhere (Doerr, Hunter and Lagoze, 2003) as a solution consisting on the "process of modifying two ontologies, preserving their intended functionality, but integrating them into a coherent wider model". Nonetheless, harmonization is considered to "require considerable intellectual effort", since it deals with integrating often conflicting conceptualizations. Even though harmonization processes have an intrinsic value in that they help in clarifying the connections and knowledge assumptions of disparate schemas, it is essentially a pairwise study of schemas that is potentially affected by a form of "combinatorial explosion". Even in the case that the integrated schema resulting from a first harmonization process is used for the alignment of others, this still requires a continuous revision of the "core model" (Doerr, Hunter and Lagoze, 2003) for each schema considered. In consequence, the core model is always tentative in that additional viewpoints added by newly considered schemas may potentially affect the core structure of the ontology. This has also the drawback that the partial nature of community-specific schemas does not guarantee that the model increases in generality, since it may result in a collection of disparate views that do not uncover the essential aspects of the entities being modeled.

As a different approach to address semantic alignment problems, general-purpose metadata models as ABC (Lagoze and Hunter, 2001) have been crafted with the specific objective of facilitating interoperability between metadata ontologies for different domains. But it is interesting to note that if the target use of the generic model becomes broader, at the end it will be similar to a process of "upper ontology" engineering (Uschold and Gruninger, 1996), which is known to be a laborious, time-consuming process. From a pragmatic viewpoint, it may be desirable to have a common core or "upper" metadata schema to map existing and future metadata specifications into a coherent whole. In fact, the ABC model is an important result in that direction. Nonetheless, these schemas are often built "from scratch" in the sense that they do not reuse previous ontological structures. Even though they obviously build on existing ontology research and studies, and they often borrow concepts form previous schemas, they propose a new stand-alone ontology, instead of providing an extension of existing ontologies. An alternative view on schema integration may be that of agreeing on a common existing *base ontology* and then developing the semantics of each community-specific metadata schema as an extension of it. This will provide several benefits, including the following: (a) the reuse of existing work on ontology engineering is guaranteed, (b) the relationships of each new schema engineered on top of the base ontology with existing ones is "as-is" clearly defined by virtue of their common underlying model. This is well-known in ontology integration (Noy, 2004). (c) applications built on the knowledge provided by the base ontology are at least partially equipped to deal with the information of future schemas, reducing software engineering maintenance costs, and (d) the inference and consistency checking mechanisms that are built-in ontology description languages can be exploited. Having considered all these benefits, the main question that remains is if 'appropriate' base ontologies for metadata are currently available. Several efforts on "upper ontology" are currently being integrated as part of the work of the IEEE P1600.1 *Standard Upper Ontology* Working Group (SUO WG)1. This would eventually provide a foundation for metadata interoperability, combined with a provision of broad and rich commonsense semantics. But even if it fails to do so, the use of any of the upper ontologies considered as input by SUO-WG as base ontology for metadata schemas is still rewarding from a methodological and practical perspective. This is because it enables a significant amount of reuse of ontological engineering work distilled in the course of the years. Surprisingly, the adequacy of existing upper ontologies to frame existing metadata schemas is still largely unexplored. Some work exists (Sicilia, et al., 2004) but much more thorough studies are required. In any case, the exercise of linking existing schemas with large upper ontologies will provide the necessary insights to judge the appropriateness of using a base ontology for metadata description.

In this paper, the use of an existing upper ontology as a framework for metadata is explored. Concretely, the linking of *Dublin Core* elements to the *OpenCyc2* knowledge base is described. *OpenCyc* is the open source version of *Cyc* (Lenat, 1995), which contains over one hundred thousands atomic terms, and is provided with an associated efficient inference engine. It attempts to provide a comprehensive ontology of "commonsense" knowledge, including what are usually considered "upper definitions". It is important to highlight that the use of base ontologies as mentioned before is not restrictive of the creativity or particular views on metadata required by different domains or communities, since the only requirement for new schemas is that they are *explicitly framed* in the existing base ontology. A useful departure assumption for this approach is that the resources to be described and the values or other resources used to describe them are all **reified** in the formal ontology, i.e. the ontology provides a representation of the describing and the described elements, even though the actual contents or information of that described may be external to it – see (Sicilia et al. 2003). This provides a coherent representation and enables the use of inference and other facilities that are standard in modern formal ontology frameworks. This paper describes a concrete case of such kind of integration.

2 Linking *Dublin Core* Elements to *OpenCyc* definitions

The Dublin Core metadata standard is defined as "a simple yet effective element set for describing a wide range of networked resources". DCMI metadata terms3 are the elements and element refinements (i.e. elements that narrow the semantic of others) that can be associated to resources. In what follows, a tentative mapping of some of the main definitions in the DCMI model to *OpenCyc* elements is provided.

2.1. Representing resources

The *DCMI Abstract Model*4 mentions *resources* and *descriptions* as the two main elements of the standard. Resources are described as "anything that has identity" and represent also aggregations of resources. If we consider metadata inside a knowledge base as *OpenCyc*, the broader interpretation of what resources are could be that of considering that the term oc_Thing5, which "contains everything there is. Every thing in the *Cyc* ontology - every Individual (of any kind), every Set-Mathematical, and every Collection -

[1] http://suo.ieee.org/

[2] http://www.opencyc.org/

[3] http://dublincore.org/documents/dcmi-terms/

[4] dublincore.org/documents/abstract-model/

is an instance of Thing". Nonetheless, this flexibility does not appear to match the current scope of metadata practice, which is focused on describing Web resources (including both public and private ones). An alternative may be that of defining resources as a subset of the digital entities subsumed by oc_ComputerDataArtifact, defined as "The collection of all pieces of computer data stored in hardcopy form or in computer memory. This collection includes all such information, such as data streams [......] and more complex objects such as files in a filesystem". This could be complemented by considering also resources to running processes, which would allow the description of agents and other running software entities like services. Another alternative would be that of assimilating resources to a subsuming entity like oc_InformationBearingThing that refer to elements with "interpretable contents", for which metadata elements as "keywords" seem more applicable.

Nonetheless, if we strictly adhere to DCMI definition "a resource is anything that has *identity*", then the previously described definitions are not completely satisfactory. The concept of identity on the Web is usually assimilated by "anything addressable through a URI", so that it could be considered that every URI-identified element is a resource. If we adhere to an ontological notion of *identity* (Welty and Guarino, 2001), then URIs may serve as extrinsic *identity conditions*. These could be different from other possible ontological identity conditions, specific to each kind concept. Either way, the identification through URIs inside ontologies can easily assimilated to the subset of oc_Things identified by URIs for maximum flexibility, and narrower definitions can be used whenever required for concrete metadata elements. In addition, *OpenCyc* provides the oc_UniformResourceLocator term modeling these identifiers, and a number of predicates for specific URL-mapping like oc_urlOfCW that link URLs to digital instantiations of oc_ConceptualWorks.

Resources in the DCMI abstract models may belong to one or several *classes*, which are used for the refinement of "semantics". The declaration of classes in the case of *OpenCyc* is simply the definition of new terms, and the common subsumption semantics is a flexible capability of "refined semantics". Moreover, the concept of "property/value pair" in terms of ontology languages is simply the instantiation of a predicate for a concrete value, and describing a concrete instance of a term or class. Even though *OpenCyc* allows predicates of arbitrary arity, they can always be expressed in terms of binary ones by the introduction of defined terms, e.g. the oc_SkillRequired predicate relates one "activity type" to other required one and associated with a level. This can be expressed in two steps by relating the first "activity type" with an instance of a new term, let's say "activity type level", being the latter in turn related to both an activity type and a level.

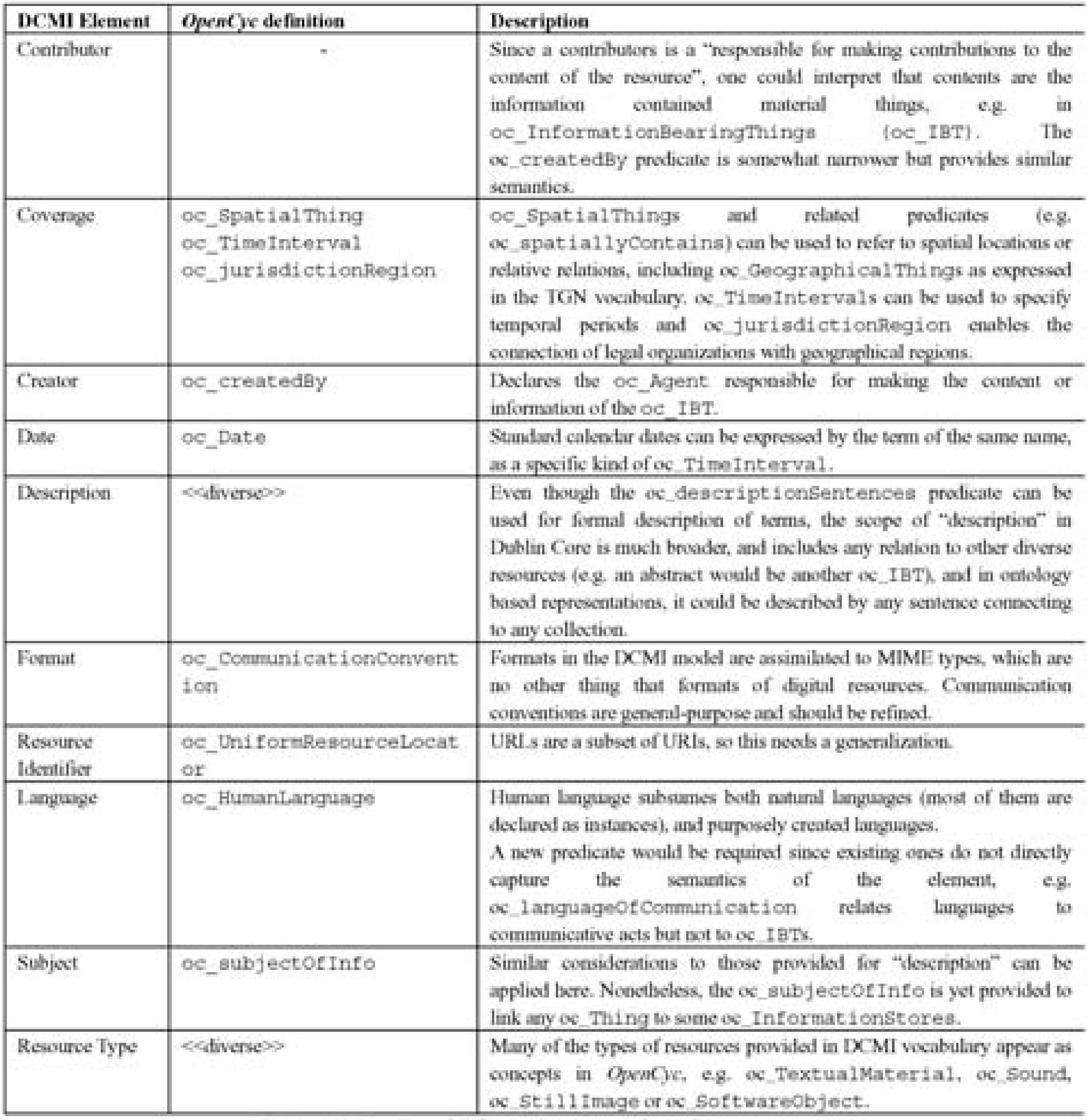
A *property* in the DCMI abstract model is described as "specific aspect, characteristic, attribute, or relation used to describe resources". This in *OpenCyc* can be assimilated to the notion of predicate, and the definition of sub-predicates that are common to many ontology description languages can be assimilated to "sub-properties" in the DCMI. Once again, the semantics of these predicates are no other thing that the surrounding concepts and definitions that are related to the given predicate inside the ontology. For example, the predicate oc_ authorOfLiteraryWork-CW can be considered a concrete property with semantics associated to the concept of literature, which is in turn connected to genres and to editorial constraints as part of its semantics. In consequence, a flexible interpretation of the concept of DCMI resource leads to consider URI-identified ontology elements as resources, which in turn are related to others by instantiations of predicates. The elements in Table 1 are some of the predicates that are directly linked to the current DC semantics.

Table 1 provides possible mappings for Dublin Core terms of type "element", i.e. for properties maintained currently by the DCMI. While some elements have not a direct correlate, all the facets described by them are actually contained in the knowledge base. It should be noted also that an additional benefit of framing elements into *OpenCyc* is that the domain of the properties is made explicit and narrower inside a logical framework.

2.2. Representing descriptions

Descriptions in the DCMI abstract model are "one or more statements about one, and only one, resource", with *statements* being "a **property URI** (a URI reference that identifies a property), zero or one **value URI** (a URI reference that identifies a value of the property), zero or one **vocabulary encoding scheme URI** (a URI reference that identifies the class of the value) and **zero or more value representations** of the value". From the viewpoint of ontology integration this would entail that: (a) properties (i.e. predicates or slots) must also be URI-identified, (b) the value associated with the property may have alternate representations. Alternate representations are a matter of lower-level interoperability and could be handled in the ontology through *ad hoc* techniques. For example, various language strings can be incorporated as extra-logical predicates in a similar way to the approach

[5], OpenCyc elements are prefixed with "oc".



DCMI Element	*OpenCyc* definition	Description
Contributor	-	Since a contributors is a "responsible for making contributions to the content of the resource", one could interpret that contents are the information contained material things, e.g. in oc_InformationBearingThings (oc_IBT). The oc_createdBy predicate is somewhat narrower but provides similar semantics.
Coverage	oc_SpatialThing oc_TimeInterval oc_jurisdictionRegion	oc_SpatialThings and related predicates (e.g. oc_spatiallyContains) can be used to refer to spatial locations or relative relations, including oc_GeographicalThings as expressed in the TGN vocabulary. oc_TimeIntervals can be used to specify temporal periods and oc_jurisdictionRegion enables the connection of legal organizations with geographical regions.
Creator	oc_createdBy	Declares the oc_Agent responsible for making the content or information of the oc_IBT.
Date	oc_Date	Standard calendar dates can be expressed by the term of the same name, as a specific kind of oc_TimeInterval.
Description	<<diverse>>	Even though the oc_descriptionSentences predicate can be used for formal description of terms, the scope of "description" in Dublin Core is much broader, and includes any relation to other diverse resources (e.g. an abstract would be another oc_IBT), and in ontology based representations, it could be described by any sentence connecting to any collection.
Format	oc_CommunicationConvention	Formats in the DCMI model are assimilated to MIME types, which are no other thing that formats of digital resources. Communication conventions are general-purpose and should be refined.
Resource Identifier	oc_UniformResourceLocator	URLs are a subset of URIs, so this needs a generalization.
Language	oc_HumanLanguage	Human language subsumes both natural languages (most of them are declared as instances), and purposely created languages. A new predicate would be required since existing ones do not directly capture the semantics of the element, e.g. oc_languageOfCommunication relates languages to communicative acts but not to oc_IBTs.
Subject	oc_subjectOfInfo	Similar considerations to those provided for "description" can be applied here. Nonetheless, the oc_subjectOfInfo is yet provided to link any oc_Thing to some oc_InformationStores.
Resource Type	<<diverse>>	Many of the types of resources provided in DCMI vocabulary appear as concepts in *OpenCyc*, e.g. oc_TextualMaterial, oc_Sound, oc_StillImage or oc_SoftwareObject.

Table 1. Mapping of DCMI elements to OpenCyc terms.

taken to link *WordNet* terms to *OpenCyc* elements. Since value URIs and vocabulary encoding schemes URIs are ways to refer to resources and their classes, they can simply be assimilated to that related notions. Nonetheless, from that viewpoint every statement will include (implicitly) possibly several vocabulary encoding schemes, which are simply the classes to which the resource belong. This seems not to represent a flaw in the mapping, since the representation of resources through ontology instances removes possible "representation" problems.

The URI-centric definitions of the DCMI abstract model entails a bias to markup languages that is not required in a ontology grounding approach. Since modern ontology description languages as OWL use URIs to describe every ontology element, we could remove the restriction that properties and resources are URI-identified, since they will be so described later. In the case of resources that have "external" URIs (e.g. instantiations of Web resources), the ontology instances reifying them will simply provide the external URI to associate all the knowledge (i.e. descriptions) connected to them from any ontological representation. The discussion above evidences that DC semantics are currently covered by *OpenCyc* in a general sense and their associated representational issues could be handled through a number of simple extensions.

3 Conclusions and Outlook

This paper has reported an example study on the alignment of Dublin Core and related metadata aspects with a mature large knowledge base. The result provides evidence on the suitability of the approach, since the base elements can be directly linked to *OpenCyc* definitions with little semantic alignment effort, and in many cases providing with a rich surrounding conceptual structure that would help in improving the descriptive power of metadata records. Even though the mapping described here does not attempt to be definitive, it may serve as a foundation for further studies in the area. The most straightforward direction for future work is that of evaluating the suitability of different upper ontologies to serve as the base for metadata integration. The supporting "library system" should also be subject to evaluation for practical purposes (Ding and Fensel, 2001). After this, it would be required a process of alignment of current schemas, and future ones could benefit of common grounding with previous work.

References

1. Y. Ding and D. Fensel (2001). Ontology Library Systems: The key for successful Ontology Reuse. *The first Semantic web working symposium* (SWWS1), USA.
2. Doerr, M., Hunter, J., Lagoze, C.: Towards a Core Ontology for Information Integration. *Journal of Digital Information*, 4(1), 2003.
3. Greenberg, J. (2003). Metadata and the World-Wide-Web. *Encyclopedia of Library and Information Science*, 1876-1888.
4. C. Lagoze and J. Hunter, "The ABC Ontology and Model," *Journal of Digital Information*, 2 (2), 2001.
5. Lenat, D. B. Cyc: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM* 38(11): 33—38 (1995).
6. N. Noy (2004). Semantic integration: a survey of ontology-based approaches *SIGMOD Record*, Vol. 33, No. 4.
7. Sicilia, M.A., García, E., Aedo, I. and Díaz, P. (2003). A literature-based approach to annotation and browsing of Web resources. *Information Research*, 8(2), paper no. 149
8. Sicilia, M. A., García, E., Sánchez-Alonso, S. and Rodriguez, E. 2004. On integrating learning object metadata inside the OpenCyc knowledge base. In *Proceedings of the 4th IEEE International Conference on Advanced Learning Technologies*, pp. 900-901.
9. Uschold, M. and Gruninger, M. Ontologies: principles, methods and applications. *Knowledge Engineering Review*, Vol. 11:2, 1996, pp. 93–136.
10. Welty, C. and Guarino, N. Supporting ontological analysis of taxonomic relationships. *Data and Knowledge Engineering* 39(1), 2001, pp. 51-74.