An Ontology Design Pattern for Data Integration in the Library Domain

Patrick OBrien,[1] David Carral,[2] Jeff Mixter,[3] Pascal Hitzler[2]

Montana State University,[1] Wright State University,[2] OCLC[3]

Abstract. A university's institutional repository (IR) contains the intellectual output of its faculty, staff and students. Its content is extensive and heterogenous, which complicates data aggregation and discovery tasks. To address these challenges, we propose the use of a conceptual ontology design pattern to model information for the IR domain which is general enough to be reused across different IR datasets.

1 Introduction

A university's institutional repository (IR) contains the intellectual output of its faculty, staff and students. Content can be diverse and may include theses and dissertations, proceedings, books, preprints and post-print journal articles, as well as grey literature and datasets that support research conclusions. While there are a number of Linked Open Datasets (LOD) with structured bibliographic records on the web (i.e., DBLP, CiteSeer, Semantic Web Dog Food, etc.), none have open access to a full text version of the scholarly article or a robust view of the academic output for an entire University.

Currently there are more than 2,400 IR affiliated with universities or disciplinary societies that are built on the principle of open access [**?**]. Most IR include full text versions of the scholarly work encoded as media objects (PDF, CSV, etc.). IRs contain a vast amount of data encapsulating information that can provide unique perspectives on institutional research activities, such as the interdisciplinary collaboration among researchers, departments and colleges.

However, this valuable information is typically locked in bibliographic records as simple text strings, or blobs, that are difficult for machines to isolate, ingest and interpret. Unstructured IR data also hinder discovery by making indexing by scholarly search engines difficult [**?**].

To unlock the full potential of open access IR, it is necessary to dissect each bibliographic record to identify, and link together, the entities contained within. The research question, then, is whether a repeatable structured data model can improve access and discovery of IR content by improving the quality of IR data.

This paper describes a generic Ontology Design Pattern (ODP) based on a project to convert bibliographic records from Montana State University's Open Access Institutional Repository (IR) into linked data and still improve access and discovery by services such as Google and Google Scholar. Like most libraries, Montana State University's IR metadata was maintained in multiple production systems using various formats to describe and access the same scholarly papers encoded as full text PDF files. Specifically, MAchine Readable Cataloging (MARC) and Metadata Object Description Schema (MODS).

The challenge was producing a single accurate, and robust, description of the materials contained within the IR. This required staff to extract, consolidate, and parse records into individual text strings and transform them into RDF. This was done using a model based upon Schema.org, Dublin Core and extended using the Citation Style Language for granular details. Once converted into RDF, the data were reconciled against the university's internal Faculty Activity Database to establish instance data of people with their Colleges and Departments. The RDF data were then linked to the external sources of DBpedia and the Library of Congress Subject Headings (LCSH). While the process was successful in publishing Montana State University's IR as LOD[**?**], this process required significant ad hoc and manual processes to identify and address data quality issues.

We propose a generic Ontology Design Pattern (ODP) developed with the three characteristics below would help IR managers improve the speed and efficiency for publishing IR content as quality LOD:

1. Directly applicable to a variety of IR datasets and, thus, reduce the initial hurdle for IRs to publish Linked Data [**?**].
2. Easily extensible, e.g., by aligning with existing library ontologies, foundational ontologies, and other domain specific vocabularies.
3. Help IR data managers improve the quality of IR metadata by reducing the practice of manually reviewing bibliographic records for accuracy.

Deriving such an ODP requires a generic use case which captures recurring problems in different application domains. *Competency questions* are queries that a domain expert would be expected to run against a knowledge base and are recognized as a good approach for modeling requirements from multiple domains. For the proposed ODP, such competency questions include:

1. *Which records violate existing conditions required for scholarly citation?*
2. *What is the topic diversity of an organization intellectual output?*
3. *What is the depth of an organization's intellectual output?*
4. *Are their authors with "weak ties" to my domain of expertise I can explore for "novel ideas" or collaboration in my research?*

2 Formalization

This section discusses the more interesting classes, properties, and axioms of the library pattern. Description Logics (DL) notation has been used to present the axioms. To encode the pattern, we make use of the logic fragment $\mathcal{SROIQ}$ as defined in [**?**], which is the basis for the OWL 2 DL standard [**?**]. The proposed ODP has been formally encoded using the Web Ontology Language (OWL).[1] A schematic view of the pattern is shown in Figure **??**.

[1] The pattern can be downloaded from `www.dropbox.com/sh/88jh5qwdgpxueqz/AAAj_kgmL5ErPL2JaPWtCvEsa?dl=0`.

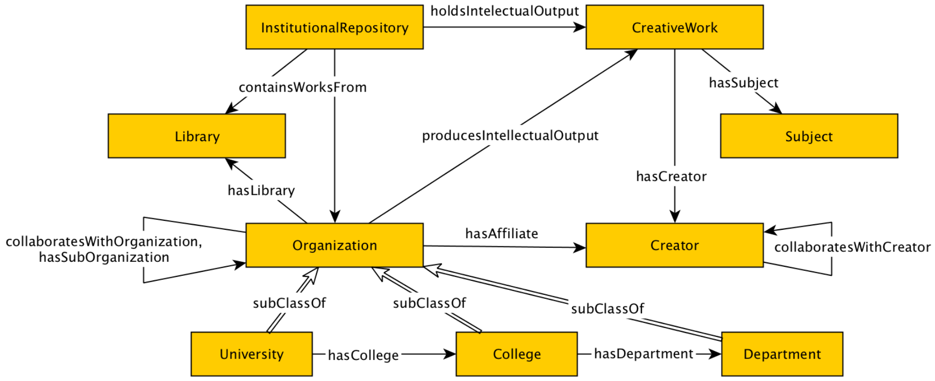

Fig. 1. A schematic view of the Library ODP

CreativeWork: a generic class of creative work that includes things like books, movies or software programs. A subclass of CreativeWork, ScholarlyWork, contains all creative works related to scholarly research. The CreativeWork and ScholarlyWork class relationship is enforced by axiom (1). Axiom (2) indicates that every scholarly work must have some author and exactly one publication date.

$$\mathsf{ScholarlyWork} \sqsubseteq \mathsf{CreativeWork} \tag{1}$$

$$\mathsf{ScholarlyWork} \sqsubseteq \exists \mathsf{hasCreator.Creator} \sqcap = 1 \mathsf{hasPublicationDate.Date} \tag{2}$$

Creator: some person or organization responsible for generating some creative work. All creators must have created at least some CreativeWork (3).

$$\mathsf{Creator} \sqsubseteq \exists \mathsf{isCreatorOf.CreativeWork} \tag{3}$$

InstitutionalRepository: a repository which contains a set of creative works. It is related to some organization. An institutional repository must contain some type of scholarly work from some creator.

$$\begin{aligned} \mathsf{InstitutionalRepository} \sqsubseteq\ & \exists \mathsf{containsWorksFrom.Organization} \sqcap \\ & \exists \mathsf{holdsIntelectualOutput.CreativeWork} \end{aligned} \tag{4}$$

Organization: An entity that formally links a group of people to a common goal. A relevant class of Organization for our context is ScholarlyOrganization (5). Universities, colleges, academic departments, and libraries are scholarly organizations (6-9).

$$\mathsf{ScholarlyOrganization} \sqsubseteq \mathsf{Organization} \tag{5}$$
$$\mathsf{University} \sqsubseteq \mathsf{ScholarlyOrganization} \tag{6}$$
$$\mathsf{College} \sqsubseteq \mathsf{ScholarlyOrganization} \tag{7}$$
$$\mathsf{Department} \sqsubseteq \mathsf{ScholarlyOrganization} \tag{8}$$
$$\mathsf{Library} \sqsubseteq \mathsf{ScholarlyOrganization} \tag{9}$$

Universities have at least one college and one academic department (10). Colleges are part of at most one university (11). Academic departments are part of at least one and only one university (12).

$$\mathsf{University} \sqsubseteq \exists \mathsf{hasCollege.College} \sqcap \exists \mathsf{hasDepartment.AcademicDepartment} \tag{10}$$
$$\mathsf{College} \sqsubseteq \leq 1 \mathsf{isCollegeOf.University} \tag{11}$$
$$\mathsf{Department} \sqsubseteq = 1 \mathsf{isDepartmentOf.University} \tag{12}$$

We introduce subproperty statements (13-14) and declare the subproperty hasSubOrganization as transitive with the following axioms:[2]

$$\mathsf{hasCollege} \sqsubseteq \mathsf{hasSubOrganization} \tag{13}$$
$$\mathsf{hasDepartment} \sqsubseteq \mathsf{hasSubOrganization} \tag{14}$$
$$\mathsf{hasSubOrganization} \circ \mathsf{hasSubOrganization} \sqsubseteq \mathsf{hasSubOrganization} \tag{15}$$

The following role chain enables automatic determination of some organization's intellectual output:

$$\mathsf{hasSubOrganization} \circ \mathsf{hasAffiliate} \sqsubseteq \mathsf{hasAffiliate} \tag{16}$$
$$\mathsf{hasAffiliate} \circ \mathsf{isCreatorOf} \sqsubseteq \mathsf{producesIntellectualOutput} \tag{17}$$

3 Conclusions and Future Work

Applying an ODP to IR data will improve the efficiency and effectiveness of library metadata management workflows by quickly identify issues with data that are currently done manually. Improving the quality of IR metadata and publishing it for syndication on the Semantic Web will aid machine assisted discovery and help address the limited availability of datasets that contain adequate information linked to full-text scholarly research capable of supporting semantics-driven Literature-Based Discovery [**?**].

We are planing future iterations that extend the axiomatization and populate the pattern using previous domain modeling and a real-world dataset from Montana State University [**?**].

[2] Many axioms which are intuitively derived from labels such as $\mathsf{isCollegeOf}^{-} \equiv \mathsf{hasCollege}$ are omitted. For a comprehensive list see out submission at `www.dropbox.com/sh/88jh5qwdgpxueqz/AAAj_kgmL5ErPL2JaPWtCvEsa?dl=0`.