Developing a Collection Registry for IMLS NLG Digital Collections

Sarah L. Shreeves, Timothy W. Cole
University of Illinois Library at Urbana-Champaign, USA
{sshreeve, t-cole3}@uiuc.edu

Abstract

The University of Illinois at Urbana-Champaign has been funded to create a collection registry and item-level metadata repository for digital collections and content created under the auspices of the Institute of Museum and Library Services (IMLS) National Leadership Grant (NLG) program. The first stages of developing the collection registry are described: defining a collection; surveying current collection description schemas and registries; and determining which schema to use and whether to modify it. Next steps will include populating the registry, testing the efficacy of the schema and designing internal and public interface.

Keywords: *Interoperability, Collection Description, Metadata Schemas, Digital Collections*

1. Background

Through its NLG program, IMLS is enabling the development of hundreds of significant new digital collections. The visibility, adaptability, and interoperability of these collections and their content have become increasingly important as new paradigms of digital libraries have developed. In September 2002, IMLS awarded the University of Illinois at Urbana Champaign a National Leadership Grant for a three-year research and demonstration project. A collaborative effort between the University of Illinois Library, the Graduate School of Library and Information Science, and IMLS, the primary goals of the IMLS Digital Collections and Content (IMLS DCC) project are to:

- Create a registry of digital collections funded by the IMLS NLG program since 1998.
- Implement a searchable item-level metadata repository for these collections using the Open Archives Initiative Protocol for Metadata Harvesting (OAI PMH).
- Assist NLG recipients in setting up OAI-PMH compliant data provider services.
- Research how resource developers can best represent collections and items using metadata.

The creation of a well designed collection registry has been a major focus of the IMLS DCC project thus far. The stages of developing a collection registry are defining what constitutes a collection (for purposes of the registry), creating or adopting a metadata schema to describe digital collections created under the NLG program, designing a registry database architecture, developing internal and public interfaces, populating the registry, and thoroughly testing its effectiveness and usability. The first two stages of work are described below.

2. Defining a Collection

We began with a model that equated NLG funded projects with collections for purposes of the registry. In December 2002 IMLS provided us with grant proposals of 94 NLG projects with digital content funded since 1998. An analysis of these proposals and project websites, a discussion with our project steering committee, and further research into the concept of a collection revealed a far more complicated picture. While there exists a one to one correlation between many NLG funded projects and the digital collections they create, many of the projects created multiple collections, or created a large collection with many distinct sub-collections. Furthermore, for many projects the important outcome was not access to the digital collection, but education, preservation, training, or collaboration. In such cases equating the project with the collection misrepresented the project. Steering committee members also noted that projects are transient; but that collections are meant to persist. However, they stressed the need to maintain links between a project and its collection(s).

Surprising little has been written on defining a collection – whether digital or physical. Johnston and Robinson have defined a collection as 'any aggregation of individual items" and have noted that the aggregation can be based on any number of things, including institution, use, format, and donor [1]. Lee conceptualizes collections in the digital world as information environments or contexts of selected and ordered resources which facilitate information seeking by users [2]. The key elements in this definition are the context, the selection and organization of resources, and a focus on the user. We identified several criteria for designating a collection in the registry:

- cohesive (whether by topic area, holding institution, type of material, etc.);
- searchable as a distinct collection or subcollection;
- a unique point of entry (URL).

This last criterion is largely practical and based on the following user scenario. Imagine that a large collection has multiple sub-collections without distinct URLs. If a search retrieves several of these sub-collections but the entry is

always to the same top-level URL, a user may not understand the distinction between these various sub-collections. Requiring a unique URL would ease users' access to the sub-collection. In the course of our project we will test and refine these criteria for collection definition.

3. Development of the IMLS DCC Collection Description Metadata Schema

To develop a collection metadata schema for this project, we began by surveying what work had already been done on collection description (CD) metadata schemas and registries. We identified three emerging standard schemas of interest. The Research Support Library Programme (RSLP) CD Schema [3] contains descriptive attributes about a collection, its location, agents associated with the collection, and relationships with internal or external collections. The RSLP CD schema is very well documented and has been implemented by RSLP projects throughout the UK. However, it has not been well tested for use in describing digital collections. The Dublin Core (DC) CD schema [4],being developed by the Collection Description working group of the DC Metadata Initiative, is based on the RSLP CD schema, but adapted and simplified for digital collections. It does not include the location and agent information. There are no known implementers. The Encoded Archival Description (EAD) [5] is used primarily to encode archival finding aids, although it has been used to describe museum collections. [6] It can describe both a collection as a whole and the discrete contents of that collection. This schema is used primarily by the archival community. After the analysis of these schemas and discussions with the authors of the RSLP and DC collection description schemas we determined that an adaptation of the RSLP schema would best fit our needs.

As mentioned above the RSLP CD schema describes the collection, its location, and any associated agents (such as owner, collector, or administrator). We examined each of these entities closely and adapted each slightly to fit the needs of our registry. We changed location to institution. We added a project entity to maintain the link between NLG funded projects and collections. In addition we dropped some attributes which were of use only for physical collections (for example, hours of access). **Table 1** lists attributes for our proposed schema and shows correspondence to RSLP elements and DC refinements.

4. Next Steps

A database has been developed to implement our project's collection description metadata schema. A survey instrument has been created to help obtain information needed to populate the registry. Next an initial collection registry entry will be created for each project. Web forms will then allow individual projects to enhance these initial entries and/or add entries for sub or sibling collections associated with a project. (By default certain attributes will

be inherited by all collections associated with a given project.) Concurrently a search and discovery interface will be created to support end-user use of the registry. Feedback from the projects will help us refine our schema and registry design. Focus group testing with selected groups of end-users also will be conducted

Table 1. IMLS_DCC and RSLP CD elements

Proposed Elements in IMLS_DCC Schema	Equivalent Element in RSLP CD Schema	Refinement of DC
title.collection	dc:title	—
has.Service.URL (the URL of the collection)	cld:hasLocation	relation
identifier.IMLSGrant	—	—
format	dc:format	—
language	dc:language	—
type	dc:type	—
audience	—	—
accessControl	cld:accessControl	rights
legalStatus	cld:legalStatus	description
institution.holding	cld:owner	relation
institution.contributing	—	relation
description	dc:description	—
strength	cld:strength	description
subject.topic	dc:subject	—
subject.objectName	cld:objectName	subject
subject.personName	cld.agentName	subject
coverage.spatial	dcq:spatial	coverage
coverage.temporal	dcq:temporal	coverage
relation.hasPart	dcq:hasPart	relation
relation.isPartOf	dcq:isPartOf	relation
relation.hasAssociation	cld:hasAssociation	relation

References

[1] Johnston, P. & Robinson B. (2002) Collections and Collection Description. *Collection Description Focus Briefing Paper*. No. 1. Retrieved May 12 2003. http://www.ukoln.ac.uk/cd focus/briefings/bp1/bp1.pdf

[2]Lee, H. (2000). What is a collection? [Electronic version]. *Journal of the American Society for Information Science*. 51 (12): 1106-1113.

[3]RSLP Collection Description Schema. Retrieved Jan. 28 2003.http://www.ukoln.ac.uk/metadata/rslp/schema/

[4]Dublin Core Collection Description. Retrieved May 12 2003. http://homes.ukoln.ac.uk/~lispj/dc-cd/rslpcd.html

[5]Encoded Archival Description. Retrieved May 14, 2003. http://www.loc.gov/ead/ead.html.

[6]Rinehart, R. (2003). MOAC: A report on integrating museum and archive access in the Online Archive of California. *D-Lib Magazine*. 9(1). Retrieved Jan 17 2003. http://www.dlib.org/dlib/january03/rinehart/01rinehart.html
.