Open Source Software to Manage Digital Collections in a Decentralized Environment

Scott A. Gunn
Education Development Center, Inc., Newton, MA, USA
Tel. +1 617 969 7100
Fax. +1 617 527 8906
E-mail: sgunn@edc.org

Abstract: This paper describes the design, creation, and prototyping of a new open source software application to manage digital collections. The software is standards based, being strongly rooted in Dublin Core. This paper highlights the particular design objective of this software: it allows decentralized maintenance of sub-collections ("virtual collections") within a larger collection.

Keywords: digital library; metadata applications; open source; digital collections.

1   Introduction

Education Development Center, Inc. (EDC) needed a means of developing digital collections that would allow centralized administration while also allowing delegated, decentralized management of many aspects of collections and sub-collections. Our survey of extant solutions found that there were large solutions that were too complicated for our needs, and there were other solutions that did not provide adequate features. We undertook to develop an open source package that would meet our needs and perhaps the needs of other organizations.

We began with several requirements:
- Minimal dependencies for installation
- Provides web-based interface for users and administrators
- Allows one central collection with multiple "virtual collections" or sub-collections that appear to be freestanding, and can even be assigned to different domain names or URLs
- Can catalog internal items (owned by an organization; on their own servers) or external items (web sites, tools, papers on other web sites)
- Allows delegated role-based permissions to allow decentralized maintenance of a large collection
- Allows areas to be marked private such that one must be logged in to view items
- Uses Qualified Dublin Core, including data storage and presentation
- Provides RSS[1] feed of new items or "featured items" within a "virtual collection."
- Allows bulk ingesting of many items at once
- Allows end users to submit items for approval via a simple workflow system

2   Background Information

EDC is an international non-profit organization, building bridges between research, policy, and practice. Today, EDC manages 325 projects in 40 countries. Our work

strengthens nearly every facet of society, including early child development, K-12 education, health promotion, workforce preparation, community development, learning technologies, basic and adult education, institutional reform, and social justice.
EDC is very decentralized, and the organization is project-based in its financial and management structure. Consequently, company-wide practices can be difficult to mandate. It was important, therefore, for us to allow each of EDC's centers or projects to make decisions about their digital collections. For example, keyword controlled vocabularies and taxonomies might vary from project to project. At the same time, we wanted a single system to ease administrative burdens and to make organization-wide searches possible.

Past experience with decentralized creation of digital libraries at EDC had led to a variety of systems and standards, yielding a burden in continued maintenance, as well as difficulty in providing organization-wide searching (metasearch). Custom solutions were built for particular projects, and the systems themselves were difficult to maintain and the data were difficult to manage or to extract. Because Dublin Core was not used in many of these collections, standards across collections varied.

For these reasons, we wanted a system that would allow each project or center to maintain "virtual collections" within a larger system. Our core organizational metadata namespace could be maintained centrally, while allowing managed extensibility for particular needs. Keyword vocabularies and taxonomies needed to be maintained by delegated managers.

We found that a growing number of our funders were requiring that digital collections meet particular standards, especially Dublin Core and related standards such as the Gateway to Educational Materials (GEM)[2].

The initial prototype was created for Project LEAD, which is funded by The Wallace Foundation. This project is housed within EDC's Center for Leadership and Learning Communities[3].

3   Technical and Design Notes

3.1   Platform
It was our intention to release this software as open source software, free of charge, for use by other organizations. Therefore, we wanted a system that would be easy to install and have few if any requirements for commercial software.

We elected to design the system for Linux, although it could probably be ported to Windows. The code is written in object-oriented Perl (with minimal requirements for CPAN or other modules), and it stores its data in MySQL. The web server is Apache.

The database schema is based in Qualified Dublin Core, with particular influence from the DC-Lib application profile[4]. We have an internal application profile that specifies

controlled vocabularies and usage notes, with appropriate references to RFCs and other standards, alongside some internally devised schemes and adaptations.

The application is coded in Perl such that each item in the collection is treated as an object, with instance methods and properties. There are also class methods for items, categories, and users. This careful design allowed for rapid prototyping, and it further allows straightforward additions of new features.

3.2   Virtual Collections

It was important for us to be able to allow each project within our organization to appear to have its own collection. This means that we wanted our projects to be able to define their own look and feel, their own branding, and their own URL. By reading the URL, the system dynamically renders collections such that visitors may not even be aware they are visiting a sub-collection within a larger collection. Each collection may have its own graphics, funder logos, and style sheets.

Here we show examples of two virtual collections, both within the same instance of software and database. The first example (Figure 1) is from the Education Leadership Resource Library (http://cllc.edc.org/rl/). The second example (Figure 2) is from the Suicide Prevention Resource Center Library (http://library.sprc.org/).

Figure 1: Education Leadership Resource Library -- welcome page

EDUCATION LEADERSHIP RESOURCE LIBRARY

Search the Library

Enter search words

[Search] [Clear]

Advanced Search

Home

Leadership Development
Preparation and Training
Recruitment

Leadership Policy Environment
Partnerships and Networks
Policymaking and Reform
Sustainability

Project LEAD
Meetings

Show All Categories

Log in

User registration

About the Library

Frequently Asked Questions (FAQ)

Contact Us

The Wallace Foundation™
Supporting ideas.
Sharing solutions.
Expanding opportunities.™

Welcome to the Education Leadership Resource Library

Important work and research are occurring in the field of education leadership. The Education Leadership Resource Library collects resources to support the development of current and aspiring administrators, coaches, researchers, and others. The Library is a tool that provides an efficient, quick way for leaders to seek useful information. It is our hope that this library will allow the lessons being learned to be shared across the education community.

All resources on this site have been submitted by its users. The library depends on its users for submissions of new resources. New submissions from a wide range of participants guarantee a diversity of current resources to inform and improve leadership in education. To submit, you must first go through our free user registration process. Full details are available on the registration page.

The content of the Library is focused in two broad areas -- Leadership Development and Leadership Policy Environment. The Library is initially designed to support The Wallace Foundation Education Leadership Initiative. Available resources are currently organized to correlate to the goals of the Initiative: leadership skill development, working conditions for leaders, alignment of leadership policies and programs, and partnerships and networks for sustainability. You may explore our site by clicking these links, by using the expanding category list in the left-hand navigation bar, or by using our search feature.

This Resource Library was developed at the Education Development Center (EDC). The development and launch of the Education Leadership Resource Library is made possible with the funding and support of The Wallace Foundation. This library is hosted by the Center for Leadership and Learning Communities at EDC. For further information, visit About the Library. We welcome your feedback and questions; for this you may use our Contact Us page.

Featured Items

Richard Laine Opening Session: Part I Video

Richard Laine Opening Session: Part II Video

School-Based Staff Development

Transforming High Schools

The Careers of Public School Administrators

What's New?
Most recent additions to the Library

The Careers of Public School Administrators

Transforming High Schools

Tutors For Kids

Six Principles of Effective Accountability

A 'Perception Gap' on Working Conditions

Partnering with Families and Communities

Edutopia Professional Development Module: View of the Principal and the Job

The Seven Principles of Sustainable Leadership

Figure 2: Suicide Prevention Resource Center Library -- welcome page

3.3   Taxonomy

One distinguishing feature of our system is that it is based on non-rigid taxonomies. For example, suppose that we have a taxonomy in which we have placed resources about food. Under "fruit" we might have "apples" and "oranges" and then further levels of specificity. In a strict taxonomy, an item would appear in only one node. In our system, we allow items to appear in multiple locations. This means, in our example, that we could also have a taxonomy in which we list "breakfast ingredients" and "pie ingredients"; oranges and apples could appear in each one respectively, as well as their place under the fruit section. In a decentralized organization, several departments might choose to list an item in their own choice of frameworks. For example, if a mistake in a creator's name is corrected, that will be reflected for everyone.

Another implication of this feature is that we will be able to maintain a virtual collection for official EDC publications and products, and each center or project can submit items for approval in this collection, viewable and searchable by the general public via our EDC web site. Absent this integrated approach, we would maintain a wholly separate database, requiring duplicate entry and maintenance for each item. Now those responsible for maintaining the external catalog (i.e. our Communications Office) can

have control over what appears in the catalog, and the creators of each product or item (i.e. the project or center) have control over the metadata.

We allow role-based delegation – with inheritance – within the taxonomy. That means that an administrator can allow someone to maintain one section of a taxonomy, and that person will not have the ability to change other sections. Likewise, permission can be given for someone to add items within a given section of the taxonomy, but they will not be able to add items to other sections. Finally, sections of the taxonomy can be marked "private" forcing explicit reader permissions to be granted in order to view items. That means that one can have a public view, and then when the visitor logs in, they will see items for which they have been given viewing rights.

The taxonomy also allows an easy browsing interface for users. They may select a "Show All Categories" feature to view the entire taxonomy and the number of items in each node. See Figure 3.

Figure 3: "Show all categories" feature on ELRL

Show all categories

Leadership Development
    Preparation and Training
        Adult Learning (10)
        Coaching Mentoring and Internships (12)
        Instructional Leadership (19)
        Other School Leaders (9)
        Principal Academies (3)
        Professional Development (16)
        University Preparation Programs (8)
    Recruitment
        Diversity In The Workforce (6)
        Selection (13)
        Succession Planning (12)
Leadership Policy Environment
    Partnerships and Networks
        Business and Community Partnerships (19)
        School Boards (6)
        Unions and Labor Relations (9)
    Policymaking and Reform (1)
        Federal Government (12)
        Policymaking Process (4)
        School Reform Models (19)
        State Departments of Education (1)
        State Legislatures (8)
    Sustainability
        Capacity Building (25)
        Evaluation (21)
        Learning Communities (21)
        Organizational Change (14)
        Performance Standards (12)
        School Culture (8)
Project LEAD
    Meetings
        Eugene Grantee Meeting (6)
        Tampa Grantee Meeting (16)

Users can also browse through the taxonomy. On the index page, the top two levels are present, along with some text that frames the entire collection. See Figure 4.

Figure 4: Two-level navigation and explanatory text on ELRL

If the users clicks on one of the top level categories, some further framing (or scoping) text is displayed, along with a list of immediate subcategories. The text is maintained via the administrative web interface. See Figure 5.

Figure 5: First-level categories on ELRL

If the user clicks on one of the second level categories ("Partnerships and Networks" in this case), the next level of immediate subcategories is displayed, along with further text about that category. See Figure 6.

Figure 6: Second-level categories on ELRL

Finally, if the user clicks on the next level, the items are displayed. See Figure 7.

Figure 7: Third-level categories on ELRL, with collection of records

At this point, the user can read summary descriptions of each item, click through to the item itself (e.g. the PDF on another web server), or click "Show details" to view the metadata for that record. See Figure 8.

Figure 8: Details of an item in ELRL

View item details

Title:	An Impossible Job? The View From The Superintendent's Chair
Creator:	Howard J. Fuller with Christine Campbell, Mary Beth Celio, James Harvey, John Immerwahr and Abigail Winger
Description:	Nine out of 10 urban school superintendents say they need more authority to fix bad schools and boost student achievement, according to a survey of the superintendents of the nation's 100 largest districts.
Subject keywords:	Central office administration
Publisher:	Center on Reinventing Public Education
Contributor:	
Date Published:	28-Jul-03
Resource Type:	text/html
Format	text
Identifier:	http://www.wallacefoundation.org/newsroom/news_detail.cfm?id_news=29329412&type=1
Language:	en
	Categories:
	/Leadership Policy Environment/Partnerships and Networks/Unions and Labor Relations
	/Leadership Policy Environment/Partnerships and Networks/School Boards
Information:	Added on: 2003-09-24 17:44:55 Last modified on: 2003-12-12 08:42:24

3.4   Other features

The system also allows users to submit items to a virtual collection; if approved the metadata for the items can be edited, and then the new items will appear.

Users can use an "email a friend" feature to send links to items in collections. Users can subscribe via RSS feeds to see the latest items in a collection or those items that have been featured by the administrators. See Figure 9 for an example of the featured items view from a free web-based RSS aggregator, Bloglines.com.

Figure 9: RSS feed from ELRL, as displayed on bloglines.com

3.5 Metadata

We use Qualified Dublin Core, along with some metadata from an EDC namespace for the collections. Because most collections are built and maintained by people without formal training in library science or metadata, we have sought to keep the metadata is simple as possible. Our current core element set is limited to: Title, Creator, Description, Publisher, Contributor, Date Published, Resource Type, Format, Identified (usually a URL), and Language. Our administrative metadata includes Status (active or inactive), Publish (Boolean, controls whether the item appears in the web catalog), Featured (Boolean, controls whether the item appears on "Featured Items"), Comments, and information about who added and last modified an item, and when these updates took place. Our production system will include our organizational metadata, including Center Name, Center Code, Project Name, Project Code, Funder, and so forth.

4 State of the Project

Development began in the summer of 2003, and the first collection went online in August 2003. The second collection followed two months later. These two collections may be viewed now:

http://cllc.edc.org/rl/
http://library.sprc.org/

We have continued to add minor features and to apply bugfixes as needed. At this writing, two more collections are actively preparing to move onto the system, and four additional collections are planning to do so.

Based on our successful prototyping, we have begun a rewrite of the entire system to ensure that it is fully optimized. We expect this to be complete in summer of 2004. The prototype system can be made available to other institutions upon request now, and we expect to release the final production system via SourceForge and other standard channels once it is complete.

Details on acquiring the current system (as of this writing) or the production system can be obtained from the author. There will be no charge for the software, and the production version will be released on a standard open source license.

5   Conclusion

This software has already begun to yield dividends in terms of ease of management of diverse collections. It suits our decentralized structure well, and may suit the needs of others also. By employing open standards, we meet the needs of our funders and leverage all the strengths of the semantic web. Collections with professional librarians find a comfortable environment to describe their metadata, and collections with content experts adding items are more likely to follow a successful strategy in the use of standard metadata, as opposed to devising one-off systems for each collection.

REFERENCES:

[1] RSS 2.0 Specification. Retrieved May 15, 2004, from http://blogs.law.harvard.edu/tech/rss.

[2] GEM Project Site. Retrieved May 15, 2004, from http://www.geminfo.org/.

[3] See the web site at http://cllc.edc.org/

[4] DC-Library Application Profile. Retrieved May 15, 2004, from http://dublincore.org/documents/2002/09/24/library-application-profile/