Building a Workforce of Information Professionals for 21st Century Global Information Access
(IMLS award 2014-2017)

Jung-ran Park, Chris Yang, Yuji Tosaka, Houda El Mimouni, Qing Ping

College of Computing and Informatics, Drexel University
Current developments and emerging trends in data standards and technologies which play essential roles in managing digital information:

- Newer models and tools for resource discovery and access and data communication such as RDA, BIBFRAME,
- Semantic Web technologies such as Linked Data, RDF, OWL, SKOS and SPARQL.
Agreement on whether the implementation of data standards & Semantic Web technologies represent an opportunity for the cataloging/metadata profession

Strongly agree, 52%
Agree, 37%
Neutral, 6%
Not sure, 4%
Disagree, 1%
Strongly disagree, 0%

N = 678
## Participants’ perceived potential benefits of the Semantic Web to their institutions and their user community

<table>
<thead>
<tr>
<th>Benefits</th>
<th>Very &amp; important</th>
<th>Not important</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved user services</td>
<td>616</td>
<td>4</td>
</tr>
<tr>
<td>Improved data/resource discovery</td>
<td>601</td>
<td>7</td>
</tr>
<tr>
<td>Enhanced discovery services through federated or Web-scale searches</td>
<td>555</td>
<td>11</td>
</tr>
<tr>
<td>Increasing the value of library data and their presence on the Web</td>
<td>554</td>
<td>12</td>
</tr>
<tr>
<td>Reducing redundancy and improving efficiency of bibliographic descriptions</td>
<td>507</td>
<td>27</td>
</tr>
<tr>
<td>Having a richly linked metadata description</td>
<td>506</td>
<td>15</td>
</tr>
<tr>
<td>Providing authority data for names and subjects with unique identifiers</td>
<td>450</td>
<td>39</td>
</tr>
<tr>
<td>Linking multiple domain-specific knowledge bases to support interdisciplinary research and creation of new knowledge</td>
<td>415</td>
<td>26</td>
</tr>
<tr>
<td>Supporting multilingual functionality for data and user services</td>
<td>294</td>
<td>87</td>
</tr>
<tr>
<td>Re-using and/or combining data contributed by non-library communities</td>
<td>281</td>
<td>92</td>
</tr>
</tbody>
</table>
Issues in Accessing CE Resources

- Information about continuing education resources is highly fragmented and scattered on the Web.
- Mainly compilations of resources; limited scalability and maintenance.
- Subject coverage mostly focuses on traditional cataloging but not emerging data models like the Semantic Web and Linked Data.
- Not designed to feature any mechanism for developing virtual communities.
Objective 1: Searchable Digital Repository

- Allows practitioners to seamlessly access professional development resources and opportunities by converging and linking scattered resources on the Web for self-directed lifelong learning.
- Searchable repository through metadata description and management.
- Designed with self-sustaining mechanisms to ensure its scalability, maintenance, and sustainability.
- Focuses on new and emerging data standards and technologies encompassing Semantic Web and Linked Data related topics.
Top 10 CE topics in which participants have expressed personal interest

<table>
<thead>
<tr>
<th>Rank</th>
<th>Topic</th>
<th>Interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Overview of current/emerging data standards and technologies</td>
<td>593</td>
</tr>
<tr>
<td>2</td>
<td>RDA (Resource Description and Access)</td>
<td>579</td>
</tr>
<tr>
<td>3</td>
<td>Linked data applications</td>
<td>553</td>
</tr>
<tr>
<td>4</td>
<td>Metadata standards (e.g., EAD, MODS)</td>
<td>540</td>
</tr>
<tr>
<td>5</td>
<td>BIBFRAME</td>
<td>533</td>
</tr>
<tr>
<td>6</td>
<td>Metadata project management (planning, implementation, and quality control)</td>
<td>509</td>
</tr>
<tr>
<td>7</td>
<td>RDF (Resource Description Framework)</td>
<td>505</td>
</tr>
<tr>
<td>8</td>
<td>Semantic Web applications in libraries (e.g., projects, techniques)</td>
<td>498</td>
</tr>
<tr>
<td>9</td>
<td>Data management</td>
<td>473</td>
</tr>
<tr>
<td>10</td>
<td>Semi-automatic metadata generation and tools</td>
<td>472</td>
</tr>
</tbody>
</table>
It is essential that CE practices at the individual level be embedded within a collaborative framework necessary to help diffuse the latest advances among metadata communities.

- Collaborative virtual space for professionals to communicate, mentor, and share library projects, applications and best practices on new and emerging standards and technologies.
- Incorporated into the project repository
## Perceived barriers to continuing education in new standards and technologies

<table>
<thead>
<tr>
<th>Barriers</th>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>The cataloging and metadata communities will benefit from some kind of mentoring system to bring more professionals up to date in this subject area</td>
<td>516</td>
<td>14</td>
</tr>
<tr>
<td>We need a portal for continuing education resources and opportunities in this subject area</td>
<td>513</td>
<td>11</td>
</tr>
<tr>
<td>The cataloging and metadata communities need to collaborate more in sharing knowledge and experience in this subject area</td>
<td>487</td>
<td>34</td>
</tr>
<tr>
<td>Available online resources offer limited coverage, organization, and sequence for effective continuing education in this subject area</td>
<td>390</td>
<td>58</td>
</tr>
<tr>
<td>My institution is not sure about how to take advantage of new standards and SemanticWeb technologies, so I am not sure whether I should invest the time and effort in this area</td>
<td>336</td>
<td>125</td>
</tr>
<tr>
<td>I have difficulty finding best practices, project reports, and new initiatives in this subject area</td>
<td>328</td>
<td>131</td>
</tr>
<tr>
<td>My institution does not provide sufficient support (financial support or release time) for continuing education in this subject area</td>
<td>306</td>
<td>207</td>
</tr>
<tr>
<td>I am not aware of available continuing education resources and opportunities in this subject area</td>
<td>232</td>
<td>264</td>
</tr>
</tbody>
</table>
What participants would like to see in the Project Repository?

- Online tutorials for self-paced learning
- A series of self-paced learning modules that cover emerging...
- Learning resources for free download
- Basic and advanced searching capabilities
- Self-paced learning modules that are properly sequenced and...
- Information on CE opportunities
- Browsing capability
- Learning community that fosters communication, collaboration,...
Objective 3: Open Source Webinars

- Short self-paced online courses and programs are the preferred CE methods for practitioners.
- Professional development materials especially in emerging standards and technologies tend to be disjointed and may lack sequence of the subject which hinders self-paced learning.
- Three asynchronous open source webinars with sequential and progressive learning modules will be instrumental for metadata professionals and students to build in-depth knowledge and skills necessary for organizing and managing burgeoning data and digital information.
Project Repository

• Built on *DSpace* software and hosted on a server at CCI.
• Developed automated workflow will continually populate the repository with continuing education resources.
• Used a *rapid prototyping* approach to develop the automatic crawling and monitoring system.
• Designed with self-sustaining and governing mechanisms to ensure its scalability, maintenance, and sustainability.
Intelligent techniques:

- Web crawler that utilizes the starting Web page of a search as the basis of finding other Web sites related in subject matter to the first. (Chen & his colleagues, 1998)
- Link structure of Web documents and the similarity of contents between relevant Web sites to improve search results (Jamali et al., 2006)
- Utilization of Tf-Idf and TIDS algorithms in conjunction with repeated crawlers to improve search results (Kumar & Vig 2013)
- Automatically map disparate category trees into a single hierarchy, thereby improving search. (Yang and Liu, 2009; Yang et al. 2010)
• Has the ability to identify high-quality Web objects relevant to current and emerging metadata standards, tools, Semantic Web technologies.

• Seed information—URLs based on Google search using four keywords: RDA; RDF; XML; Web Ontology Language.

• Filtered authoritative web sites relevant to continuing education resources submitted to the crawling system.

• Intelligent content-based algorithm to select optimum paths to extract, retrieve & classify high-quality relevant Web objects based on content, anchor similarity with the seed-pages and the popularity of crawled page among seed pages.
1. Content similarity between crawled pages and seed pages

We first take the whole set of seed pages as one seed document. Each new crawled document (Web page) was compared with the seed document to determine their content similarity.

2. Popularity of crawled pages among seed pages

Indicates their structural characteristics in the local network; the more direct links there are from the seed pages to the crawled page, the higher scoring is assigned to the crawled page.
Select Seed Pages → Outgoing Links → Crawled Pages

Stopping Criterion:
- Ranking scores below threshold
- Sufficient crawled pages

Select top K pages to seed pages

Ranking Criterion:
- Votes from precedent pages
- Content similarity
- Anchor similarity

Selection Criterion:
There is great drop in the ranking score
Continuing Education in New Data Standards & Technologies

Prototype
Conclusion and Future Study

• The development of this crawler is only one aspect of this project

• Next steps:
  o include the development of mechanisms for the ingest of identified resources within the repository,
  o the generation of metadata for each resource,
  o the development of a user interface to facilitate the use of the repository and
  o the development of a collaborative virtual space

• In its final form, the repository will constitute a self-sustaining continuing education resource for the metadata community.