DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2010

Use of Community Metadata: Public Policy Research Documentation in PolicyArchive

Sarah Buchanan
Center for Governmental Studies
United States
sarahab@ucla.edu

Abstract

PolicyArchive collects public policy research from over 800 known research publishers and makes these documents accessible in a navigable digital library. The contributions of thousands of publications from these providers enable in-depth secondary source materials to be utilized by policymakers, legislators, foundations, scholars, journalists, and educators. The functionality of this digital repository is discussed, including the use of terminologies, subject navigability, and Special Collections. PolicyArchive unites subject content with metadata and is openly accessible; the application of these principles not only provides coordinated access to previously unavailable resources, but also allows the reader to place a given document in multiple contexts. Analysis of this information environment illuminates ongoing digital library initiatives regarding the creation of navigable, accessible learning resources.

Keywords: policy documentation; policy research; legislative research; public policy; public affairs; digital publication archive; digital library.

1. Introduction

A wide range of institutions including think tank, university, government, and foundation-funded organizations engage in policy research, and produce a staggering number of publications per year. Among this community of practice, valuable research is distributed through a diverse network of libraries, institutions, databases, and websites both behind and in front of secure networks. Reflecting this, policy researchers have not had access to sources in a centralized database comparable to that which is found in the law, medicine, and science disciplines. A 2005 survey of 39 foundation-funded policy research organizations performed by the non-profit Center for Governmental Studies showed that individual organizations publish on average 73 documents each year (range 2 to 750), or about 2,850 documents annually (Rivera, 2008). These are the results of some $1.5 billion spent per year by philanthropic foundations on research.

Yet preservation and long-term access concerns have been absent or sporadic in strategic plans, and until recently, the lack of cross-institutional access has impeded the sharing and awareness of resources. The institutional barriers that have until recently precluded broad-scale participation in a union repository are numerous, but most importantly, they reflect the singular missions of each organization rather than the needs of the field as a whole. The underlying problem of silo-based data presentation becomes most pressing when researchers both within the academy and in government seek sources and data produced by an ever-broadening range of organizations. The need for a coordinated, accessible repository has been expressed among policymakers, legislators, foundations, scholars, journalists, and educators in communications with the Center for Governmental Studies (CGS, est. 1983). The present project originated as a collaboration between a nonprofit organization and an academic library.

PolicyArchive [http://www.policyarchive.org] is a repository built on the principle that open access to secondary source material - research reports, data, community analysis - enables more effective and accountable public policymaking. PolicyArchive enables legislators and policymakers to enact well-informed measures that are grounded in consultation with published studies. The platform is the first and most comprehensive free, searchable archive of public

https://doi.org/10.23106/dcmi.952109768

policy research. On June 19, 2008, PolicyArchive officially launched its Web site, then featuring more than 12,000 research documents in 24 topics and over 300 subtopics. Active efforts among contributors and archive staff now provide electronic access to 30,000 publications from over 800 distinct publishers, representing all shades of the organizational and political spectrum. Two other repositories exist which might provide an environmental barometer with which to evaluate the archive. These repositories offer discrete breadth (FOLIO [https://folio.iupui.edu/] provides foundation-supported research; IssueLab [http://www.issuelab.org/] provides nonprofit organizational research) and depth of content (FOLIO: 3,000 documents as of 2008; IssueLab: 3,700 currently). PolicyArchive serves as an aggregator of social policy research, a digital library of information resources, and an effective leader in archival outreach to new content contributors. With these current operations, several key adoptions of infrastructural support and application of community contributors' expertise have enabled the archive's growth and continued provisioning of added value.

2. PolicyArchive Collection Methodology

Through research and outreach to organizations involved in public policy, PolicyArchive acquires new documents for inclusion in the archive. In addition to the general repository, PolicyArchive develops Special Collections, which serve as directories for subject-specific collections. In support of the overall goals of PolicyArchive, contributors are encouraged to submit content for permanent storage, by supplying bitstreams and metadata through registration online and manual upload of items. If such deposit is not possible, contributors can supply external links back to the originating Web site, according to a distribution agreement. PolicyArchive provides a searchable interface which accesses standard parts of the document, such as its abstract, and provides subject indexing and full-text search for held items. Content deposited in PolicyArchive has grown at the approximate rate of 10,000 items per year; there is currently no cost to contribute content. Following contributor submission, quality control managed by staff during the workflow process, includes supplying information missed initially yet necessary from a retrieval standpoint, and typically includes analysis of subject and publication, standardizing proper names, and applying relevant terminologies. Terminology efforts have included the creation and use of "community-based metadata," as well as the coordination of metadata terms with content providers such that documents are made accessible within the context of the originating organization. Community-based metadata reflects the archive's focus on providing access, maintaining intact provenance, and serving as a bridge between research producers and research seekers.

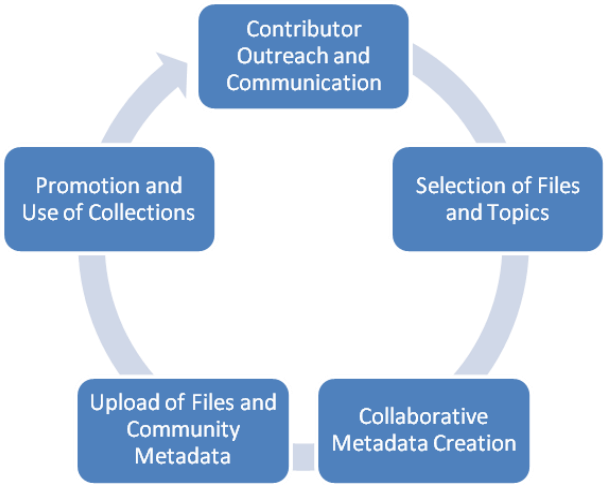

Fig. 1. Data flow in PolicyArchive

https://doi.org/10.23106/dcmi.952109768

3. Ingest procedures

Methods for depositing content include the use of an online, menu-based form for individual documents, and a batch upload form for multiple documents in aggregate. Supplementary files (e.g. multimedia, press releases, newsletters, data sets, and previous versions) can be included. Hard-copy documents are accommodated through scanning. Following transcription and analysis of the item in-hand, documents and/or a cover image (PDF, JPEG, GIF, or PNG) are supplied. For items in Special Collections, an additional collection-level field, dc.relation.ispartof, is completed. Figure 2 illustrates the procedure for submitting content and associated metadata through the PolicyArchive Web site.

Fig. 2. Item submission procedure

Once reviewed and posted by PolicyArchive staff, the documents are fully searchable in terms of both supplied metadata and full-text content. Currently some organizations prefer to retain documents on a separate public site for varied reasons (e.g. content not suited for PDF; interrelated content pages hosted elsewhere, extent, and presentation preference). For some 3,000 items, PolicyArchive provides metadata and the item's bitstream is available through linked documents via a URL in the final "View Publication" feature. Linked documents are subject to the possibility of external URL renaming, and file text is not indexed locally. The archive's HandleServer configuration and profile facilitates individual URI assignments (see Sauermann and Cyganiak, 2007 and Summers, 2008 for discussion of element linkages in the semantic Web).

3.1. Field Selections

The Dublin Core (DCMI) metadata schema (DCMI, 2005) enables information providers worldwide to exchange data in a common format. The archive's metadata registry is a version of the qualified Dublin Core schema (DCMI, 2008), which is a necessary component of the software utilized, the open source DSpace. As public policy research calls for a unique set of identifying information, the archive utilizes a modified qualified DC schema to express the descriptive elements of its documents. [Note: elements expressed below follow DSpace practice, which in many cases is not compliant with DCMI recommendations.] The following fields (first column, Table 1) are chosen to be indexed and retrievable through site searches. A localized DC schema (second column) frames the population of fields with controlled vocabulary terms specified locally.

Table 1. Vocabulary Architecture: Item Metadata Table

Public fields	**Local encoding of metadata value (DC field)**
1. Title	Supplied by document (dc. title)
2. Publication date	yyyy-mm-dd (dc. date. issued)
3. Authors	SONF procedure (dc. contributor. author)
4. Abstract	Supplied by document (dc. description. abstract)
5. Series/Report No.	Series information note supplied by document (dc. relation. ispartofseries)
6. Publisher	SONF procedure (dc. publisher)
7. Funder	SONF procedure or leave blank if same as Publisher (dc. description. sponsorship)
8. Document provider	SONF procedure (dc. contributor. other)
9. Special Collections	Currently 4 possible assignments (dc. relation. ispartof)

https://doi.org/10.23106/dcmi.952109768

10. Coverage Area	SPN; U.S. states during submission; country and regional names via manual edit (dc. coverage. spatial)
11. Subject Keywords	Collaborative by archive staff and contributors (dc. subject)
12. Topics	PolicyArchive Topic List (dc. subject. other)
13. Type of Item	PolicyArchive Controlled Type List (dc. type)
14. Identifiers	ISBN, ISSN, ISMN, Gov't Doc #, URI, Other (dc. identifier. xxxx)
DSpace automatic fields	
	Time of ingest (dc. date. accessioned)
	Time of administrator approval (dc. date. available)
	Submitter and approver ID and date (dc. description. provenance)
Item's Permanent Link	URI assigned per Handle System (dc. identifier. uri)
External File Link	Hyperlink (dc. relation. uri)

Notes: The SONF (Standardized Organization Name Format) procedure (nos. 3, 6, 7, 8) is a local method of determining organization/corporate names, developed in coordination with librarians at IUPUI (Indiana University-Purdue University Indianapolis). It entails utilizing the following resources in order of preference: 1) Library of Congress Authority Files [http://authorities.loc.gov/], and 2) the Guidestar database [http://www2.guidestar.org/Home.aspx] of 1.8 million IRS-recognized organizations. If name is not found in either source, PolicyArchive determines the standardized heading from 3a) the name on the document, 3b) the organization's website, and 3c) AACR2 naming rules.

SPN (Standardized Place Names) (no. 10) is a local term equivalent to our use of Library of Congress Authority Files.

The PolicyArchive Topic List (no. 12) is a local, dynamic list of categories applied to documents. It uses elements from PAIS (Public Affairs Information Service), an information organization system maintained by CSA Illumina (Cambridge Scientific Abstracts) [http://www.csa.com/factsheets/supplements/paisbroadtopics.php], as well as from MESH (NIH Medical Subject Headings) [http://www.nlm.nih.gov/mesh/], for health policy content. Additions are made through a global change.

The PolicyArchive Controlled Type List consists of the following local types: Audio, Book, Book chapter, Book review, Brief, Fact sheet, News release, Newsletter, Other, Report, Speech, Testimony, Thesis/Dissertation, Video.

The above quantity of metadata is completed in both manual and batch uploads of new content. A manual upload typically takes about three minutes to complete, with all relevant data in hand. Some administrative metadata fields are not part of the public view. Following approval in workflows by the PolicyArchive staff, the submitter receives a confirmation email containing a link to the item.

4. Community Metadata Application: Subject Keywords

In the determination of subject keywords for items, PolicyArchive consults both controlled vocabularies and the contributing organization's programs and emphases (Harper and Tillett, 2007). In many cases, the general topic derived from the PAIS vocabulary - e.g. "housing" - may still be too broad to provide users with a focused set of key documents related to a more focused area of inquiry. The archive's innovative approach lies herein with the use of community-based metadata. In creating these item-level keywords - an applied vocabulary - the archive and organizational representatives collaborate, guided by knowledge of specific and intended research inquiries. The results of this cooperation are distinct organizational vocabularies applied manually to documents. The "housing" topic, in the example above, would be complemented by new application of the keyword "foreclosure prevention," a term which would aid a researcher seeking documents on this topic alone [see https://www.policyarchive.org/handle/10207/7379].

https://doi.org/10.23106/dcmi.952109768

Through this process, metadata is applied which reflects the granularity of the document, and allows the document to become retrievable in more focused search sets. Because the archive is a repository for primarily written and narrative documents, the use of controlled vocabularies developed for bibliographic formats is appropriate. In particularly statistical or data-rich documents, subject analysis can provide a crucial avenue for exposing the relevance of the document to other similar publications.

4.1. Special Collections

In addition to review of submissions and associated metadata, staff facilitates the assignment and implementation of Special Collections. In particular, the Presidential Advisory '08 collection [http://www.policyarchive.org/collections/presidential/] presents notable policy recommendations around key social issues produced by several organizations. Four other Special Collections present documents by a single organization; each collection interface is designed collaboratively. Continued expression and evaluation of research need facilitates the creation of distinctive Special Collections, noted for their interconnectivity with the PolicyArchive general collection.

5. Records Retrieval

Both users and metadata administrators can browse complete listings of topics, subtopics, publishers, and funders for consistency through the "Browse by," Advanced Search, and Quick Links features. Batch uploads (approx. 50 items) can be immediately surveyed to resolve both small and large issues related to metadata - including proper date displays and standardized organization name, abbreviations, and punctuation. Smaller sets of manual items can also be retrieved for editing.

6. Future Work and Conclusion

PolicyArchive engages in continued outreach to new and existing content providers. PolicyArchive provides information seekers of public policy research with an archive of documents representing varied publishers and subjects. The archive faces challenges and opportunities in the form of developing additional Special Collections, acquiring new quality *full-text* contributions, expanding the use of subject keywords to a fuller portion of the archive, and facilitating storage capacity. PolicyArchive is active on several social media platforms, including a Twitter page which provides synchronous and automated posting of new content via URL shorteners. A periodic newsletter provides updated developments, and a small staff engages in communication with future contributors. Attention in the popular press (Scribner, 2010) to librarians' and metadata specialists' role in providing access to information has gained broad visibility - and a new wave of popular support may yet keep libraries and archival repositories connected to citizens' daily lives.

Acknowledgements

The author gratefully acknowledges Romulo Rivera and Tracy Westen for their support, and the anonymous reviewers for their helpful comments.

References

DCMI. (2005). Using Dublin Core: Usage guide. Retrieved February 2, 2010, from http://dublincore.org/documents/usageguide/.

DCMI. (2008). DCMI Metadata Terms. Retrieved February 2, 2010 from http://dublincore.org/documents/dcmi-terms/.

Harper, Corey, and Barbara Tillett. (2007). Library of Congress controlled vocabularies and their application to the semantic web. Cataloging and Classification Quarterly, 43(3/4), 47-68.

Rivera, Romulo. (2008). PolicyArchive Plan. Los Angeles: Center for Governmental Studies.

https://doi.org/10.23106/dcmi.952109768

Sauermann, Leo, and Richard Cyganiak. (2007). Cool URIs for the semantic web. Retrieved February 2, 2010 from http://www.w3.org/TR/cooluris/.

Scribner, Sara. (2010). Saving the Google students. The Los Angeles Times, Mar. 21, 2010. Retrieved Mar. 22, 2010 from http://www.latimes.com/news/opinion/commentary/la-oe-scribner21-2010mar21,0,764753.story.

Summers, Ed. (2008). Following your nose to the Web of Data. Information Standards Quarterly, 20(1). Retrieved February 2, 2010 from http://inkdroid.org/journal/following-your-nose-to-the-web-of-data.

https://doi.org/10.23106/dcmi.952109768