Integrating Harvesting into Digital Library Content

David A. Smith, Anne Mahoney, Gregory Crane
Perseus Project
Tufts University
Medford, MA 02155
{dasmith,amahoney,gcrane}@perseus.tufts.edu

ABSTRACT

The Open Archives Initiative has gained success by aiming between complex federation schemes and low functionality web crawling. Much information still remains hidden inside documents catalogued by OAI metadata. We discuss how subdocument information can be exposed by data providers and exploited by service providers. We discuss services for citation reversal and name and term linking with harvested data in the Perseus Project's document management system and a proxy service for automatically adding these links to OAI documents outside Perseus.

Categories and Subject Descriptors

D.2.12 [**Software Engineering**]: Interoperability—*Data mapping*

General Terms

Design

Keywords

Open Archives Initiative, automatic linking

1. INTRODUCTION

From its roots in the scientific e-print community, the Open Archives Initiative (OAI) has grown to hold a significant place in digital library interoperability efforts. By lowering the barrier for *data providers* to expose metadata, the OAI Metadata Harvesting Protocol provides a consistent testbed for new *service providers* [6]. The OAI has thus successfully positioned itself between high-functionality, complex federation schemes and low-functionality, web crawled search services [1].

Harvesting document metadata has proved successful especially where documents are about a single topic and relatively short. Although these conditions hold for most scholarly articles in e-print archives, the primary data, for example, on which such articles are based are not as tractable. First of all, primary sources are often quite long and amorphous and not about any single topic. (Is Stevenson's *Kidnapped* really *about* the 1745 uprising in Scotland?) Secondly, scholarly arguments often depend on citing specific passages, lines, or words in a source. Links from text span to text span — rather than document to document — are needed to follow many chains of reasoning. Finally, readers may want linking and visualization services while browsing.

Since its inception in 1987, the Perseus Project has worked on integrating services directly into reading environments and on searching and visualizing complex documents. In this paper, we describe first how our document management system extracts information from various structured documents and how we expose that information as a data provider. We then discuss our experiments with harvesting OAI repositories and integrating the harvested data into the Perseus Digital Library. We show how information extraction services can also be provided for content outside the digital library.

JCDL'02, July 13-17, 2002, Portland, Oregon, USA.

2. DL AS DATA PROVIDER

In the past two years, we have been developing a document management system, the Hopper, for indexing and retrieving information in a variety of formats [7]. The Hopper is in production use on the Perseus website (`www.perseus.tufts.edu`), which receives over nine million page views per month. Since the Hopper uses the Resource Description Framework as its metadata model and follows Dublin Core for most semantics, it was easy to become a registered data provider, as well as a founding member of the Open Language Archives Community. We expose metadata in OAI-standard unqualified Dublin Core and in the OLAC schema.

In addition to storing document level metadata, the Hopper extracts several kinds of information from document contents. First, it detects markup indicating subsections about particular topics. These divisions are most obviously useful for encyclopedias and dictionaries: our metadatabase is aware of an article about "coinage" in a classical encyclopedia or one on "Fleet Street" in a survey of London. Second, the Hopper records citations in document content so that the cited documents can be linked back. Finally, the system identifies and disambiguates place names and dates to produce map and timeline visualizations of documents and corpora [2]. Since many documents contain tens of thousands of citations, place names, and dates, we do not currently expose these metadata through the OAI protocol. We are, however, experimentally exporting subdocument information for articles and chapters within larger documents.

extreme in the world (Hidy and Klieforth 1990). The Sierra Nevada is primarily responsible for creating this arid, continental climate by capturing moisture from Pacific storm fronts before the moisture reaches the desert (Houghton et al. 1975); similarly, the Rocky Mountains intercept storms from the Gulf of Mexico (Hidy and Klieforth 1990). Local weather patterns are complicated by the mountain ranges that uplift the dispersed moisture, creating mountain storms (Hidy and Klieforth 1990). Thus, precipitation increases with elevation (Billings 1951), and average annual precipitation can be highly variable over small distances.

granite ranges occur in the west, basalt ranges occupy the northwest, rhyolite mountains form the center, and limestone mountains dominate

Figure 1: USGS PDF document with term links

3. DL AS SERVICE PROVIDER

The Hopper also functions as a service provider, integrating harvested metadata into searching and browsing. When a reader ventures into an unfamiliar discipline, unknown names and terminology can hinder comprehension. Users can also browse terms and named entities to discover documents covering similar topics [4]. The Hopper culls names and terms from the metadata of its documents and creates links in texts that use those words or phrases. Harvested metadata increases the set of terms to be linked to include the titles, authors, and subject keywords of federated documents. When reading a speech by Demosthenes, for example, the user can click on the highlighted term "Areopagus" and link to pictures of the site of this law court in Athens and also to an article in the Stoa Consortium (`www.stoa.org`) about the history and procedures of the court.

As a member of the Open Language Archives Community, Perseus is interested in linking together disparate resources for the study of historical and minority languages. Latin text in a Stoa publication, for example, can link each word to grammatical help and a dictionary in Perseus. Simply noting the presence of a Latin dictionary and linking all Latin words to it is not enough, however; the Renaissance Latin texts published by the Stoa contain many words not in the classical Lewis and Short dictionary in Perseus. Many links are thus made to non-existent dictionary entries. To give the user a better idea of what words can be successfully glossed, either the Perseus data provider would have to expose all dictionary headwords as metadata (68,000 entries for just one of Perseus' Latin dictionaries), or the Stoa would need some other mechanism to obtain this information.

Citation extraction is a service of interest to many digital libraries, as evidenced by the popularity of the ResearchIndex (CiteSeer) citation linking service [5]. While the links a document makes could be exposed as metadata, this has some disadvantages. The context in which a link is made can carry important information about its usefulness. A citation made in the body of a text may be more relevant to the main argument than a link in a footnote; a citation that also quotes a passage from a text or occurs in the abstract may indicate an even closer connection between source and target. Also, the citations in many heavily-used reference works take up more than half the space of the whole file.

The kinds of information that the Hopper extracts from documents are certainly not exhaustive. Also, as noted above, some types of metadata approach the size of the full content. In order to encourage new, more complex services, as well as for reasons of efficiency, we are experimenting with extracting information from full text in other digital libraries. Most OAI data providers, while they link to HTML displays, do not provide links from their metadata records to full content, even when a structured SGML/XML file underlies the data. While a full-text file is often available somewhere on the data provider's site, finding it would revive the need for web crawling, which the OAI is meant to obviate. Although we have demonstrated the Hopper's ability to manage content in disparate formats by the integration of SGML documents from the Library of Congress' "American Memory" and UNC's "Documenting the American South" and of PDF documents from the US Geological Survey, these will remain ad hoc examples without the ability to discover and acquire content through a standard protocol.

Even though we can seldom harvest full text for analysis, our service provider allows readers, once they have discovered OAI documents in our search interface, to browse them through a proxy service. The proxy inserts links to other Perseus and OAI resources into HTML and PDF documents (figure 1). The system adds links anchored on the types of phrases mentioned above: subject labels, section headings, document titles, etc. It is interesting to compare this system to [3], which makes links based on existing links in other documents. Phrases are placed in a flat key-lookup database indexed by the first non-stopword of the phrase. As with documents in Perseus, remote documents are tokenized into markup and words. The system skips over text that is already linked and tries to match the longest possible phrase at each word position. Currently, the service allows users to restrict the set of phrases to be matched to terms from the whole digital library or from individual collections.

The proxy service only sends information in one direction. We are now experimenting with providing linking, term detection, and name disambiguation as SOAP-based web services for content providers. By exposing full content to digital library services, data providers can create a richer testbed for services and greater enhancements for users.[1]

4. REFERENCES

[1] W. Y. Arms, D. Hillmann, C. Lagoze, D. Krafft, R. Marisa, J. Saylor, C. Terrizzi, and H. Van de Sompel. A spectrum of interoperability: The Site for Science prototype of the NSDL. *D-Lib Magazine*, 8(1), January 2002.

[2] G. Crane, D. A. Smith, and C. E. Wulfman. Building a hypertextual digital library in the humanities: A case study on London. In *Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries*, pages 426–434, Roanoke, VA, 24-28 June 2001.

[3] S. R. El-Beltagy, W. Hall, D. D. Roure, and L. Carr. Linking in context. In *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia*, pages 151–160, Århus, Denmark, August 2001.

[4] R. Gaizauskas, P. Herring, M. Oakes, M. Beaulieu, P. Willett, H. Fowkes, and A. Jonsson. Intelligent access to text: Integrating information extraction technology into text browsers. In *Proceedings of HLT*, San Diego, CA, 2001.

[5] C. L. Giles, K. Bollacker, and S. Lawrence. CiteSeer: An automatic citation indexing system. In *Proceedings of the 3rd ACM Conference on Digital Libraries*, pages 89–98, Pittsburgh, PA, 1998.

[6] C. Lagoze and H. V. de Sompel. The Open Archives Initiative: Building a low-barrier interoperability framework. In *Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries*, pages 54–62, 2001.

[7] D. A. Smith, A. Mahoney, and J. A. Rydberg-Cox. Management of XML documents in an integrated digital library. *Markup Languages: Theory and Practice*, 2(3):205–214, 2000.

[1] A grant from the Digital Library Initiative Phase 2 (NSF IIS-9817484), with particular backing from the National Endowment for the Humanities, supported this work.