S

Searching Digital Libraries

PANAGIOTIS G. IPEIROTIS
Department of Information, Operations and Management Sciences, Leonard N. Stern School of Business, New York University, NY, USA

Synonyms

Federated search

Definition

Searching digital libraries refers to searching and retrieving information from remote databases of digitized or digital objects. These databases may hold either the metadata for an object of interest (e.g., author and title), or a complete object such as a book or a video.

Historical Background

The initial efforts to standardize and facilitate searching of digital libraries date back to the 1970s, when the development of the Z39.50 protocol started. The Z39.50 protocol is an ANSI standard and defines how to search and retrieve items from a remote database catalog. The Z39.50 protocol was widely deployed within library environments, allowing users to perform searches to remote libraries.

With the advent of the Web, libraries started digitizing and making contents available on the Web, and the Z39.50 protocol started losing its importance. Many libraries made their content "searchable" through standard Web forms, allowing users to search and retrieve content using simply a Web browser. However, due to the lack of a link structure, the contents of the libraries remained "hidden" from the modern search engine crawlers, forming part of the "Hidden-Web" (also known as Deep Web, or Invisible Web). Searching across multiple Hidden Web databases, despite the tremendous progress since 2000, is still an open research problem.

However, achieving interoperability across all Web databases is inherently harder than achieving interoperability across library databases, which are relatively more homogeneous. Therefore, a set of efforts focused on introducing protocols to facilitate integrating and searching digital libraries. The Open Archives Initiative focused on defining a protocol for exporting metadata about the objects in the collections hosted by each library. The SRU protocol aims to modernize the Z39.50 by making it similar to modern Web services. Such efforts allow programmers to leverage their existing skills and develop easier tools for the library market.

Scientific Fundamentals

Digital libraries host a variety of digital objects, including, but not limited to, textual documents, images, sounds, videos, or even multimodal objects that combine the above. The concept of searching digital libraries may refer either to the action of searching a *single* digital library or to the action of searching across *multiple* digital libraries.

Searching a *single* digital library typically refers to the action of searching and browsing the contents of the underlying relational, textual, or multimedia database. The interested reader may review the corresponding entries of this encyclopedia for further details on this topic.

Searching across *multiple* digital libraries is a concept that evolved significantly over the years. The development of these efforts is broadly divided in three periods:

- *The pre-Web period (late 1970s–mid 1990s)*: Development of the Z39.50 standard.
- *The early-Web period (mid 1990s–early 2000s)*: Emergence of the Web, and increased accessibility of libraries over the Web.
- *The Web-services period (early 2000s–now)*: Definition of protocols for Web services, and development of library-focused search and discovery protocols.

The Pre-Web Period

The first attempts to define a standardized, common protocol for searching library databases date back to the 1970s. Then, the "Linked Systems Project" examined how to provide support for standardized access method to a small set of homogeneous, bibliographic databases. This effort led to the formation of a NISO committee in 1979, which after years of efforts defined the "American National Standard Z39.50, Information Retrieval Service Definition and Protocol Specifications for Library Applications" in 1987. The protocol was later revised in 1992, in 1995, and in 2003. (See [1] for a detailed history and timeline of the development of Z39.50.)

The Z39.50 protocol was designed as a client-server protocol, defining how the client can search and retrieve information from a remote database. The protocol supports a significant number of actions, including searching across individual fields, such as author, abstract, title, and so on. Unfortunately, the protocol did not mandate the implementation of several aspects of the specifications, allowing the developers to choose the aspects of the protocol to implement. This led to unexpected behavior of some systems, as the same query, executed over the same underlying content, could return very different results, depending on the implementation. Furthermore, the extremely heavy specification made it difficult for vendors to develop systems that were fully compatible with each other.

The Early-Web Period

The emergence of the Web changed significantly the way that digital libraries make their content available. Many libraries, perhaps encouraged by the *Digital Libraries Initiative* in 1994, started digitizing and making their content available over the Web. This meant that user could simply visit the Web site of a library and then, using simply Web forms, could query and browse the holdings of the library.

A significant fraction of these new digital libraries are only accessible via a search interface and the ability to browse through a static hyperlink structure is often missing. This means that the contents of these libraries are "hidden" from search engines, since traditional crawlers, which discover new pages by following links, cannot discover the contents of the library. Such libraries are part of the *hidden-Web* [2]. On the other hand, libraries that provide a link structure for accessing their holdings, are part of the *surface Web*, which is accessible by using general search engines, such as Google.

For libraries with content available as part of the *surface Web*, the common model for searching is through vertical search engines. The vertical search engines create topically-focused indexes of the material available on the Web by using *focused crawlers* [3] to identify and index the pages about a given topic. Under this model, the distributed digital libraries become searchable through a centralized search interface that indexes the remotely stored content. When a user issues a query, the vertical search engine identifies the most relevant pages in the index and returns to the user the URLs of the pages, which are stored remotely.

For libraries with *hidden Web* content, the typical way of searching their contents is through *metasearchers*. A complete metasearcher has to perform the following tasks:

- Discover the available digital libraries. This involves crawling the Web to identify pages with Web forms that are search interfaces for underlying databases [4].
- Understand the capabilities of the available query interface [5–7].
- Characterize the contents of the underlying database, typically by extracting a small sample of the stored contents through query-based sampling. The characterization may involve classifying the database into a topic hierarchy [8], extracting a statistical summary of the content [9,10], or it may involve keeping the actual sample as a surrogate for the contents of the database [11,12].
- Use the database characterization to select the most promising databases for evaluating a given query [11,13].
- Evaluate the queries in the selected databases, retrieve, and merge the results from multiple databases into a single list [14].

An alternative approach to the distributed search technique adopted by metasearchers is to try to download *all* contents of a hidden Web database [15]. Once all the contents of the remote digital libraries are retrieved and stored locally, the problem of searching multiple digital libraries is reduced to the problem of searching a single, centralized database. One of the issues in this case is the need to periodically refresh the local copy with the most recent contents of the remote database [16].

The Web-Services Period

During the early-Web period, the problem of integrating and searching across digital libraries was similar to the problem of integrating Web databases at large. The vision of the *semantic Web* promised a solution for this problem, and the implementation of a *Web services* framework was a first step towards this direction.

Inherently, though, the library integration problem is much easier than the problems involved in the full implementation of the semantic Web. Therefore, a set of niche solutions were developed for the library integration problem, focusing on the one hand on library-specific needs, but building on top of the existing tools for general Web services that are being developed and rapidly improved.

One of the first attempts to make effortless the discovery of the contents of a library database was the development of the *Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)*. This protocol defines how a library can export metadata descriptions of its holdings. Then, *metadata harvesters* can easily collect the contents of the database and make these contents searchable through a centralized search interface. The OAI-PMH protocol is now widely adopted by many libraries and a set of OAI registries facilitate even further the discovery of libraries that support this protocol. Notably, major search engines, such as Google and Yahoo! also support the protocol, as an alternative of the *sitemaps protocol*. This support allows libraries to be an integral part of the general Web and at the same time use a protocol developed and customized for their own needs.

Beyond OAI, there are also attempts to modernize the Z39.50 protocol and make it part of the larger family of Web protocols. First, the *Bath profile* specifies the exact query syntax that Z39.50 clients should use, so that clients can interpret the results returned by Bath-compliant Z39.50 servers. A more significant development is the agreement for the *Search/Retrieval via URL (SRU) protocol*. SRU is a standard XML-focused search protocol for Internet search queries that uses *Contextual Query Language (CQL)* for representing queries. The SRU uses the REST protocol and introduces a standard method for querying library databases, by simply submitting URL-based queries. For example, consider the following URL-encoded query: http://z3950.loc.gov:7090/voyager?version=1.1 &operation=searchRetrieve &query=dinosaur&maximumRecords=10

This example is a search for the term "dinosaur," requesting that at most ten records to be returned. The SRU protocol is easy to support and implement, and is familiar to programmers that also use such syntax to interact with other popular Web services.

Key Applications

Digital libraries are increasingly becoming part of everyday life. The book digitization projects undertaken by corporations (e.g., Google, Microsoft) and by many universities will generate enormous digital archives accessible over the Web. Similarly, the high-quality holdings of the existing libraries are becoming increasingly accessible over the Web, allowing users to reach easier authoritative sources of information.

Cross-references

- ► Bioinformatics Data Management
- ► Digital Libraries
- ► Health Informatics Databases
- ► Metadata Management
- ► Multimedia Databases
- ► Multimedia IR
- ► Querying over Data Integration Systems
- ► Scientific Databases
- ► Semantic Web and Ontology
- ► Semi-structured Text Retrieval
- ► Structured and Semi-structured Document Databases
- ► Text Retrieval
- ► Web Search and Crawl
- ► Web Services and Service Oriented Architecture

Recommended Reading

1. Lynch C.A. The Z39.50 information retrieval standard. D-Lib Mag., 3(4), April 1997.
2. Bergman M.K. The deep Web: surfacing hidden value. J. Electron. Pub., 7(1), August 2001.
3. Chakrabarti S., van den Berg M., and Dom B. Focused crawling: a new approach to topic-specific web resource discovery. Comput. Netw., 31(11–16):1623–1640, May 1999.
4. Cope J., Craswell N., and Hawking D. Automated discovery of search interfaces on the web. In Proceedings of the 14th Australasian Database Conference (ADC). 2003, pp. 181–189.
5. Raghavan S. and García-Molina H. Crawling the hidden web. In Proceedings of the 27th International Conference on Very Large Databases (VLDB). 2001, pp. 129–138.
6. Bergholz A. and Chidlovskii B. Using query probing to identify query language features on the web. In Distributed Multimedia Information Retrieval, SIGIR 2003 Workshop on Distributed

Information Retrieval, Revised Selected and Invited Papers (LNCS). 2004, pp. 21–30.

7. Zhang Z., He B., and Chang K.C.-C. Understanding web query interfaces: best-effort parsing with hidden syntax. In Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data (SIGMOD). 2004, pp. 107–118.
8. Gravano L., Ipeirotis P.G., and Sahami M. QProber: a system for automatic classification of hidden-web databases. ACM Trans. Inf. Syst., 21(1):1–41, January 2003.
9. Callan J.P. and Connell M. Query-based sampling of text databases. ACM Trans. Inf. Syst., 19(2):97–30, 2001.
10. Ipeirotis P.G. and Gravano L. Distributed search over the hidden web: hierarchical database sampling and selection. In Proceedings of the 28th International Conference on Very Large Databases (VLDB). 2002, pp. 394–405.
11. Si L. and Callano J. Modeling search engine effectiveness for federated search. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR. 2005, pp. 83–90.
12. Hawking D. and Thomas P. Server selection methods in hybrid portal search. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR. 2005, pp. 75–82.
13. Ipeirotis P.G. and Gravano L. When one sample is not enough: improving text database selection using shrinkage. In Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data (SIGMOD). 2004, pp. 767–778.
14. Si L. and Callan J. A semisupervised learning method to merge search engine results. ACM Trans. Inf. Syst., 21(4):457–491, 2003.
15. Ntoulas A., Zerfos P., and Cho J. Downloading textual hidden web content by keyword queries. In Proceedings of the Fifth ACM+IEEE Joint Conference on Digital Libraries (JCDL). 2005.
16. Ipeirotis P.G., Ntoulas A., Cho J., and Gravano L. Modeling and managing content changes in text databases. In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE). 2005, pp. 606–617.