

WebBiblio Subject Gateway System: An Open Source Solution for Internet Resources Management

Jack Eapen C.¹

1. Introduction

With the advent of the Internet, the rate of information explosion increased rapidly. Today the main problem confronted by the information professionals and information seekers is not the lack of information resources, but the abundance of them. The quantum of information available on the Internet is bewildering. A study conducted by University of California's School of Information Management and Systems estimated that in 2000, the volume of information on the public Web was 20 to 50 terabytes and in 2003, it was 167 terabytes - at least triple the amount of information.¹ Any normal search on popular search engines generates thousands of results. This causes a great challenge before the new age information professionals and users: filter out relevant information from this vast sea of resources. A variety of tools have been developed to tackle the Internet information explosion- search engines, subject directories, subject gateways and the like. Let's evaluate quickly, the usability of each of these tools.

2. Search Engines

Search engines are the first resort of normal users to find information on the Internet. Search engines use special programs known as crawlers or spiders to index the existing web pages and rank them using different algorithms. The quality of results generated by a search engine can't be assured to a great extent. It depends on the nature of algorithm used. If a search engine is considering only the meta tags of web pages, web designers can fool the crawler by including unnecessary keywords in the meta tags that are not relevant to the content of the page.

We are more than happy when a search retrieves thousands of results. But the worst part is that it's only a tip of the iceberg. A large portion of the Internet is still out of reach of the search engine spiders, the reason being now-a-days static web pages are giving their way to dynamic web pages. (A dynamic page is the one, which really doesn't exist as the user sees it, but created on-the-fly when a user requires for it). Typical crawlers are stumped by such pages and fail to index them. Most of the data in online databases and gateways are thus hidden to the search engines. This portion of the Internet is known as the Invisible Web. Javed Mostafa² (2005) points out that the size of invisible web is 500 times the normal web.

3. Subject Directories

These are hierarchical listings of web sites. Users can browse the subject hierarchy and find the sites of their interest. Some directories offer search feature also. These searches are performed against the catalog record and not against the web pages as in search engines. Subject directories index very few sites as compared to search engines. If a search feature is absent, the user has to browse the entire depth of subject hierarchy to find relevant results. This may be a tedious task. Also sometimes

¹ The author is affiliated to the USAID-EHP Urban Health Program as a Consultant: Library & Website Management. Email: jack@jackeapen.org

a user may find it difficult to determine to which category his/her search topic belongs. There is no way to evaluate the usefulness of the result set also.

4. Subject Gateways

Subject gateways- also known as information gateways- are developed to solve some of the problems associated with search engines and subject directories. Subject gateways are catalogues of Internet resources, which have been selected, evaluated and classified by subject specialists in accordance with criteria designed to ensure resource quality. They provide links to high quality subject- specific online resources. Most information gateways contain summaries of the contents of resources and provide an index. According to Bradley³ (1999), main characteristics of subject gateways are:

- they use the expertise of information professionals and subject experts in collecting and organizing web information resources
- information is checked for authority
- emphasis is on the content of a source rather than its location and
- the information is current and sometimes value-added

Social Science Information Gateway available at <http://www.sosig.ac.uk> is a good example of a subject gateway.

5. Tools for setting up subject gateways

Many organizations and individuals tend to develop subject gateways. Many times lack of an intuitive tool prevents the sustained development of such projects. Once created, they remain static, as the continuous updating may be not so easy. A very few tools are available today for creating subject gateways. Availability of specialized tools will enable information professionals to create high quality subject gateways, which will result in enhanced management of the Internet resources.

6. WebBiblio Subject Gateway System

WebBiblio is an open source software developed by this author. The code is not written entirely from the scratch, but a modification of existing code of the open source library automation package OpenBiblio (<http://obiblio.sourceforge.net>). This is done true to the spirit of open source philosophy. WebBiblio helps to create subject specific gateways of Internet resources quite easily. Records can be arranged into different material types and collections. WebBiblio is currently hosted by SourceForge.Net, world's largest open source development platform. It can be accessed at <http://webbiblio.sourceforge.net>. Current version available is 1.0.

In an organization, each staff member uses a lot of websites for their information needs. Other colleagues may not know all these sites. If organizations can collect a list of websites usually accessed by their staff create subject gateways, this will result in effective knowledge management and thereby enhanced productivity of the organization. New staff members also can perform better by consulting such subject gateways for their information needs.

7. Technical Specifications

WebBiblio is written in PHP, the powerful open source programming language. It uses a MySQL database at the backend for storing data. WebBiblio is a web-based

application- that is, users can access it through a web browser, while the application runs on a server. Apache web server is recommended though it should work with any web server, which can run PHP applications. There is no requirement at the user-end except a standard browser.

7.1 Software Requirements

Web Server- Since WebBiblio is a web-based application, we need a web server to serve the pages to the user. This web server should be capable of interpreting PHP pages as WebBiblio is written in PHP programming language. All the popular web servers such as Apache, Microsoft IIS, Microsoft PWS are capable of doing this. Apache is recommended due to its robustness, security and free and open source availability. WebBiblio is tested with Apache 1.3.* and 2.0.* . Apache can be downloaded from <http://httpd.apache.org/download.cgi>.

PHP- This open source programming language is becoming popular and maturing at a very fast rate. PHP is a server side scripting language, which means all the processing of the code is done at the server side and only the output is thrown to the user side. WebBiblio is tested with PHP 4.3.* and 5.0.1. PHP can be downloaded from <http://www.php.net/downloads.php>

Database- All data the user wants to store in a WebBiblio system is kept in a database. WebBiblio uses the powerful open source database MySQL as its backend. WebBiblio works well with MySQL versions 3.23.* and 4.*.*. It hasn't been tested with MySQL 5 series. MySQL is available for download at <http://dev.mysql.com/downloads/index.html>

Operating System- PHP applications are platform-independent. So WebBiblio should work on any operating system which is capable of running MySQL and PHP. GNU Linux versions comes bundled with Apache, MySQL and PHP. On Windows, users can download and install them as they are freely available at the above mentioned web sites.

Table 1 shows the various environments on which WebBiblio has been tested.

Operating Systems	RedHat Linux	Mandrake Linux	Windows 98	Windows XP
PHP Versions	4.3.*	4.3.*	4.3.3 5.0.*	4.3.10 5.0.*
MySQL Versions	3.23.58	4.0.*	3.23.58 4.1.9	3.23.58 4.1.9
Web Servers	Apache	Apache	Apache	Apache

Table 1

7.2 Hardware Requirements

The hardware requirements for WebBiblio arise directly from the requirements to run Apache, PHP and MySQL. These programs run on any Intel-based architecture so as WebBiblio. It has been tested on PCs varying from the ones with Pentium processor with 48mb RAM to Pentium IV with 256mb RAM.

8. Technical Architecture

Fig. 1 shows the technical architecture of WebBiblio.

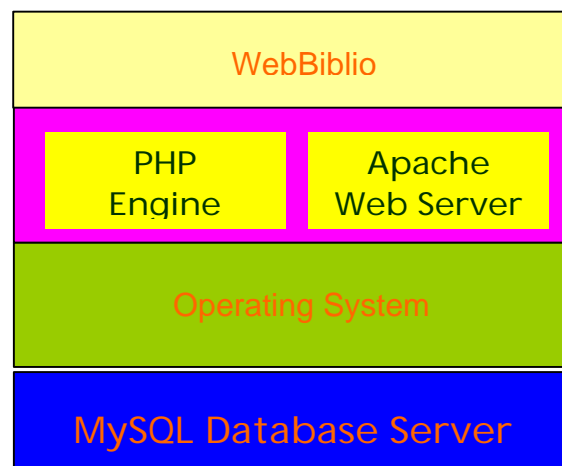


Fig. 1

WebBiblio has three components or modules-Administration, Cataloging and Online Public Access Catalog (OPAC). Fig. 2 shows the modular diagram of WebBiblio. Access to first two modules is restricted by password. WebBiblio administrator can create users and assign them access to either of the two protected modules or to both of them. Administrator can define the material types and collections to which the records can be classified.

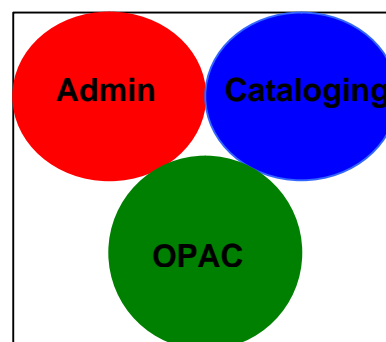
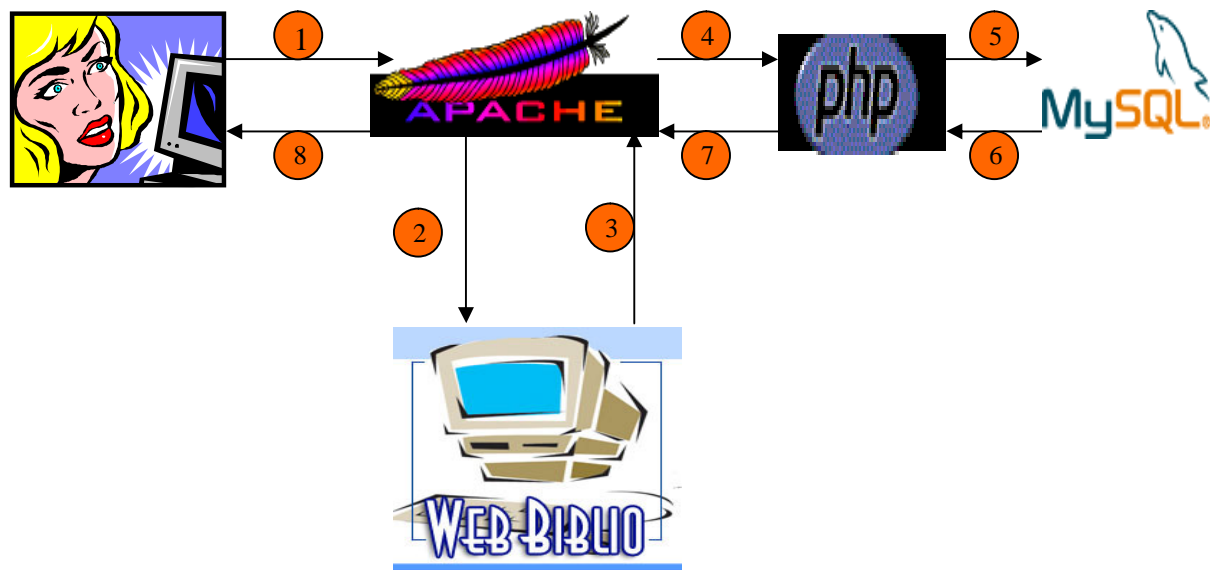


Fig. 2

When a user invokes the WebBiblio interface through a browser, web server sees that the requested page is of type php and transfers the request to the PHP engine. PHP engine reads the code written in the requested file and executes it. If there are any sql queries included in the page, those are sent to the MySQL server and results

obtained. Then PHP engine creates a standard HTML output and sends it back to the web server, which in turn throws it back to the browser. Fig. 3 is a diagrammatic representation of actions taking place when a user interacts with WebBiblio.



Description of steps

1. User accesses WebBiblio through a browser. Browser request is send to Apache (web server)
2. Apache requests the required page from WebBiblio
3. WebBiblio returns the page to Apache
4. Apache sees that the page is a php script, so transfers it to the PHP engine
5. PHP engine processes the code, sends database queries to MySQL server
6. MySQL executes the queries and returns results to PHP engine
7. PHP generates a HTML output and sends it to Apache
8. Apache throws the HTML page back to the user

8. Salient Features of WebBiblio

8.1 MARC Compatible Data Structure

WebBiblio uses USMARC standard for its records. This helps in detailed cataloging of the record, if desired and to exchange data. Presently WebBiblio can import MARC data into it, though MARC export facility is not currently available. The MARC tag 856u is used for the URL field, which appears as a hyperlink in search records.

8.2 Various search options

WebBiblio provides title, author and summary field searches. Search operator is Boolean AND. Search can be limited to particular material types and collections. Search results show how many times a resource link has been clicked so far which helps users in determining the popularity of the resource.

8.3 Supports Collaborative Development

WebBiblio supports the participation of users in developing the subject gateways created by it. Users can submit any useful sites to the system through a simple form

available in the OPAC. That will be automatically e-mailed to the designated administrator and he/she can decide on the inclusion of that resource according to the collection policy.

8.4 Well-documented

WebBiblio provides detailed installation instructions and a web-based installation procedure. It provides online help to system administrators, catalogers and normal users. A demo site is available at <http://webbiblio.sourceforge.net/WebBiblio/> where interested users can experiment with WebBiblio and verify the features.

8.5 Compact

The entire software occupies only 1mb of disk space out of which files of 200kb are sql files used for installation, which can be deleted after the installation. Size of the MySQL database is in addition to this, which depends on the amount of data stored. Maximum size of WebBiblio databases is limited only by the allowed size of the MySQL database, which in turn related to the operating system constraints and not directly related to the MySQL limitations.

8.6 Free and Open Source

WebBiblio is freely available to anyone along with the source code. This ensures the quality of code, as it is open to scrutiny. Interested developers can improve the code and add more functionality into it. All the required software needed for WebBiblio are also open source.

9. Why WebBiblio, why not a digital library software?

WebBiblio is not a digital library software. It just manages the information about information resources and points the user to where the resource is available. Digital Library softwares like Greenstone or DSpace actually store resources within it and manage it. They are more complex in nature. WebBiblio offers a very easy way to create and manage a catalog of Internet resources. As open access publishing is booming up these days, a lot of quality resources are freely available over the Internet. So a library need not store local digital copies of such documents, instead use subject gateways to guide users to the relevant locations, provided good internet connections are available. This may save a lot of storage space.

10. Sample subject gateways using WebBiblio

10.1 Relativity Resource Library

2005 marks the centenary of Einstein's miraculous year in which he published 3 papers in Physics, which completely changed our understanding of the universe and its phenomena. So 2005 is celebrated as the World Year of Physics. To celebrate the occasion, this author is setting up a sample subject gateway called Relativity Resource Library at <http://www.physics.jackeapen.org> using WebBiblio. This is an effort to collect information regarding quality web sites related to theory of Relativity and its inventor Einstein so that students and researchers seeking information on this topic can easily find them.

10.2 Urban Health Library

Environmental Health Project India (EHP India) is the USAID supported urban health program in India (where this author is affiliated to). One of the major components of EHP India's activities is to generate and disseminate information related to urban child health in India. EHP India is presently developing a compendium titled 'Urban

Health Library' (UHL) which is a compilation of 200 articles and reports related to urban health. UHL will be delivered in CDs and also through Internet. The Internet component of UHL is being created using WebBiblio. As most of the items included in the UHL are available on the Internet, storing them on EHP India's web space is wastage of disk space. Moreover this warrants copyright permissions from the respective publishers. Using WebBiblio we are preparing a catalog of the UHL resources and providing links to the publisher's website where the resource is available. Users are benefited from getting information about quality resources on urban health and search and browse mechanisms to access them. The organization is benefited from the saved disk space and hassles of obtaining copyright permissions. Resource publishers are benefited from greater visibility to their resources and the increased traffic on their websites.

11. Conclusion

The new-age information professional must be equipped with tools necessary to manage information resources on the Internet. Open source softwares offer a wide spectrum of such tools that enables them and their organizations to achieve this. WebBiblio is such a tool, which provides cost-effective and technically feasible solution for creating high quality subject gateways.

¹ Lyman, Peter and Hal R. Varian (2003). How Much Information. Available at <http://www.sims.berkeley.edu/how-much-info-2003> [visited 30-05-2005]

² Javed Mostafa (2005). Seeking Better Web Searches, *Scientific American*. Available at <http://www.sciam.com/article.cfm?chanID=sa006&colID=1&articleID=0006304A-37F4-11E8-B7F483414B7F0000> [visited 05-05-2005]

³ Bradley, P (1999). The Advanced Internet Searcher's Handbook. London: Library Association Publishing