APPENDIX 18. THE WORLD WIDE LIBRARY

Written by Christopher Burns, Member, NCLIS Group of Experts for This Assessment

WHY CAN'T THE WEB BE MORE LIKE A LIBRARY?[10]

Why can't the Web be more like a library? If you have a library card you can see anything in the collection, regardless of the publisher, format or age of the document. But on the Web you often have to go from publisher to publisher, agency to agency, site to site to find the document you need. And at each step you have to identify yourself, present credentials, and request access.

In a library, everything available to you is in the card catalog. You can search on author or subject, and choose the material you want by looking at the standard information on the card. But on the Web there is no comparable set of metadata, no good way to look up an object. You can search across the web looking for a description of the object, but the description is not fielded, it doesn't define the date or structure of the document, and because it is a broad, general search, it brings back many more candidates than the user can handle. You can't narrow your search to a "computer sciences" library, or a "personal health" library or a K-12 education library, or a "government documents" library. The state of the art is to search on the words appearing on the HTML page, if it is accessible for anonymous searching.

If the information you need is not in the library you can locate it in an affiliated library by searching the interlibrary network catalog. But on the Web site the search tools are unique to that site; you can't search more than one site at a time at any level of precision. The thousands of individual document collections now available on the Web have no standard catalogs that can be searched together, and they share no common search protocol.

If the book or document you want is in the library, it is really there—or will be returned within a predictable time. But on the Web it is common to find that the information has moved to an unknown address, or that it has been superseded by a new and different document, or that it has since been withdrawn, lost, or "revised". The Web is ephemeral; there is little sense of preservation or accountability especially in critical areas like scientific, technical or government documents. As often as we may have criticized librarians for emphasizing preservation and circulation management, we can see now that life without those disciplines is chaotic and unreliable.

If you can't find it in a library, ask the librarian. But on the Web there are no authority files or cross-indexes. There are no tools like a list of publications recently added to a community of sites, no standard dictionary of author's names and pseudonyms, no catalog of sites. We get "links lists." No one evaluates the authenticity or usefulness of a site—each one asserts its authority through mere existence. Time and again we have seen that the most valuable information retrieval device is a helpful colleague whose knowledge and judgment one trusts—and there are none of those on the Web.

In spite of the rich profusion of knowledge now available to anyone on the simplest terms, the Web is more like a flea market than a department store. You walk from stall to stall, adjusting to this

[10] Available at http://www.nclis.gov/govt/assess/asess.appen18.pdf. This appendix was last revised on September 5, 2000.

organizational scheme and that eccentric standard. But the pleasure it offers in serendipity is lost for some in its lack of organization and precision. The searcher sinks deeper and deeper into that most modern paranoia: knowledge that the exact information needed is out there somewhere, coupled with the certainty that it will never be found.

We can change this. If the Web or some portion of it is to become the organized network of knowledge that our libraries now represent, then we will have to find a way to preserve the flourishing independence and accessibility of Web sites as we know them, but align them through standards, protocols and procedures so document and publication catalogs can be searched more precisely together. How can we bring the thousands of emerging document sites to a higher state of organization?

MOVING TOWARD DECENTRALIZED COLLECTIONS

Historically we have approached the task of managing diverse document collections by putting all the materials into a single database, running under a single search protocol and a single access management regime. This is still the obvious solution when the documents are all of the same type and format and when all the users belong to the same organization. When the documents types are the same and the users are willing to accept strong central systems management, it is even possible to have distributed databases operating in separate but identical environments. This works for hospitals owned by a single group, for example, who keep their patient records in a group-specified format on separate but interconnected systems. Branch libraries in a large city, a chain of retail stores or regional offices of a major government agency may also benefit from centralized management of decentralized resources.

But when the interests and activities of the user group begin to diversify, and the collection of documents comes to include many different formats the right architecture is less clear. Different file types may be best stored in different systems. User groups may have different requirements for access and security. Separate systems are able to evolve more quickly than a large document database on which many groups depend. More important but harder to rationalize, decentralized collections keep the information under control of the organization that cares most about it; the collection is nourished. Over and over we find that when diverse collections are gathered together in a central corporate or government system, maintenance declines and controls grow to favor efficient management of the system instead of service for the user.

Recent efforts to manage diverse document collections have taken a different approach: leave the collections on local systems under local control, but create a single shared catalog located on the Web. Users can search the catalog for documents and other information, just as one searches a library catalog. When the document is found, the user clicks on the link icon and is connected over the Internet to the local system where the information is available. Access and security can be managed specifically by each collection, and each system can be configured independently for the needs of the user and the type of files it contains. This architecture allows rapid growth in the number and diversity of collections available to the user. The tools are simple and local control over content is preserved. To make a document available beyond the local system, the collection manager puts a "card" in the central catalog, but the object itself remains at home.

But that is the problem. For a shared catalog system to be useful, the independent collection managers must keep it up to date. While some automation is possible to help send updates to the central catalog, the burden inevitably falls on the local collection management staff, and they usually have more

immediate duties. We have tried having the corporation send subsidies to the operating unit to pay for this cataloging, but that is a pale monetary incentive. Various organizations have tried creating document librarians, centrally funded, who scour the local collections and do the cataloging. But this, too, gets displaced in the daily business of creating and using the information. If the cataloging is not somehow made integral to the production workflow, it doesn't get done.

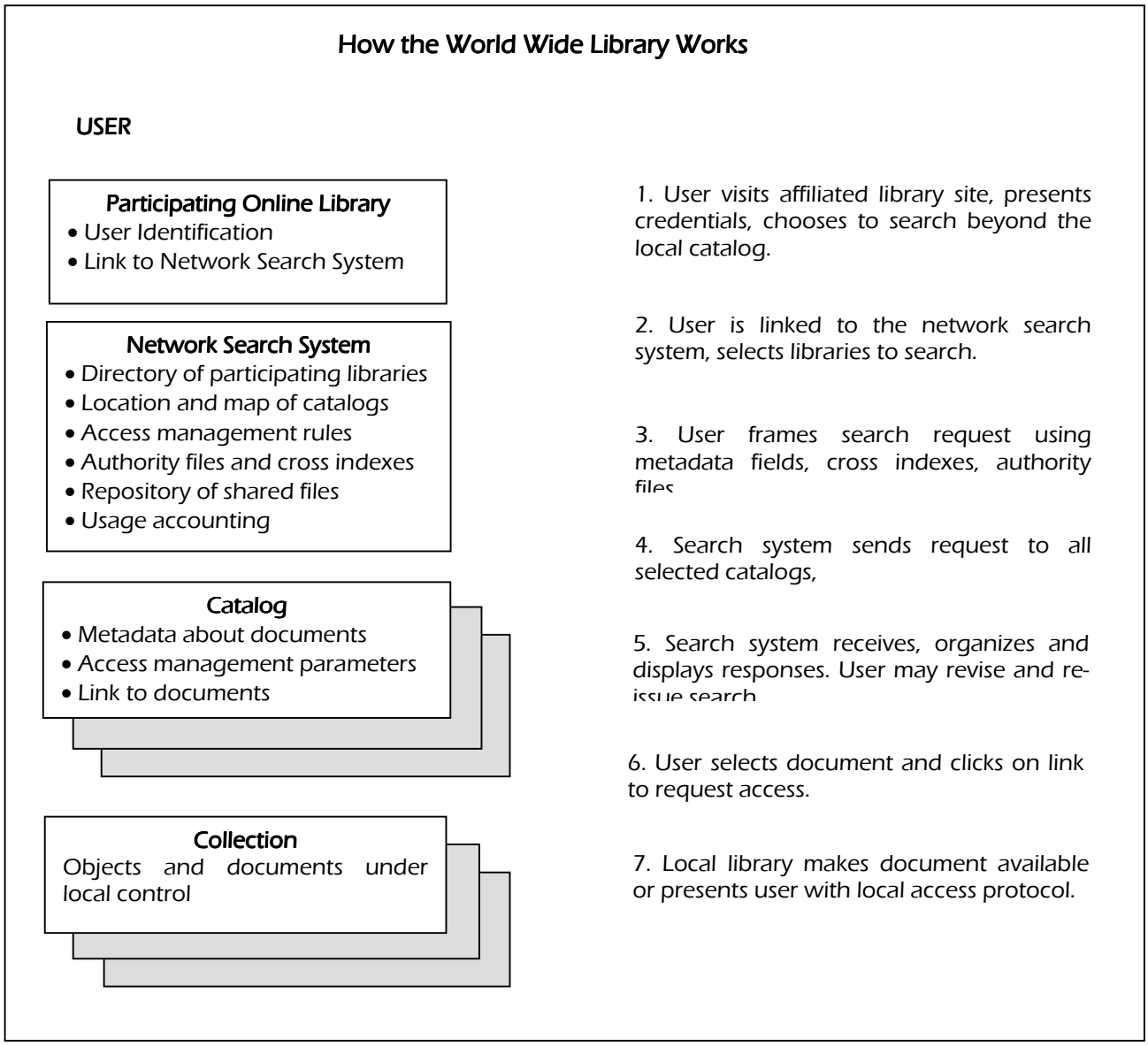
RECIPROCAL CATALOG ACCESS

The third alternative, proposed here, is to decentralize the collections, decentralize the catalogs and create a centralized search mechanism by which the users of each collection can search other participating collections at the same time. This is not a general utility available to all users of the World Wide Web. It is a "network" or affiliated group of online collections and libraries that have agreed to provide each other's users with reciprocal catalog access. The network is implemented by (a) adopting certain information standards and protocols specific to their group, and (b) creating a shared search mechanism which all their users can access. A network might be shared by all the colleges in a statewide university system, or all government agencies, or all the public libraries and museums in a region. It may be a research consortium in biological sciences, or independent suppliers of parts to the aviation industry. By creating a standard "public" catalog format, a standard search request, and a protocol for reciprocal access, the group can maintain independent document collections, each cataloged by its own staff, but each accessible to the users of all other participating collections. It will behave as a federation of collections, accessible as a virtual database, functioning like a network of libraries.

Three major systems elements are necessary for such a network to function: (a) each participating library would have to create and maintain a public catalog that follows a standard format. (b) Together the participating libraries would have to install and support a central system for searching all the catalogs in the network. (c) The members of the network must agree on a protocol for extending user privileges, allowing access, and reporting usage.

The Public Catalog: Central to the concept of a network library is that each participating library or collection maintains an online catalog of the documents it has chosen to make available. Where a simple HTML page might list all the documents available, the catalog provides fielded metadata so the user can search more deeply and more specifically on date, author, document number, format or language. A participant could choose to make some but not all objects accessible this way, and could keep an internal catalog in an entirely different format for users within the local organization.

The catalog must be in a structure and location accessible over the Internet and should contain basic metadata in a standard format for each object or document. While there are several existing metadata standards that might be employed (for example the ONIX system develop by book publishers or GILS developed primarily for government use) the Dublin Core standard seems a good starting point for the metadata, and offers an insight into the catalog's likely complexity. Developed by the library community, the standard specifies the definition and general format of sixteen descriptive elements designed to aid in finding the document. Those elements include:



Title: Title of the document or resource

Subject: Series of key terms that describe the document. These may be from a general thesaurus, or from a specialized thesaurus maintained by the affiliated libraries. For example, a community of document collections related to aerospace might agree to share a specialized thesaurus of terms useful to that community.

Description: A description, abstract, table of contents or excerpt of the document.

Coverage: An optional additional description, usually to deal with geography.

Creator: The person or organization responsible for creating the document.

Contributor: Additional persons or organization who contributed to the content.

Publisher: The organization responsible for primary distribution.

Source: The original source of the material, if not the author or publisher.

Relation: Relationship of this document to a previous document or set of documents.

Type: The Dublin Core proposal recommends using very general types, such as dataset, sound, image and text.

Format: File format, medium, dimensions.
Rights: A rights management statement, for example copyright date and owner.
Date: Date of publication.
Language: Language in which the document is written.
Identifier: May be a unique number within a known set such as a government publication number, or it may be a more general identifier such as the Digital Object Identifier used in the publishing community.

The catalog would need two additional elements:

Location: A link to the document itself, or a description of its location.
Access: An indication of what limitation there may be on access.

These elements would need to be refined by the participants in a particular network, including development of syntax and conventions for each one (last name first?, format of the date?) so that they could be efficiently searched. Specific differences between one catalog and another in the same network can be mediated by the search engine. The online catalog is like the union catalog compiled by several libraries, or like the online catalog they now maintain together.

The Network Search System: Users of the network would have to send their search requests to all participating sites in a common format. This can be best done by an intermediate site which establishes the identity of the user, helps the user construct the search in the most efficient form, sends the search request out to participating catalogs and provides the user with an aggregated response. The search system provides a directory of network sites, information about how those catalogs are maintained and authority files to help the user clarify or expand certain kinds of searches. It permits the user to search all the catalogs in the network with a single command and enforces any access restrictions in place at individual libraries. It redirects any searches or links to new URL's in case they have been moved and reports the results of the search to the user. The user may then access the document directly from the participating collection, or request access from the collection manager. It begins to behave as a librarian's assistant, providing news, advice and help in searching across the network.

In an expanded role, the network search system might support a shared repository of documents or objects that all the libraries use. It may also store documents that are no longer in the individual library's collection but which the entire network agrees should remain available.

The Access Management Protocol: The same network search system would also determine that the user is member of the authorized user group. Individual libraries and collection managers who participate in the network may issue their members user identities, or they may connect the user to the central search system through their own site. But basic to the notion of an online library network is that the participants are not individuals but libraries or online document collections who manage the network together, and who affirm the identity of the individual users to whom these access privileges have been extended. The access management system would also record the usage of documents by user or participating library. It is the equivalent of libraries in a region that honor each other's library cards.

Libraries who wish to join an existing network may do so by creating a public catalog on their site and agreeing to the reciprocal access management protocol. As networks form around specific types of documents, specific topics, regions or types of organizations, a library may belong to more than one network without creating more than one public catalog. To form such a network, a group of libraries would have to agree to acquire and support the shared search system.

Users who wish to gain access to the network may request affiliation with a participating library, and while this seems to present an obstacle to a user accustomed to ranging freely across the Internet, it allows a community of libraries to share a level of security, user identification and usage accounting that might otherwise be costly for an individual library.

The World Wide Web was conceived as an open system allowing any user access to any information on any site. But the lack of a structured catalog and the absence of an adequate mechanism for searching multiple sites means that while the Web is wonderful for reaching "pages" of information, it isn't equipped to handle the higher form of information objects, "documents". The World Wide Library allows sites of similar interest to create affiliations with common metadata and reciprocal catalog searching so that the user of one site can find documents on other sites as well with a single search.