Guidelines for Collecting Aggregations of Web Resources

William H. Walters, Samuel G. Demas, Linda Stewart, and Jennifer Weintraub

The best aggregations of Web resources combine authoritative information, effective retrieval mechanisms, and dependability of access. The worst can be incomplete, inaccurate, and difficult to use. Libraries can expand their Web holdings and maintain high-quality information services by selecting aggregations based on objective standards of content, coherence, and functionality. These three criteria can also be used to plan the most effective presentation of aggregated resources in the library's catalog or Web pages.

As Internet resources have grown in number and importance, public and private organizations have begun to develop Web-based aggregations of journals, conference reports, data files, and other resources. *JSTOR,* for example, includes back issues of nearly fifty scholarly journals in eleven academic fields.[1] Similarly, *GPO Access* serves as a single point of entry to the reports and publications of over twenty-five government agencies.[2] While aggregations vary in format and content, all meet two criteria:

- They are distributed by a single agency or group (an aggregator) with responsibility for maintaining the collection and selecting the resources included within it.
- They are prepared and distributed as a collection; the payment of subscription fees or access charges (if required) allows access to the entire aggregation.

Some aggregations include only those resources produced by the distributing agency, while others provide access to information generated by other organizations. *IDEAL* is devoted to Academic Press journals, for example, while *JSTOR* includes the output of many different publishers.[3] In all aggregations, however, a single agency, group, or person is responsible for the selection, organization, and delivery of information resources. The responsible agency designs and maintains the Web site, establishes standards for the inclusion or removal of items, and acts as liaison with publishers and subscribers.

An effective aggregation combines materials closely related by function or topic, providing a single point of entry to a set of resources of known quality and scope. The best aggregations also include search mechanisms or contextual material not available within any single resource. For these reasons, the selection and cataloging of high-quality aggregations can improve the accessibility of individual resources. In contrast, ineffective aggregations can impede access to resources that would be accessible if collected separately. These aggregations may provide incomplete or inaccurate versions of the resources they include, for example. This is especially likely when the aggregator provides access to information originally collected or distributed by other agencies. In addition, many aggregations rely on ineffective information retrieval mechanisms. An interface designed for use with a particular database, for instance, may be adopted for all the databases within the aggregation—including those for which it isn't appropriate.

Our experience suggests that certain objective criteria can be used to distinguish between effective and ineffective aggregations. For libraries that normally provide links to entire Web sites or collections, these criteria can be used to identify the most useful sites. For libraries committed to the selection and cataloging of individual Web resources, these standards can be used to identify those aggregations that warrant special consideration—those that are suitable for selection and cataloging *as aggregations.* At the Albert R. Mann Library, we maintain our commitment to the accessibility of individual Web resources. Rather than establishing a single link to the Bureau of Labor Statistics site, for example, we provide separate catalog entries and access points for those BLS Web resources that meet our selection standards—titles such as the *Occupational Outlook Handbook* and the *Consumer Price Indexes.*[4] At the same time, we realize that certain aggregations provide capabilities and advantages not otherwise available—that the best aggregations can increase the accessibility of the resources they include. These guidelines were therefore developed in an attempt to identify and select the most useful Web aggregations in the Library's subject areas.

Standards for the Selection of Aggregations

Three criteria—content, coherence, and functionality—are central to the evaluation of aggregated collections. While designed for use in the selection of Web resources, many of these standards can be applied to telnet, microform, and print aggregations as well. These criteria are meant to supplement, rather than replace, existing standards for the collection of Internet resources.[5]

William Walters and **Jennifer Weintraub** are bibliographers in the Albert R. Mann Library at Cornell University. **Linda Stewart,** also at Mann, is Acting Head of Collection Development and Preservation. **Sam Demas** is College Librarian and director of the Laurence McKinley Gould Library at Carleton College.

Content

Evaluations of content should be based, first of all, on an assessment of the individual resources within the aggregation. Specifically, most of the resources within the aggregation should meet the usual criteria for selection—criteria such as appropriateness of subject, authority of the author, projected use, scholarly value, instructional value, archival value, and general utility. Ideally, the assessment of content should be based on an evaluation of every resource within the aggregation.

The content criterion also requires evaluation of the aggregation as a whole. Five considerations are especially important:

- The proportion of resources in the aggregation that would be added to the collection (or listed in the library's Web pages) if they were available separately
- The existence of essential resources available only through the aggregation
- The proportion of relevant resources available elsewhere (outside the aggregation)
- The cost of the aggregation relative to the cost of purchasing the relevant resources separately, and
- The possibility of replacing current subscriptions or purchases with resources included within the aggregation

If resources contained within the aggregation duplicate those available elsewhere, comparisons of content are especially useful. In particular, the aggregation should fully and accurately represent the content of every resource it includes. More specifically:

- Technical documentation and explanatory notes present in the original resource should be available through the aggregation as well.
- For serial resources, the aggregation should include current issues and updates as well as backfiles; updates or new issues should be available within the aggregation soon after their original release
- The aggregation should include all necessary tables, graphs, and related materials, especially if they present information not available in the text.
- The aggregation should maintain the context of the original source; statistical documents retrieved through an aggregated database, for example, should identify the original source and provide the context necessary for accurate interpretation of the data.

Aggregations are preferred, of course, when they provide authoritative content or context not available in the original source.

Coherence

The coherence criterion specifies that the resources within an aggregation should be related by subject or function. *Subject* can be interpreted in a broad or narrow sense—from a collection of instructional materials for an introductory economics course to a set of journal articles on the microeconomics of industrial location. The concept of coherence by function allows for the collection of useful materials not necessarily dealing with a single subject—aggregations of daily newspapers or job announcements, for example. (Many aggregations intended for reference use may be coherent by function rather than subject.)

In each case, the evaluation of coherence is also an evaluation of utility—"Does this aggregation exhibit the coherence needed to make it a useful resource for library patrons?" An aggregation, regardless of its content, will not be useful if patrons are unable to identify it as a potentially relevant source of information. Coherence is necessary so that patrons can gauge whether a particular aggregation is relevant to their needs. Similarly, the most coherent resources are those that contain a significant proportion of the useful Web materials on a particular subject (or with a particular function). Patrons will be unlikely to make use of an aggregation unless they have a reasonable chance of finding what they want.

Unfortunately, many aggregated collections fail to maintain coherence by subject or function. *IDEAL*, for example, provides access to 175 Academic Press journals covering a wide range of subjects. While the resources within this aggregation are certainly valuable, library patrons can be expected to find them only if they know (1) which journal they need and (2) whether that journal is published by Academic Press. (This is not true, of course, if the library provides a catalog entry for each individual resource within the aggregation.) Similarly, *Stat-USA* contains a large number of publications and data files related by agency (U.S. Department of Commerce) but not necessarily by subject or function.[6]

Functionality

Functionality refers to the organization and features of the aggregation itself—among other things, the effectiveness of its interfaces and search mechanisms. Because most aggregations have been developed in an attempt to facilitate access, the best aggregations have much to offer in this area. They may provide keyword search capabilities not present in the original publications, or combine several data files into a more useful product. The *Climate Visualization System (CLIMVIS)*, for example, allows users to map and graph climate data

from four different sources.[7] While the same data are available elsewhere on the Web, *CLIMVIS* provides added capabilities and convenience.

An effective aggregation should provide complete, reliable, and consistently structured access to the resources it includes. In particular:

- The aggregation should provide an accurate index of the resources it contains, and, if possible, a global search capability—a means of searching several related resources simultaneously; if there are several search engines, each should specify the resources covered.
- The hierarchy of files should be logical and consistent.
- Each resource should be identified by a single name that is used consistently in links and labels.
- The aggregation should provide links to specialized helper applications required for use with particular resources—image viewers, file compression utilities, etc.
- The links to each resource should be reliable; users should be able to connect to each resource even during high-use hours.

The best aggregations provide functions not otherwise available, such as links from bibliographic databases to full-text articles. They also maintain an interface format that conforms to the expectations of users. While the features of each interface are dictated largely by the content of the aggregation, certain common elements can help put users at ease. Although the interface for an archive of genetic sequence data will be fundamentally different from the interface for a general-purpose bibliographic file, all bibliographic files should maintain certain common elements—a standard symbol for the truncation of search terms, for example, and a uniform set of searchable fields. The Z39.50 interface, among others, may be a useful step toward this goal.

Ensuring Access to Aggregated Resources

The standards of coherence and functionality are designed to facilitate access to aggregated resources. Even the best and most coherent aggregations require careful presentation in the library's catalog or Web pages, however. While we encourage the cataloging of Web resources along with print titles in the online catalog,[8] other access mechanisms may be appropriate in certain cases. (Some libraries use subject-based Web pages with links, for example, while others maintain a separate catalog of Web resources.) Regardless of its form, an effective access mechanism should meet three criteria:

- Searches for information on a particular topic should retrieve the records of those aggregations containing relevant resources. This type of search should not normally require the assistance of a reference librarian.
- Well-known resources should be accessible by title even if they are part of an aggregated collection.
- The scope and function of each aggregation should be represented in a way that allows patrons to evaluate the relevance of its contents. The bibliographic record should include enough information for patrons to judge the aggregation.

Patrons searching for a particular resource by title, for example, should first be guided to the aggregation containing the relevant resource, then presented with enough information to decide whether the aggregation satisfies their needs.

An aggregation that meets the standards of content, coherence, and functionality is suitable for cataloging as a single title—a single database record. Effective access by subject and title can be achieved through the assignment of all relevant subject headings, the provision of added title entries for major resources within the aggregation, and the inclusion of notes describing the aggregation's scope, content, and function. Both subject and title access are important, since neither is an adequate substitute for the other. The thoughtful cataloging of aggregated resources may also help patrons understand the range of information available within a particular subject area. A patron searching for the *HealthSTAR* database, for example, may be directed by the library catalog to *Internet Grateful Med*, an aggregation of bibliographic resources in the biomedical sciences.[9] Within *Grateful Med*, the patron will encounter *HealthSTAR* along with several other databases that may be even more useful—*MEDLINE* and *PREMEDLINE*, for example. In this manner, the best aggregations support a method of resource discovery not otherwise available on the Web; they allow users to browse a set of books (resources) that meet specific quality standards within a specified call number (subject) range.

An aggregation that fails to meet the standard of coherence or functionality is not suitable for cataloging as a single entity. A collection of unrelated resources, for example, cannot be described adequately in a single bibliographic record. Likewise, an aggregation with an unsatisfactory interface will not facilitate information access even if the catalog record is complete. High-quality resources within an aggregation may be cataloged separately, however, even when the aggregation as a whole is deficient. This practice—individual cataloging

of resources acquired through an aggregation—is most appropriate when the aggregation meets content standards but fails to satisfy the criteria of coherence or functionality.

Conclusion

While many aggregations are free, others can be expensive. Moreover, indirect costs—the costs of selection, description, and support—will often exceed the price of the aggregation itself. In many cases, however, the investment required to collect and maintain aggregated resources is justified by a corresponding set of benefits.

Aggregations can be especially valuable to libraries just beginning to identify and catalog Web materials. In particular, the systematic selection of aggregations can improve knowledge of the resources available on the Web and encourage the utilization of free resources that might have otherwise gone unnoticed. The selection and cataloging of aggregations can also promote efficiency in library operations. An aggregation of one hundred resources, for example, can be acquired and maintained more easily than one hundred separate titles. Moreover, the best aggregations are already organized for effective information retrieval, and many provide access to specialized applications, interfaces, or services that would not be available otherwise. Because many of the agencies responsible for selecting and organizing aggregated resources have considerable subject expertise, aggregations in a particular subject field are often tailored to the needs of users in that discipline. Most collections of genetic sequence data, for example, include interfaces particularly well-suited to the needs of biologists. Libraries, operating alone, cannot maintain this level of expertise in subject-specific information delivery.

As Web aggregations grow in number, size, complexity, and specialization, several important problems will need to be addressed. Most importantly, neither aggregators nor librarians have accepted responsibility for the preservation of aggregated resources. This problem is of special importance for those aggregations that provide only current information—the latest issue of a periodical, for example. In some cases, information is irretrievably lost when replaced by a new or updated file. A second problem is that many aggregations are marketed only to individual subscribers. Some sponsoring agencies have yet to establish subscription and pricing policies for libraries. More generally, the relationship between libraries and aggregators has not been clearly defined; there is no typical vendor/client relationship in this arena. While agreements between individual libraries and aggregators may address some of these problems, long-term solutions are likely to require the adoption of coherent preservation and pricing standards by a large number of cooperating institutions.

Acknowledgements

We are grateful for the advice and comments of Gregory W. Lawrence and Janet A. McCue.

References and Notes

1. *JSTOR*, URL http://www.jstor.org/.
2. *GPO Access*, URL http://www.access.gpo.gov.
3. *IDEAL*, URL http://www.idealibrary.com/.
4. Bureau of Labor Statistics, URL http://www.bls.gov; *Occupational Outlook Handbook*, URL http://www.bls.gov/ocohome.htm; *Consumer Price Indexes*, URL http://stats.bls.gov/cpihome.htm.
5. Rachel Cassel, "Selection Criteria for Internet Resources," *College & Research Libraries News* 56, no. 2 (Feb. 1995): 92–93; Samuel Demas, Peter McDonald, and Gregory Lawrence, "The Internet and Collection Development: Mainstreaming Selection of Internet Resources," *Library Resources & Technical Services* 39, no. 3 (July 1995): 275–90; Dan C. Hazen, "Collection Development Policies in the Information Age," *College & Research Libraries* 56, no. 1 (Jan. 1995): 29–31; Julia Ann Kelly, "Collecting and Accessing 'Free' Internet Resources," *Journal of Library Administration* 22, no. 4 (1996): 99–110; Gregory F. Pratt, Patrick Flannery, and Cassandra L. D. Perkins, "Guidelines for Internet Resource Selection," *College & Research Libraries News* 57, no. 3 (Mar. 1996): 134–35.
6. *Stat-USA*, URL http://www.stat-usa.gov/.
7. *CLIMVIS*, URL http://www.ncdc.noaa.gov/onlineprod/drought/xmgr.html.
8. At Mann Library, Internet resources and aggregations that meet our selection criteria are added to the Mann Library Gateway, a Web-based catalog that provides bibliographic information as well as direct access to networked resources. (The Mann Library Gateway, URL http://www.mannlib.cornell.edu/, has recently been incorporated into the Cornell University Library Gateway, URL http://campusgw.library.cornell.edu.) Networked materials can also be found along with print materials in the Library's NOTIS catalog.
9. *Internet Grateful Med*, URL http://igm.nlm.nih.gov/.