Introduction

ANDREW G. TOROK

THE THEME OF "ORGANIZING THE INTERNET" brings to mind the late 1950s folk-rock singer Jimmie Rodgers's song titled "The World I Used to Know." A great many developments have transpired in the world of information science since the seminal works of S. C. Bradford, Claude Shannon, Vannevar Bush, and numerous other pioneers. To those of us who have been in the information science field for several decades, the peek-a-boo devices such as Termatrex, Mortimer Taube's Uniterm cards, and discussion of pre- and postcoordinate indexing have given way to the world of browsers, HTML, XML, and numerous other ways of coding text and multimedia. The Internet and the World Wide Web have had a profound impact on how we go about storing and retrieving information. Document integrity has become transient, with little assurance that the location, existence, or even the content of a publication will be the same tomorrow as even a few minutes ago. We are often hard-pressed to determine if the failure to retrieve a publication is one associated with network infrastructure or the publisher. The dream of universal bibliographic control seems quite remote. By being able to bypass traditional publication channels, anyone can publish virtually at will. The situation becomes more chaotic when we consider the increasing redundancy of knowledge and the rampant proliferation of misinformation and disinformation, to say nothing of social concerns with pornography, copyright violations, and other flagrant obtrusions into personal rights. Nevertheless, it behooves the information worker and the information user to make some sense of order if good information is to remain the basis of learning and decision making, and if documents are to continue as an archive of human knowledge.

Andrew G. Torok, Professor, Department of Educational Technology, Research, and Assessment, Northern Illinois University, DeKalb, IL 60115
LIBRARY TRENDS, Vol. 52, No. 2, Fall 2003, pp. 203–208

As I reflected on writing this introduction, I began to ask myself just how far have we come from the world I used to know. The biggest paradigm change has not been that of technological development. Rather, the Internet has enabled virtually anyone with access to a computer to become intimately involved with the entire information cycle, namely, publishing, acquiring, organizing, and retrieving information, thereby bypassing information intermediaries such as indexers, reference librarians, and publishers. There is no question that the technology is vastly different from the early days of information retrieval. At the same time, the paperless office never materialized, nor are libraries being phased out as a result of the public's ability to access information directly from the desktop. More importantly, we still do not understand what constitutes information or how people make relevance judgments. Information retrieval (IR) to most searchers consists of character string matching between a query posed to a data source. In some ways, IR has even regressed, since now the trained search intermediary is no longer needed. The Internet consists of a vast unchecked sea and searching is referred to as "surfing." The issue is further complicated by the proliferation of document formats, incompatibility between generations of hardware, and questionable scalability of software. Even in doctoral seminars that I teach, I find the need to explain Boolean logic and patiently teach students how to develop search strategies, formulate queries, and even how to compute the precision of searches. While the Internet has empowered the general public to perform tasks once done by professionals, it has also created a large body of knowledge needing organization. Vocabulary control is extremely limited at best. The average Web searcher has little understanding of the search process much less a fundamental ability to determine the effectiveness or exhaustivity of a search. People rely on a limited set of search tools, especially general search engines such as Google, not realizing that less than 20 percent of all indexable documents are being accessed. Beyond that, there are many electronic text and multimedia publications that are not indexed at all by Web crawler software. This part of the Internet is called by many names, such as the Invisible Web, the Opaque Web, the Hidden Web, the Dark Web, and so on.

In all fairness, the Internet, especially the Web, is still in its infancy. Techniques for publishing, organizing, and accessing content are changing rapidly as a result of new technological developments, the competitive information marketplace, and the growing sophistication of searchers. As always, libraries are instrumental in promoting access to online publications, especially to those that belong to the invisible Web. Librarians are also educating users through the cooperative development known as information literacy. Developed by AECT (the Association for Educational Communications and Technology) and AASL (American Association of School Librarians) electronic information literacy standards are being

taught to children and teachers alike. The ACRL (Association of College and Research Libraries) supports similar standards for higher education. The dynamic nature of the Internet is going to require methods of organization way beyond the relatively static classification schemes that have served libraries for many years. New methods of organization must take into consideration more sophisticated techniques for content description in order to minimize such problems as retrieving pornography or to be able to detect plagiarism and copyright violations. Eventually the exponential growth of the Web will itself subside. The Internet is not free. Market regulations will eventually restrict the free ride enjoyed by Web publishers. Publication patterns will be easier to recognize as publication activity becomes more linear. The end result will be that users will be able to discriminate in terms of specifying what they want or avoiding the retrieval of unwanted items.

In terms of what "organization" means, I took a fairly broad approach. As in many natural systems, information on the Internet is self-organizing. For example, some search engines determine what is important to index or in what order items are viewed from a search based on link counts that point to a site. Other knowledge bases define themselves by document type, such as usenets, or come into existence by their uniqueness—blogs (Web Logs) come to mind. It seems that for many Web users, ease of use and access appear to dictate knowledge sources. At the same time, there are more organized efforts to identify and make Internet sources accessible. These efforts may simply be a subject sampler of links to relevant sites supporting a subject, area, field, or discipline. For example, the invisibleweb.com site provides classified links to Web-based databases that are not indexed by general search engines. Other sources, such as the Internet Public Library (http://www.ipl.org/ or http://www.libraryspot.com/), are portals that offer classified access to information on a much broader basis. The Open Directory project, also referred to as DMOZ, attempts to create a definitive catalog of the Web. The Open Directory is the most widely distributed database of Web content classified by humans. The Open Directory powers the core directory services for the Web's largest and most popular search engines and portals, including Netscape Search, AOL Search, Google, Lycos, HotBot, DirectHit, and hundreds of others.

Ad hoc classification systems are offered by directory search engines such as Yahoo, and other search engines like Google permit users to search by media type or document format, such as newspapers. Efforts are underway to improve basic document description beyond the limitations of HTML. Xtensible Markup Language (XML) and various permutations are but one example. In the library field, the Dublin Core Metadata Initiative (DCMI) is a notable example. Beyond large-scale efforts to identify and organize Internet content, many local efforts structure learning tools that provide quality information filtering of relevant Web information. They go

by names such as WebQuests, scavenger hunts, and Tracer Bullets. Perhaps someday these efforts will fuse into clear-cut methods of organization that lead to the development of information standards by which Web content can be created. At this time, all such projects can be construed as efforts to organize the Internet.

The purpose of this issue of *Library Trends* is to describe some of these efforts. Leading educators, librarians, and researchers have contributed articles that represent an integrated set of ideas but also serve to reflect the diversity embodied in the theme of "Organizing the Internet." The articles consist of general surveys designed to inform as well as in-depth investigations of specific issues and services.

It is appropriate to have the first article by John Carlo Bertot address the contributions and activities of libraries in a networked environment. Ever since ancient times, libraries have acted as organizers and caretakers of recorded knowledge. In addition to creating and maintaining major classification schemes such as Dewey, Library of Congress, and UDC (Universal Decimal Classification), libraries also pioneered the first major foray into electronic information retrieval. The Dialog system at the Lockheed facility in Palo Alto laid the groundwork for online searching and related software utilities that provide unique indexing capabilities for electronic files. Libraries have also contributed to knowledge organization through a variety of OPACs (Online Public Access Catalogs) and other public and technical services innovations. As libraries move away from these traditional systems grounded in service quality and outcomes frameworks, Professor Bertot discusses the challenges information professionals face in the networked environment.

To continue on the track developed by Bertot, the contribution from Adrienne Franco focuses on finding quality information on the Internet. She makes the point that librarians have long sought to select, organize, and evaluate information on the Internet. Her discussion includes the initial production of "webliographies" by librarians and then focuses on librarian-produced portals and portals with a high level of librarian participation.

Jerry D. Campbell examines portals from a more theoretical perspective. He discusses the Scholar's Portal project that builds on the need for a research library portal. Essentially, a scholar's portal (SP) describes efforts to create specialized subject portals for researchers, until such time as the Web becomes a digital library with seamless access to scholarly information. He builds on an earlier article by outlining the larger context within which SP falls.

As mentioned earlier, document organization is often by media type or even by domain name. A particularly good example of this is government information. Greg R. Notess provides a history of the government on the Web. He makes the point that the government is not only a major con-

tent provider on the Internet but also a source for the organization of the content. Patricia Diamond Fletcher continues the discussion of the government's involvement in organizing the Internet by providing a firsthand analysis of FirstGov.com based on a recent National Science Foundation–funded research project. FirstGov is the portal to U.S. government information and services. Her case study analyzes the reasons leading to the success of the portal.

Quite often the value of portals is to expose users to sources that they might not normally encounter in using general search engines. Even the best search engines index less than 20 percent of what is termed the indexable or "visible" Web. Many persons, even professional researchers, are not familiar with the invisible Web. Any discussion of organizing the Internet needs to address the invisible Web. The invisible Web consists of major databases and document formats that are not indexed by most general search engines. Less familiar, even to experienced searchers, are terms such as the "opaque Web" and the "Private Web." Chris Sherman and Gary Price discuss various permutations of the invisible Web. Their article should be of interest especially to end-users of the Web.

Classification of Web-based information is often determined by popularity, thus user preferences often prompt new methods of organization and access. Amanda Spink provides an overview of recent research exploring what we know about how people search the Web. Her paper reports selected findings from studies conducted from 1997 to 2002 using large-scale Web user data provided by Excite, AskJeeves, and AlltheWeb. The results of the research will have an impact on subsequent methods of organizing the Web according to use.

Any discussion of publication activity or use cannot avoid the topic of copyright. More than ever before, Web publishers are blatantly ignoring intellectual property rights, especially with respect to multimedia. This leads one to ask if organizers of Web publications are also contributing to copyright violations by inadvertently facilitating access to questionable material. Part of the problem lies in attempting to interpret current legislation regarding ownership of electronic publications. Rebecca P. Butler discusses implications for organizing the Internet from the viewpoints of both the owners/publishers and users. She analyzes several strands within the dilemma of the Internet and copyright. Web-based copyright issues are also addressed by Jane L. Hunter in the context of XML-based vocabularies developed to define usage and access rights associated with digital resources.

The next two contributions focus on specific aspects of organization, including discussion of metadata standards and issues of access based on document structure and content. Jane L. Hunter provides an overview of key metadata research issues and current projects and initiatives for improving our ability to discover, access, retrieve, and assimilate information on the Internet. Of particular interest to the end user is her review of

metadata search engine research. Kevin Crowston and Barbara H. Kwasnik continue the issue of vocabulary control in a somewhat different light. Their paper discusses the possibility of improving information access in large digital collections through the identification and use of document genre as a facet of document and query representation. They begin with a framework of the information retrieval problem with respect to genre and finish by outlining a research protocol that would provide guidance for identifying, using, and representing Web document genres.

Sometimes the larger efforts to make Internet documents available fail to fit the local needs of individuals. For example, a teacher in the classroom may have his/her own idea of appropriate resources to complement a lesson plan. Also, traditional methods of classification fail to reflect the constructivist paradigm popular in some educational environments. The belief is that, in order to engage students for maximum learning, there must be some way to not only identify relevant Web sites but also develop ways to explore them. Thus, educators and librarians like to develop customized resource lists that are then also made accessible to other Web users. Don E. Descy describes a variety of tools and techniques that essentially represent an ad hoc method of organizing Internet resources. He makes the point that teachers can construct Web learning environments containing safe sites for students. These can also act as quality information filters similar to the current awareness services as implemented in special libraries in the early days of automation.

In summary, the authors have addressed several dimensions surrounding efforts to organize the Internet. The contributions are of particular value because the content should be of interest to a wide spectrum of users, including librarians, educators, and academic researchers. Furthermore, many of the topics are treated in a fashion that ensures their relevance for a significantly longer period of time than that associated with most activities in a rapidly changing technological world.