PUBLICATION REVIEWS

Creating and Documenting Electronic Texts. By Alan Morrison, Michael Popham, and Karen Wikander. Oxford, Great Britain: Oxbow Books, 2000. $18.00. 63 pp. Glossary and bibliography. Soft cover.

Creating and Documenting Electronic Texts is one in a series of publications produced as part of the Arts and Humanities Data Service (ADHS) Guides to Good Practice. Authors Alan Morrison, Michael Popham, and Karen Wikander are clear about their objectives. Their aim is to take users through the basic steps involved in undertaking a text-encoding project. Their focus is on text, broadly defined. They make several good assumptions: that the reader wants long-term options that are not software dependent, that involve open architecture, and that are "practical and cost effective." Their intent is not to be comprehensive. The authors do not focus on the specifics of rapidly changing technology but rather on the key issues and decisions involved in encoding projects. This is appropriate to the brief introductory overview they aim for. However, it is disappointing to a reader looking for a more thorough coverage of the topic.

The guide has seven chapters, each addressing key concerns, discussing the issues in roughly the same order that anyone implementing a project should. The book begins the way any good project begins: with analysis. They encourage asking such questions as, What is the project scope? What is its purpose? How will people use the result? The answers to these initial questions lead to more specific issues involving selection. They also point out the external factors that need to be considered such as community concerns, funding agencies' concerns, and the compatibility of users' needs and project goals.

This series of questions leads to ever more specific document analysis. Before trying to provide access to a group of items, the authors suggest looking at one item and at its history, condition, provenance, and circumstances of creation. Its provenance is critical for authority, especially with manuscript materials. Furthermore, they suggest selecting a sample for each type of physical object and labeling the features to encode. Both of these steps are critical in preventing future problems.

Once the intellectual process is covered, the authors go on to the physical process of digitization, including scanning, OCR (Optical Character Recognition), and rekeying. Morrison, Popham, and Wikander explain the technical process and endorse the idea that the highest quality digital images are produced from the scanning of the original object. Without falling into the trap of recommending a one-size-fits-all solution, they provide an overview of the capabilities and qualities of various digitization equipment options; provide a quick review of software, image types, file formats, and terms (resolution, bit depth); and include good information on OCR and rekeying.

In the chapter on markup, the key to reusability, they encourage the reader to evaluate goals and determine which method of markup meets those goals. The text outlines three different types of markup. The first relates to formatting instructions found on the physical item. The second has to do with formatting issues such as the use of bold, italics, centering, and bulleted lists, which are tied to the software that creates the text. Third is

generalized, nonproprietary markup, which can be used by various platforms across time and systems and which enables encoding according to content.

Next they focus on visual/presentational markup vs. structural/descriptive markup. Visual structure is paramount in presentational markup. PDF documents are examples of a visual/presentational markup. The authors point out that if presentational style is the reader's only concern, this can be a good option, but if the document "needs to cross platforms or the project objectives require control over the encoding or document preservation," these will not work. They offer a nice discussion of HTML as a manageable subset of SGML and point out that if documents have short-term value or if users are unconcerned with structure, HTML can be a good choice.

The section on SGML/XML and TEI (Text Encoding Initiative) focuses on the structural descriptive markup, which the authors feel is most suited to arts and humanities projects. The chapter provides an overview of SGML and its virtues, namely that it is a standard not tied to proprietary encoding. While providing a brief history of TEI, created to address the need for a common text-encoding scheme, they also supply resources for further information. As SGML (with the exception of TEI) has not been embraced by the academic community to which this text is oriented, they go on to focus on the potential of XML, which has built upon the lessons learned from attempted SGML implementations.

The chapter on documentation and metadata focuses on two currently used metadata models, Dublin Core and TEI Header, and goes into detail about the use and intent of specific tags. They emphasize the comprehensive nature of TEI to document electronic text and its vast potential, but point out that its wide variety in implementation has hindered its use as a standard. They are more optimistic about Dublin Core because its 15-element set is broad and comprehensive enough to be of use to wide audiences.

The book concludes with a "summary" that surprisingly does not revisit previously addressed points. Instead, they outline the 10 steps of an ideal project. This can be a good way to cover topics the authors were not able to address earlier. But it is potentially confusing to readers that new points are brought up at the end.

Creating and Documenting Electronic Texts provides a good summary of the issues involved in beginning a text-encoding project. A good introductory volume is needed for those starting to plan similar projects. The book's order mimics the decision-making process and is helpful in outlining which decisions need to be made. While users will want more specific answers, it is impossible to provide them in an introductory text because those answers depend entirely on project goals. The authors do a good job of outlining the issues and, if readers are focused on textual projects, they will find this volume extremely helpful. However, those considering projects that are more visual or that combine text and images will be disappointed that those issues are not addressed.

Laurie Gemmill
Ohio Memory Project Manager
Ohio Historical Society

Sorting Out the Web: Approaches to Subject Access. By Candy Schwartz. Westport, Connecticut and London: Ablex Publishing, 2001. $32.95. 184 pp. Index, illustrations, and references. Paperback.

Sorting Out the Web addresses the challenges of providing access to networked resources in an information environment that is "complex, rich, volatile, and frequently frustrating" (p. 112). "Sorting" is an apt analogy since Schwartz's focus is not on strategies for searching and navigating the vastness of Internet sites in general but rather on means of partitioning harvested portions of the Web to facilitate meaningful resource discovery. She offers an overview of concepts, techniques, current developments, and the literature from the perspective of the professional resource organizer.

Schwartz's thesis is that time-tested library processes for knowledge representation and information retrieval—selection, description, organization, location assistance—can be applied to on-line resources as fruitfully as to the more traditional library materials. Her discussion, in fact, is oriented toward means of capturing and taming those portions of the Web that comprise materials of substantive content and long-term importance, how to, in effect, build a "library" of networked resources and present that library to potential users. Although the context is not explicitly stated, the tenor of the discussion suggests that such a collection will most frequently be hosted by an academic or cultural institution, perhaps as an extension of its OPAC, and will present materials of interest to that community.

Any writing about the status of the Web is, of course, subject to almost instant obsolescence. Schwartz acknowledges this problem and addresses it by encouraging the reader to focus on the concepts and principles of subject access and resource organization that remain valid even as specific applications evolve.

A brief introductory chapter sets the stage by noting the changes in information availability wrought by electronic representations of materials, with the consequence that end users can now interact directly with retrieval systems without the types of mediation traditionally supplied by libraries and librarians. Rather than viewing this development as a harbinger of librarian obsolescence, she sees librarians partnering with other knowledge experts in guiding users through a chaotic environment.

Chapter 2 defines metadata, likens their creation to the library practices of cataloging and representation, and stresses their utility in supporting discovery, retrieval, and resource sharing by "providing searchable representations of Internet resources" (p.16), particularly if in multiple applications their creators adhere to the same interoperable standards. Schwartz notes in particular their use in creating and maintaining "subject gateways," and in fact the next two chapters expand on this theme by discussing means of organizing and accessing groups of selected resources whose compilers add value through applying descriptive metadata. Chapter 2 also briefly reviews some of the better known "bibliographic" metadata projects and tools, including OCLC's CORC (Cooperative Online Resource Catalog), the Dublin Core, markup languages (HTML, SGML, XML, EAD, TEI), and the Resource Description Framework (RDF). It concludes by noting that "metadata standards development has made possible a wealth of opportunity for resource sharing and interoperability" (p. 39).

Chapter 3, "Classification," discusses some approaches to classifying Internet resources, taking as its premise that "displaying Internet resources in a systematic topical arrangement" has the same effect as does a library classification scheme in permitting the user to navigate through a diverse universe by browsing groups of related materials. Schwartz argues strongly in favor of applying standard, familiar, library-based classification schemes to the organization of harvested Internet resources. The advantages of these schemes (widely used and familiar, readily browseable, able to broaden/narrow/filter, with multilingual access, richly developed and well maintained, available in machine-readable form), she feels, strongly outweigh their shortcomings (overly academic, slow to change, not always intuitive). She concludes the chapter by encouraging more research in the application of library classification principles to the characteristics and needs of Internet resources and their users. The bulk of the chapter constitutes an introduction to several projects in North America and Europe that have applied classification schemes to the organization of Internet resources, in particular the Dewey Decimal System (or verbal representations thereof) and, to a lesser extent, the Library of Congress classification structure and some topically specialized classification schemes.

The next chapter discusses the role that is or can be played by controlled vocabularies in resource discovery and, again, reviews a sampling of projects and applications. As supplements to free-text searching, Schwartz finds controlled vocabulary to be particularly applicable in databases that do not contain full texts of documents; that contain non-text items; or in situations where one wishes to achieve greater recall and/or precision in search results than may be afforded by natural language. She contrasts the browseability and precision of precoordinate indexing (the indexer characterizes the content by bringing preferred terms together) with the richness of access points attained via post-coordinate indexing (the indexer assigns individual terms, which the user's search query collocates). Despite their potential, she feels that controlled vocabularies lose effectiveness in an on-line environment. The applications presented in this chapter make it evident that they are most useful when a subject list or thesaurus is presented, either to click on a term for immediate access to relevant materials or to guide the user in selecting search terms.

Chapter 5 consists of an overview of search engines: the two basic types (classified and query based); how they acquire their content; variations in the elements of the captured Web pages that the search engine indexes; search features and defaults; the types of parameters used to rank search results by relevance; and some of the literature on search engine performance evaluation. An appendix to this chapter summarizes tips compiled by search services and users. Although Schwartz touches only briefly on how classification schemes and controlled vocabularies might affect search engine functions and options, the reader can begin to imagine some scenarios. Noting that "the product of a query in current search engine circumstances is often poorly ordered and bewildering" (p. 121), she suggests that exploring methods whereby users can customize their results and "define an information space through which they can browse" (p. 127) may be more fruitful than attempting to further refine the current generation of general-content search engines. The chapter concludes with an introduction to the concept of subject gateways.

The concluding chapter takes a quick look at current and emerging trends in information retrieval, including machine-aided indexing to improve efficiency and consistency, automated text processing to cluster and rank results, text mining to expose patterns and relationships, and visualization interfaces to sort and cluster search results graphically. Schwartz anticipates a future in which traditional and automated techniques for resource organization increasingly reinforce each other and sees this as the arena in which the principles and practices of subject analysis can play a major role.

It's a little difficult to assess the target audience for this book. It is explicitly intended as an overview of the concepts and tools that support subject approaches to networked resource discovery and, as such, is too abbreviated to appeal to those knowledgeable enough to want to delve deeply into the topic. On the other hand, in a number of places it is perhaps too opaque—or assumes too much prior knowledge—to satisfy the uninitiated. It is most useful probably to students in library and information sciences who will read it in conjunction with supporting literature and discussion and to the many library/archives practitioners who are neither ignorant of nor immersed in this area of endeavor but who want to improve their overall awareness of issues, applications, and trends.

The text is generously illustrated with charts and images of Web pages. Rather than footnotes or a formal bibliography, there is at the end of each chapter a list of sources that discuss in greater detail the specific topics and projects presented in that chapter. Quotations within the chapter are credited via a parenthetical reference to the author's name and the date of the cited work. Many of these sources are articles retrieved from the Web; the citations include their URLs. No URLs are given, however, for the projects and applications discussed in the chapters' text. Schwartz has also pledged herself to maintain (during her working life) her personal Web pages that provide links to projects, sites, references, and other Web-based resources mentioned in the text or pertinent to it. The address of the Web page relevant to each chapter appears at the end of the list of references for that chapter.

Lydia Lucas
Head, Processing Department
Minnesota Historical Society

Visualizing Subject Access for 21st Century Information Resources. Ed. Pauline Atherton Cochrane and Eric H. Johnson. Urbana, Illinois: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 1998. $45.00. 176 pp. Introduction, illustrations, bibliographies, index. Hardcover.

Using the Internet as the first example of what will happen during the next century with "globally distributed information resources" (p.1), *Visualizing Subject Access* deals with questions of how technology can respond to the problems of subject access. This volume is a collection of papers presented at the 34th Annual Clinic on Library Applications of Data Processing, held at the University of Illinois at Urbana-Champaign, March 2–4, 1997. Several individual chapters or portions of chapters are available on the Web. As suggested by the editors, to fully understand many of the presentations requires linking to the URLs listed in the volume.

The conference dealt with a specific series of questions concerning interface and navigational tools and how these coincide with traditional library classification schemes. For example, what is needed and what digital tools already exist to answer the user's needs? Many of the chapters demonstrate that subject access is a question that technology alone cannot solve. Research is required into the processes of information-seeking patterns and cognitive thinking.

Bob Zich's "Visualizing Digital Libraries" points out many of the searching drawbacks faced by computers. Current Web search engines use a linear presentation that, at best, presents the hit list ranked by percentage of relevancy. Zich suggests instead a linkage to relevant Web sites, relevant E-mail-Listserv posts, and the names of experts in that topic. The cues should include color, sound, and visualizations that are inspired from the traditional systems of libraries and card catalogs.

Zich reinforces the point made in earlier articles by Roland Hjerppe and Bryce Allen that searching styles differ for different researchers. Hjerppe notes that some may search by words, others by visual clues. Subject access, particularly as it moves into an on-line environment, needs to take these variables into account. An example of this is that while *Visualizing Subject Access* discusses the value of graphic retrieval, the examples presented in the case studies are primarily text based.

Bryan Allen's "Visualization and Cognitive Abilities" discusses the cognitive process in two parts: spatial and language/symbols. In his research Allen uses specially designed databases, indexes, word maps, and a data presentation screen to determine their effects on user searching. While the study was incomplete, early results showed that the use of a word map did not significantly reduce search time for users. There was, however, much less time spent browsing the subject-heading list. All of Allen's navigational tools are text based but more attention is paid to spatialization. Allen urges further studies into the visualization and navigation of subject access systems.

Many of the studies on retrieval confirm several previously held ideas about searching patterns. Nicholas Belkin, in his analysis of the Rutgers Information Interaction Laboratory, notes that people use varied "normal" searching strategies. Belkin also notes that people prefer systems that they understand how to operate because this leads to greater user satisfaction and because interactive systems that allow query expansion are more effective. Belkin's research echoes Raya Fidel and Michael Crandall's study that

examines filtering criteria used by individuals at Boeing utilizing a bulletin board report system. Participants in the study filtered reports based on the relevance of subject matter, the newness of information, and familiarity with the topic. There was not, however, much uniformity in what defined a relevant subject.

Thesauri and vocabulary play a large role in word retrieval in a networked environment, even as many chapters suggest that navigational tools need to step beyond a strict word-association subject retrieval. Jessica Milstead's "Thesauri in a Full-Text World" provides a history of thesauri in a print-based library past. She points out that, with a growing number of machine-aided indexing terms, the system will continue to use thesauri and it is, therefore, worth the effort to adapt them to the twenty-first century. Searchers also demand indexing, not only at a bibliographic level, but in an increasingly full-text environment. Milstead cites the thesauri's hierarchical design, particularly the use of BT and NT, which is not always known to the researcher nor necessarily relevant to the topic. Current textual analysis software will soon be able to show hierarchical relationships.

A second article by Joseph Busch looks specifically at the Getty thesauri and how data can be similarly structured to generate search terms across multiple databases. The Getty has attempted to use similar types of terms in its *Art & Architecture Thesaurus*, *Union List of Artist Names*, and *Getty Thesaurus of Geographic Names*. Some potential components recommended by RLG for a new vocabulary server have included associative links, equivalent links, notes, and other resource links (Web pages, images, etc.).

This volume—and conference—raise many more questions than answers about the problems of subject access. It does a good job in linking subject retrieval to other fields such as information technology and a study of cognitive processes. The biggest problem, however, is that because many of the presentations are designed for computer systems or are sample database projects, they are not best presented in the traditional bound-volume format.

Melinda McMartin
Assistant Archivist
American Jewish Archives

MIDWEST ARCHIVES CONFERENCE
C/O MENZI BEHRND-KLODT
KLODT AND ASSOCIATES
7422 LONGMEADOW ROAD
MADISON, WI 53717