Encoded Archival Description: The Development of an Encoding Standard for Archival Finding Aids

DANIEL V. PITTI

Abstract: Encoded Archival Description (EAD) is nearing completion and formal release as a standard. EAD attempts to overcome obstacles to intellectual access for geographically distributed primary resources by providing a standard encoding structure for archival finding aids. EAD is the most recent in a line of similar efforts to address universal intellectual access to such data, and like its predecessors, EAD applies emerging technology to the problem. The technology underlying EAD is Standard Generalized Markup Language (SGML) and Extensible Markup Language (XML). This article reviews the background of EAD and the contributions of archivists in both large and small repositories to its development.

About the author: *Daniel V. Pitti is Project Director at the Institute for Advanced Technology in the Humanities at the University of Virginia. Previously, he was the Librarian for Advanced Technologies Projects in the library at the University of California, Berkeley. Pitti is the principal architect of Encoded Archival Description.*

AS ENCODED ARCHIVAL DESCRIPTION (EAD) nears completion and formal release as a standard, it seems useful to recall the long-standing problem that it seeks to address, to survey the technology that it employs, and to recount the process by which its nature and structure have been defined.

Successful innovation does not take place in a vacuum. The intellectual inspiration for innovation comes from tradition, even if at the same time the innovation seeks to transform past practice. The chief motivation for developing EAD was to provide a tool to help mitigate the fact that the geographic distribution of collections severely limits the ability of researchers, educators, and others to locate and use primary sources. Modern attempts to overcome the obstacles presented by the geographic distribution of resources date back to at least the middle of the nineteenth century,[1] and the library and archival communities have been trying in various ways since then to tackle this problem. EAD represents but the most recent and certainly not the last endeavor in this ongoing tradition.

Attempts to address the problem of the geographic distribution of materials have focused on providing universal *intellectual* access. Attempts to solve the problem of universal physical access to the materials themselves, or, more accurately, to their intellectual content, are in their infancy, as the technological means for doing so have only recently emerged. As we began to work toward a standard computer-based data structure for finding aids—the textual analytic guides that control and describe archival collections—we believed that such a standard would be an important contribution toward realizing the long-sought goal of universal intellectual access, but also would set the stage for providing access to the intellectual content of the physical materials themselves.

There is a close relationship between endeavors to overcome geographic obstacles and the emergence of technological innovations; all efforts to improve access have been inspired by new technologies that suggest promising new solutions.[2] EAD is not different from its predecessors in this regard. Emerging computer hardware and software technology, combined with advances in standards and network communications, have stimulated the imaginations of those involved in the development of EAD.

In addition to being an intellectual and technological undertaking, the development of a standard is also a political exercise; it is a community-defining and -building activity. A successful standard must reflect a community's interests, and the community must be directly involved in the standard's development if its interests are to be served. From the beginning of the development of EAD, we have sought to involve the archival community.

Universal Access via Printed Catalogs

The attempt to overcome the geographic distribution of primary sources places EAD development squarely in the mainstream of a major movement in both the library and archives communities that has been making its relentless way throughout most of the current century. Well before the emergence of international networked computing and online catalogs, the library community was working steadfastly to overcome the challenge

[1]Charles C. Jewett, *On the Construction of Catalogues of Libraries, and Their Publication by Means of Separate, Stereotyped Titles* (Washington, D.C.: The Smithsonian Institution, 1853).

[2]Stereotype printing was the technological development behind Jewett's plan to develop a universal catalog. Instead of metal plates, however, Jewett intended to use clay. When the plan failed, it was derisively referred to as "Jewett's Mud Catalog."

presented by the geographic distribution of collections. Initially, these efforts were directed toward providing union access to published materials. In 1909 the Library of Congress began a catalog card exchange arrangement with several major libraries. Herbert Putnam, then Librarian of Congress, described the plan and its purpose as follows:

> The Library of Congress expects to place in each great center of research in the United States a copy of every card which it prints for its own catalogues; these will form there a statement of what the National Library contains. It hopes to receive a copy of every card printed by the New York Public Library, the Boston Library, the Harvard University Library, the John Crerar Library, and several others. These it will arrange and preserve in a card catalogue of great collections outside of Washington.[3]

This was the first tentative step toward what would eventually become the *National Union Catalog*. Other libraries joined the effort, and by 1926, the Library of Congress had compiled a file of nearly two million cards. In 1948 the file was officially named the *National Union Catalog* (*NUC*), and the libraries that had been only selectively reporting acquisitions were asked to report comprehensively.

Gathering the titles together was only the beginning of the effort to create a useful union listing. In order for it to be universally useful, it needed to be universally accessible. It would take until 1956 for the library to develop a solution to this problem by reviving the book catalog, a format which had not been used by most libraries for fifty years. In 1946 the library published *A Catalog of Books Represented by Library of Congress Printed Cards Issued to July 31, 1942*. Ten years later, at the urging of the American Library Association, the Library of Congress applied this approach to the *National Union Catalog* and began issuing in book form the titles acquired by the reporting libraries. This eventually led to the publication of the more than six hundred volumes of *The National Union Catalog Pre-1956 Imprints*, the largest single publication ever produced.[4] For the first time, the library world and the public it served had a system for building a national union catalog and making it universally available. But this union catalog, significant as it was, provided access only to published materials, and not to the nation's rich collections of primary source materials.

In 1951 the National Historical Publications and Records Commission (NHPRC)[5] began to compile a union register of archives and manuscript collections held by the nation's repositories. The objective was to provide central, intellectual access to the nation's primary source materials. The effort initially focused on collection-level summary description rather than on in-depth subcollection or item-level description. After gathering collection-level data from thirteen hundred repositories nationwide in the 1950s, the commission published *A Guide to Archives and Manuscripts in the United States* in 1961.[6] The commission decided to revise the directory in 1974, but, after assessing the situation, found that the number of repositories and records had increased so dramatically in the thirteen years that had elapsed from the publication of the first directory that compiling

[3]*The National Union Catalog Pre-1956 Imprints* (London: Mansell, 1968), vol. 1, vii.

[4]*The National Union Catalog Pre-1956 Imprints*, vol. 1, x.

[5]At the time, the NHPRC was named the National Historical Publications Commission.

[6]This account is based on Richard A. Noble's article "The NHPRC Data Base Project: Building the 'Interstate Highway System,' " *American Archivist* 51 (Winter/Spring 1988): 98–105.

collection-level descriptions would be prohibitively expensive. The commission decided to change the focus to repository-level information and thereby provide a coarser level of access. Despite this shift in focus, the commission continued to envision a "national collection-level data base on archives and manuscripts."[7] For a variety of reasons, the idea was abandoned in 1982.

In 1951, the same year that NHPRC began planning the directory, the Library of Congress began actively to plan the *National Union Catalog of Manuscript Collections* (*NUCMC*).[8] *NUCMC* was intended to be for manuscripts and manuscript collections what the *NUC* was for printed works. Winston Tabb at the Library of Congress has described a major factor in the decision to develop *NUCMC:*

> Scholars, particularly in the field of American history, were instrumental in urging the establishment of a center for locating, recording, and publicizing the holdings of manuscript collections available for research. They had long been frustrated by the difficulties of locating specific manuscripts and even of identifying repositories possibly containing primary-source materials.[9]

It was not until late in 1958 that the Library of Congress began to implement its plans with a grant from the Council on Library Resources. In 1959 the Manuscript Section was established in the library's Descriptive Cataloging Division and was given responsibility for initiating and maintaining the *NUCMC* program. The union manuscript catalog would provide collection-level description for collections held in U.S. repositories and, for particularly important manuscripts, item-level descriptions. Like the *NUC*, the catalog would consist of catalog cards and was to be published in book form, available by subscription. The first volume of *NUCMC* was published in 1962, one year after the NHPRC's *A Guide to Archives and Manuscripts in the United States*. After thirty-two successful years, the library announced in 1994 that volume 29 would be the last *print* publication of the *NUCMC*.

The elimination of the *NUCMC* print publication in no way suggests that it is no longer important to build union catalogs to provide access to our intellectual and cultural resources. Instead, this change was the logical and prudent response to the realization that *NUCMC*'s objective would be better served by using powerful networked computer technology instead of print technology.

Universal Access via On-line Catalogs

The advent of machine-readable catalog records, coupled with the emergence of nationally networked computer databases, provided the archives and library communities with the means to build centralized union catalogs that would be available everywhere, all the time, and in doing so, set the stage for the development of standards such as EAD. For the first time, technology enabled archives and libraries to provide universal access that was not geographically and temporally constrained and thus was far more accessible

[7]Noble, "The NHPRC Data Base Project," 99.

[8]This account is based on the "Foreword" to the *Library of Congress National Union Catalog of Manuscript Collections: Catalog 1991* (Washington, D.C.: Cataloging Distribution Service, Library of Congress, 1993), vii–ix.

[9]*Library of Congress National Union Catalog of Manuscript Collections: Catalog 1991*, vii.

and effective than printed catalogs. Technology also has greatly facilitated the compiling of union databases. Over the course of the 1980s and 1990s, the OCLC and RLG databases, by aggregating millions of machine-readable catalog records, emerged as *de facto* union catalogs to not only the nation's bibliographic holdings, but to a good share of the world's as well. Scholars, educators, and the general public, using networked computers in offices, homes, and libraries, could discover what published materials existed and where copies could be found.

As of 1983, the records in these national utilities almost exclusively represented published print materials; the primary source materials in the nation's archives and manuscript repositories were not represented. This was all to change with the work of the National Information Systems Task Force (NISTF) of the Society of American Archivists. From 1981 to 1984, NISTF paved the way, both intellectually and politically, for the development of the USMARC Archival and Manuscripts Control (MARC AMC) format.[10] The AMC format made it feasible for archives and manuscript repositories to provide brief, synoptic surrogates for collections in their care in bibliographic catalogs. The AMC format by itself, however, only specified content encoding standards; it did not provide standards for the actual content of the records themselves, and without such standards, the format was simply an empty vessel. The archives and manuscripts community found the *Anglo-American Cataloging Rules*, second edition (*AACR2*) inadequate because its chapter on manuscript cataloging abandoned longstanding archival descriptive principles. In response, Steven L. Hensen, then working at the Library of Congress, developed an alternative set of rules that was to complement the AMC encoding standard. These rules, entitled *Archives, Personal Papers, and Manuscripts (APPM)*, coupled with the AMC format, have enabled the archives and manuscripts community to contribute more than 475,000 records to the Research Libraries Group's RLIN database.[11] Through these utilities, scholars now have access to a growing accumulation of brief descriptions of the nation's archival and manuscript collections.

As important and revolutionary as these accomplishments have been, however, they represented only one major step toward enabling researchers to easily locate primary source materials. The generalized descriptions found in AMC records can only lead a researcher to a collection which *may* have individual relevant items. The researcher must next consult the assortment of inventories, registers, indexes, and guides, generally referred to as finding aids, with which libraries and archives have achieved administrative and intellectual control of archival materials in the form of in-depth, detailed descriptions of their collections. Finding aids provide hierarchically structured description, proceeding in defined stages from the general to the specific. At the most general level, they roughly correspond in scope to collection-level catalog records. At the most specific level, they briefly identify individual items. In between, in varying degrees of detail, they describe subsets or series of related items. Finding aids are the detailed maps that lead one from the main highway to the byways, and from those to one's ultimate destination, the item itself.

MARC AMC collection-level records and finding aids are intended to work together as parts of a hierarchical archival access and navigation model. At the top of the model,

[10]For a short history and evaluation of the work of NISTF, see David Bearman, *Towards National Information Systems for Archives and Manuscript Repositories: The National Information Systems Task Force (NISTF) Papers 1981–1984* (Chicago: The Society of American Archivists, 1987).

[11]RLIN database statistics were provided by Ann Van Camp at the Research Libraries Group and reflect the RLIN database as of August 25, 1997.

the AMC record represents a collection in the on-line catalog and leads, through a "finding aid available" note (field 555), to the more detailed information in the finding aid. The finding aid, in turn, leads to the materials in the collection.

In this three-tiered model, the descriptive information in the collection-level record is based on and derived from the collection's finding aid. Only a very small portion of the information contained in the registers and inventories finds it way into the bibliographic record. The summary nature of the collection-level record is dramatically illustrated by the finding aid and catalog record for the *National Municipal League Records, 1890-1991 (bulk 1929-1988)* in the Auraria Library in Denver, Colorado. The finding aid comprises more than fourteen hundred pages and thirty thousand personal names. By comparison, the AMC record for this collection is approximately two pages long and has nine personal names as access points!

Thus, as positive a development as the AMC format has been, it was not the final step in the drive for universal access to primary sources. Nevertheless, AMC was an excellent prologue to the final act. AMC records whetted our appetite for more information and, almost immediately, made us aware of where we should look for it: in the detailed inventories and registers from which the collection-level catalog records had been derived in the first place. It was clear, then, almost as soon as AMC had triumphed, that the next logical step to facilitate scholars' easily locating relevant primary source materials without buying a plane ticket or putting the completion of a research project at the discretion of the U.S. Postal Service was the creation of yet another encoding standard to complement the AMC standard, a standard for the finding aids themselves. And it was equally clear that this standard would lead to the creation of union Internet access to the nation's finding aids for archives and manuscripts.

The Value of Standards

But why insist on the development of a standard? The success of AMC itself should obviate any need to argue the necessity of standards to the archival community, but recent experience has shown that the lure of simple techniques can lead us to ignore lessons already learned. In an era of tightening budgets, it can be difficult to remember that we exploit the new information frontier best if we bring enduring value to it. In the current atmosphere, it is critical to remind ourselves of the importance of standardizing our own time-honored practices rather than rushing to embrace ephemeral digital fashions that will not stand the test of time.

MARC has successfully demonstrated the value of a community-based standard in realizing the goal of universal access to primary resource materials. We are steadily and inexorably moving toward providing comprehensive, universal intellectual access to both primary and secondary resources. This remarkable effort would not be possible without the library community's pioneering work in developing content and structure standards. With the development of AMC, the archival community joined in recognizing the paramount importance of standards. Having grown accustomed to the benefits of well-designed, community-based descriptive standards, it was inconceivable that the archival community could accept proprietary, nonstandard, or worse, substandard approaches to providing universal access to finding aids and the resources they document.

Standards are the foundation upon which individuals sharing common interests form communities, enabling them not only to coexist but also to cooperatively build shared and enduring works. While many archivists were skeptical about the adaptation and use of

bibliocentric library standards, the desire to make archival materials available to users more effectively motivated them to work with the library community. Archivists share with librarians the compelling objective of making information concerning the existence, availability, and nature of the materials in their care more readily available to users. Thus, when the means was found to create surrogates for archival collections in on-line public catalogs, enabling users to locate and identify relevant primary resources more easily, the archival community embraced it.

The Lessons of MARC

Many of the design features successfully demonstrated in MARC are also desirable in an encoding standard for finding aids, and the developers of EAD looked to MARC as a model from the very beginning.

An encoding scheme such as MARC, a computer-readable system for representing the unique, intellectual structure of cataloging data, was absolutely essential if we were ever to build large networked databases that could support sophisticated and effective control, searching, display, and navigation of library collections.[12] Merely transferring complex catalog records into networked computers as unstructured text would not in itself have enabled computers to exploit the complex distinctions and relationships among the elements of descriptive cataloging records.

Cataloging is an "order-making" activity by which complex rules are applied to a defiant, unruly information universe in order to "whip it into shape," making it appear orderly to the users of catalogs.[13] Catalogers determine the identities of authors, works, and items, and the relationships among them. To be usefully exploited by computers, all of this complex order-making data must be explicitly represented in a manner that allows machines to process it with intended, predictable results. Computers cannot reliably perform complex processing on flat, unstructured text, because programmers cannot instruct machines to process that which has not been identified. To take full advantage of network computer technology, it was thus necessary to have an encoding system for catalog data that rendered the boundaries of its intellectual components explicit to programmers and computers alike.

The original designers of MARC saw it primarily as a method for automating the production of printed catalog cards, but they wisely invented an encoding system that would support more than this one use. Given the many uses to which cataloging information would eventually be put, it was important that the encoding scheme developed be sufficiently flexible to support all potential uses. The best way to accomplish this objective was to make the scheme descriptive rather than procedural.

Procedural encoding tells the computer what to do with specified components of a text; by its very nature, it is dedicated to only one use of the information. But as we know, cataloging data is subjected to multiple forms of processing in order to provide effective control, searching, display, and navigation. To support application of multiple procedures, each component could be encoded with multiple processing instructions. This would be highly inefficient, however, because it involves a great deal of redundancy and also forecloses on new processes unforeseen at the time the encoding scheme was developed.

[12]"Control" is here intended to mean "knowing what we have and where it is."

[13]David Levy, *Cataloging in the Digital Order*, <http://csdl.tamu.edu/DL95/papers/levy/levy.html>.

An alternative approach, and the one wisely chosen by the developers of MARC, is to descriptively encode the information. Descriptive encoding involves designating *what* each important component is: a catalog record, an author, a title, and so on. If we know what a data component is, then it is possible to apply different procedures to it based on explicit knowledge of its nature. The decision to make MARC a descriptive markup system ensured that information could be exploited in multiple ways, *and* it left the door open to apply procedures unforeseen in the early stages of development. In addition to faithfully representing cataloging data, MARC's developers also recognized that the system they were designing had to be a publicly owned standard to ensure that cataloging information would endure in an ever-changing computer environment. A standard must not be based on any specific hardware or software platform if it is to endure in our rapidly changing technological environment.

The descriptive nature of MARC encoding, in addition to supporting flexible processing, also supports MARC's long-term survival through means such as mapping MARC data into other computer representations. In fact, most existing MARC systems do not store and use MARC in its native form; to comply with the standard, they simply import and export MARC records. Mapping MARC into a successor standard, if and when one emerges, will be a simple export procedure. MARC's successful survival of the unbelievably rapid transformation of computing over the course of the last thirty years is a testament to the wisdom of its designers. These aspects of the design of MARC—the fact that it is descriptive markup and that it is publicly owned—strongly influenced the developers of EAD and determined the nature of EAD's design to a large extent.

Early in the development of EAD, we surveyed options for the encoding of finding aids. The primary selection criteria were (1) that the system chosen had to be a standard, which is to say, a formal set of conventions in the public domain, not owned by and thus not dependent on any hardware or software producer, and (2) that it had to be capable of faithfully representing the complex intellectual content and structure of finding aids in a manner that would enable sophisticated searching, navigation, and display.

Because of MARC's design qualities, its success in capturing the intellectual content of bibliographic description and the fact that it had been used successfully by archivists for providing collection-level summary access to collections, MARC immediately earned consideration as an option. It was a standard in the public domain. But was it capable of representing the complex intellectual content and structure of finding aids?

After careful study and deliberation, we decided that MARC was not the best available scheme for three principal reasons. First, MARC records are limited to a maximum length of 100,000 characters. This represents approximately thirty $8^{1}/_{2}$-by-11 pages of 10-pitch unformatted text stored in ASCII. Since many finding aids are longer than this, the size restriction was a prohibitive obstacle. Second, and even more significantly, MARC accommodates hierarchically structured information very poorly. Since finding aids are inherently hierarchical documents, the flat structure of MARC makes it unsatisfactory. As archivists are painfully aware, MARC was primarily designed to capture data describing a discrete bibliographic item; complex collections requiring descending levels of analysis quickly overburden the MARC structure. At most, a second level of analysis can be accommodated, but the kind of information supplied is limited.[14] The third reason for not

[14]One possible way around this problem is to employ multiple, hierarchically interrelated and interlinked MARC records at varying levels of analysis: collection-level, subunit, and item. The use of multiple records,

using MARC for finding aids involves the marketplace. It is a gross understatement to say that libraries, archives, and museums are generally not resource-rich institutions. To put this into perspective, the cost of one B-2 bomber would fund the Library of Congress for well over three years.[15] Lacking large amounts of capital, MARC's user community has been incapable of driving state-of-the-art hardware and software development.

SGML, HTML, XML, and EAD

After determining that MARC would not provide an adequate representation of finding aid data, we shifted our attention to Standard Generalized Markup Language (SGML). SGML provides a promising framework or model for developing an encoding scheme for finding aids for a number of reasons. First, like MARC, SGML is a standard (ISO 8879). It comprises a formal set of conventions in the public domain, and thus is not owned by and thereby dependent on any hardware or software producer. Second, unlike MARC, SGML accommodates hierarchically interrelated information at as many levels as needed. Third, there are no inherent size restrictions on SGML-based documents. Finally, the SGML marketplace is much, much larger than MARC's.

While SGML is both standard and generalized, it does not provide an off-the-shelf markup language that one can simply take home and apply to a letter, novel, article, catalog record, or finding aid. Instead it is a markup language metastandard, or in simpler words, a standard for constructing markup languages. SGML provides conventions for naming the logical components or elements of documents, as well as a syntax and metalanguage for defining and expressing the logical structure of documents and relations between document components. It is a set of formal rules for defining specific markup languages for individual kinds of documents. Using these formal rules, members of a community sharing a particular type of document can work together to create a markup language specific to their shared document type.

The specific markup languages written in compliance with formal SGML requirements are called Document Type Definitions, or DTDs. For example, the Association of American Publishers has developed three DTDs: one for books, one for journals, and one for journal articles. A consortium of software developers and producers has developed a DTD for computer manuals. The Library of Congress currently is testing a draft USMARC DTD. The Text Encoding Initiative has developed a complex suite of DTDs for the representation of literary and linguistic materials. DTDs, when shared and followed by a community, are themselves standards.

While MARC is devoted to structuring a specific kind of data, namely cataloging data, SGML is very general and abstract. It exists formally over and above individual markup languages for specific document classes. Because SGML syntax and rules are formal and precise, it is possible to write software that can be adjusted with relative ease to work with any compliant DTD. Typically, an SGML software product has a toolkit that allows the user to adapt its functionality to a specific DTD. As a result, all SGML users,

though, introduces extremely difficult inter- and intra-system control problems that have never been adequately addressed in the format or by MARC-based software developers. Even if the control issues were adequately addressed in the format, the control required to make multiple record expression of hierarchy succeed would entail prohibitive human maintenance.

[15]According to the United States Air Force Web page, the unit price for one B-2 bomber is $1.3 billion. Various other sources place the figure at closer to $2 billion. The 1997 Library of Congress budget is $360,896,000.

not just library and archival users, comprise the market that drives SGML software development.

Similar to MARC, SGML is intended to support descriptive rather than procedural markup of text.[16] As discussed above, procedural markup specifies a particular procedure to be applied to a document component, while descriptive markup defines each component, leaving the processing routines up to applications.

It is useful to distinguish two kinds of descriptive markup: structural and nominal. Descriptive structural markup identifies document components and their logical relationships. Structural elements generally are components that warrant distinct visual presentation: examples include chapter titles, paragraphs, lists, and block quotes. Descriptive nominal markup identifies named entities, both concrete and abstract: examples include corporate names, personal names, topical subjects, genres, and geographic names. While a specific visual presentation of them may be desirable, such elements usually warrant being indexed in particular ways to provide access to some aspect of the document. It is also possible to use SGML to treat the descriptive components of finding aids as named entities. EAD, for example, distinguishes scope and content, biographies and agency histories, chronological lists, various types of administrative information, and many more components of archival description. By explicitly identifying these components, software can be employed to index, search, display, and navigate each component in particular ways.

SGML also supports referential markup. As its name suggests, referential markup refers to information that is not present; it is markup in the third person, so to speak. Referential markup is most commonly used for hypertext and hypermedia, providing the foundation for dynamic references or links to other text and to original digital or digital representations of manuscripts, photographs, audio and audiovisual materials, drawings, paintings, three-dimensional objects of all kinds, chemical formulae, printed pages, music, choreography, and anything else that can be digitally captured and rendered in some useful form. In addition to its many other benefits, using SGML for finding aids offers the exciting option of providing access to digital representations of the primary resources in our archival collections.

HyperText Markup Language (HTML) is an SGML DTD that has enjoyed enormous success as the encoding standard underpinning the World Wide Web. As a specific application of SGML, the HTML DTD limits itself to simple procedural encoding dedicated to on-line display and hypermedia linking. Because of HTML's relative ease of use and its ability to support on-line display of finding aids, many have suggested that it suffices for the encoding of finding aids. The EAD developers felt strongly, however, that HTML was inadequate because its procedural focus would not represent the complex intellectual content and structure of finding aids in a manner that would enable sophisticated searching, navigation, and display. Evidence of HTML's limited ability to support intelligent searching and document discovery, let alone complex display, navigation, and other processing, is not difficult to find. Many of us have used Web search engines to look for both known items and items relevant to a particular topic, and more often than not, we are overwhelmed

[16]For a detailed description of different types of markup, see James H. Coombs, Allen H. Renear, and Steven J. DeRose, "Markup Systems and the Future of Scholarly Text Processing," *Communications of the Association for Computing Machinery* 30 (November 1987): 933–47.

by voluminous results. Our patience frequently is exhausted looking for an item or two that satisfies our need.[17]

The success of HTML as a display format for the Web brings into sharp relief the one major weakness in available SGML software, namely the limited options currently available for delivering native SGML over the Internet. SGML software developers have produced very good and affordable tools for SGML authoring and editing, data conversion, and database indexing and searching; they also have produced very good publishing tools for in-house and CD-ROM publishing. Delivering SGML documents on the Web, however, has been a serious obstacle, but the prospects for this changing in the near future appear to be bright.

In 1996 the World Wide Web Consortium (W3C) founded the XML Working Group to build a set of specifications that would make it easier to use SGML on the Web.[18] The working group, in a short period of time, wrote a specification for a simplified subset of SGML named Extensible Markup Language (XML). Both Microsoft and Netscape have committed to fully implementing XML in their Internet browsers.

The motive behind the development of XML is the recognition that HTML will not support complex, community-based use of shared information on the Internet. HTML hardwires a small set of procedurally oriented tags. Constraining the set of tags has made it easy to build applications that make life relatively easy for authors and Web publishers, and ease of use has been a major factor in the Web's remarkable success. The small, closed tag set, however, has come at a price: HTML has extremely limited functionality. Jon Bosak has identified three areas in which HTML is wanting: extensibility, structure, and validation.[19] SGML is strong in all of these areas, but its strength, like HTML's weakness, comes at a price: SGML is complicated for both application developers and the users of the applications. The W3C's XML Working Group addressed this weakness by identifying and proscribing some features in SGML that are difficult to implement. The result of their work is XML, a simplified subset of SGML for use on the Web.

The ongoing development of XML and closely related standards promises to overcome the last major obstacle to the use of SGML for encoding finding aids: their easy delivery over the Internet.[20] Fortunately, most of the SGML features proscribed in XML were not used in the EAD DTD, and expressions used in EAD that do use proscribed features can easily be expressed in XML-compliant ways. Thus very little modification of the EAD DTD is required to take advantage of future Internet browsers produced and distributed by Microsoft, Netscape, and other vendors, and these changes will have been completed prior to the release of EAD version 1.0.

The decision to develop EAD as an SGML DTD still appears to have been propitious. It allowed us to incorporate MARC's strengths—descriptive markup and public

[17]In response to this problem OCLC has led an international effort since 1995 to develop a simple, generic descriptive metadata scheme that would make it possible to more intelligently index and search HTML documents on the Web.

[18]The original name was SGML Editorial Review Board. Jon Bosak of Sun Microsystems is chair of the Working Group. Other members include Jean Paoli, head of Microsoft's Internet Explorer development, and Tim Bray, representing both Textuality and Netscape.

[19]Jon Bosak, *XML, Java, and the future of the Web*, <http://sunsite.unc.edu/pub/sun-info/standards/xml/why/xmlapps.htm>.

[20]XML includes three related initiatives: XML, Extensible Linking Language (XLL) and Extensible Stylesheet Language (XSL). For current information on the status of the development and the latest drafts of each, see <http://sil.org/sgml/xml.html>.

ownership—and to overcome its weaknesses—limited record and field length, hierarchical poverty, and small market appeal. It was an article of faith when we began developing EAD that to become truly robust the Web would have to outgrow HTML, and that the likely successor to HTML would be based on SGML. This was a calculated risk, but it appears to have been thoroughly justified.

EAD's foundation in the mainstream of library and archives efforts to achieve universal access coupled with the use of emerging powerful computing and network technologies, would appear to provide EAD with everything it would need to succeed. But the most important element of any standards process: the community which will use the standard—also had to be brought into play.

Overview of EAD Development

The success of any standard depends upon broad community participation in its development, followed by widespread recognition of the standard's utility. Standards are the products of communities, not of individuals working in splendid isolation, and the development process is as much a political exercise as it is an intellectual and technical undertaking. Thus, to be successful, an encoding standard for finding aids must reflect and further the shared interests of the archival community and of the agencies and institutions that support it.

From the very beginning of the effort to develop an encoding standard for finding aids, those involved realized it would be crucial to involve the archival community in the intellectual and technical design of the standard. In 1993, when the UC Berkeley library staff was first beginning to contemplate developing such a standard, Jackie Dooley and Steven Hensen both firmly emphasized the necessity of broad community involvement if the effort was to succeed.

The Berkeley Finding Aid Project (BFAP), funded with a grant from the Department of Education's Title IIA Program, began the process that has led to EAD. BFAP's objective was to demonstrate through development of a draft DTD (initially named FindAid), as well as an Internet-accessible database employing the DTD, that an SGML-based encoding standard was both feasible and desirable. To ensure that the prototype DTD reflected the content and structure of the community's finding aids, BFAP staff solicited representative examples of finding aids, regardless of quality, from scores of repositories.[21]

Early in 1995, two developments served to transfer ownership of BFAP's work to the national community. In April the Commission on Preservation and Access (CPA) and the Berkeley library cosponsored a Finding Aid Conference at Berkeley attended by seventy representatives of special collections, archives, libraries, and museums. The purpose of the conference was to evaluate the results of BFAP and to make recommendations about what should be done next. Those gathered enthusiastically agreed that BFAP had succeeded in its limited goals and that the effort should continue, though with the active participation of archival descriptive experts.

[21]The response to this solicitation provides an interesting glimpse into the standards development process. Many repositories enthusiastically promised to send finding aids, but after several weeks, only a handful had arrived. BFAP staff began to approach each repository that had promised to send finding aids to request them once again. Over and over the response was that, while they wholeheartedly supported the effort, they were concerned about how their colleagues would view their finding aids. The finding aids they eventually submitted tended to be those in which they had the most confidence. Thus the community itself voluntarily began to normalize finding aid practice.

The opportunity for engaging archival experts more closely in the project came with the author's successful application to the Bentley Library Research Fellowship Program at the University of Michigan for a team fellowship. The team, led by the author, included noted archival description experts[22] as well as distinguished SGML expert Steven J. DeRose of Electronic Book Technologies. The team met in Ann Arbor in July 1995 to evaluate formally the BFAP finding aid model and DTD and to develop a new model. The team reached early agreement on design principles, which were called the "Ann Arbor Accords," and spent the remainder of the week developing the model on which a new DTD would be based.[23] It was at this time that BFAP was renamed Encoded Archival Description (EAD).

A flurry of activity followed the Ann Arbor meeting. In the next two months, the author wrote the first draft of the EAD DTD. At the September 1995 annual meeting of SAA in Washington, D.C., members of the team began the process of determining appropriate mechanisms for profession-wide adoption and maintenance of an encoding standard for finding aids. The design principles and revised data model were presented to SAA's Committee on Archival Information Exchange (CAIE), and CAIE was invited to become formally involved in the development of EAD. CAIE agreed and created the EAD Working Group (EADWG) chaired by Kris Kiesling and including representatives from the Library of Congress (LC), RLG, OCLC, and SAA. EADWG was charged by CAIE with: 1) assisting in developing and reviewing a data model for archival finding aids; 2) reviewing the EAD DTD; 3) testing and evaluating the EAD DTD; 4) reviewing application guidelines; and 5) initiating review of EAD by the SAA Standards Board and SAA Council. SAA also agreed to formally request that the LC Network Development/MARC Standards Office (ND/MSO) assume the maintenance of EAD once it had undergone thorough community review and was accepted as a standard. In October 1995 LC's National Digital Library (NDL) sponsored a meeting of the team in Washington, D.C. to review the model and draft DTD.

After the October meeting in Washington, ATLIS Consulting Group, under contract to LC and in consultation with the author, began revision of the DTD and creation of a tag library. In a letter to Susan Fox, SAA Executive Director, the ND/MSO formally agreed to be the maintenance agency for EAD, with SAA responsible for ongoing intellectual oversight and development of the standard.

In December 1995 SAA received funding from the Council on Library Resources to create application guidelines for EAD, and at a meeting at UCLA on January 4-6, 1996, the EAD project team met with Anne Gilliland-Swetland and Thomas LaPorte to review the draft DTD and tag library and to outline the content of the guidelines. Further changes were incorporated into the "alpha" version of the DTD, which was completed and released electronically by ND/MSO for use by early implementers in February 1996. On April 27-29, 1996 the EAD team met in Berkeley to discuss the draft guidelines drafted by Gilliland-Swetland and LaPorte and to review suggested changes to the "alpha" version of the DTD that had been suggested by team members and early implementers. Agreed-upon

[22]Other members of the group were Jackie Dooley, University of California, Irvine; Michael J. Fox, Minnesota Historical Society; Steven Hensen, Duke University; Kris Kiesling, University of Texas, Austin; Janice Ruth, Library of Congress; Sharon Gibbs Thibodeau, National Archives and Records Administration; and Helena Zinkham, Library of Congress.

[23]"Ann Arbor Accords: Principles and Criteria for an SGML Document Type Definition (DTD) for Finding Aids," *Archival Outlook* (January 1996): 12–13.

changes were incorporated into the "beta" version of the DTD, which was completed by the author on June 15, 1996 and after review by the development team, was released publicly that September. The draft guidelines, tag library, and encoded examples of a wide variety of finding aids were made publicly available on the Internet in December 1996.

During the course of the EAD development process, a variety of major research and demonstration projects began implementing EAD. From the earliest stages, UC Berkeley, Duke University, and LC's NDL began encoding finding aids using EAD to test its intellectual and technical soundness. Yale University began working with the alpha DTD as soon as it was released in early 1996, as did Harvard University. The University of California, San Diego successfully began experimenting with exporting into EAD finding aids that had been created in a database. SOLINET decided to incorporate EAD into its Department of Commerce-funded Monticello Project, and the NEH-funded Dance Heritage Coalition also made the decision to employ EAD in its archival access project.

Since the EAD beta public release in September 1996, several repositories have initiated finding aid projects of varying size and complexity. The Public Record Office in London is currently developing a strategy for conversion of its repository guide. When completed, this guide will comprise hundreds of thousands of pages describing several centuries of British public records. Several universities in the United Kingdom, including Liverpool, Oxford, Durham, and Glasgow, have substantial EAD projects underway. In the United States, UC Berkeley, with funding from NEH, embarked on the California Heritage Digital Image Access Project. The goal of this project was to demonstrate that USMARC collection-level records linked to EAD-encoded finding aids could provide effective, useful access to collections comprising more than twenty-five thousand digital representations of pictorial materials documenting California history and culture selected from the Bancroft Library's vast pictorial collections. Significantly, the California Heritage Project's prototype access system is being used in an ambitious UC Berkeley K-12 outreach program called the Interactive University Project, which is funded by a Department of Commerce grant. In this project, a team of faculty and library staff are working with K-12 teachers and curriculum planners from the San Francisco and Oakland public school districts to create a teaching program and lesson plans that will use the digital archives to teach subjects related to California history and cultures during the 1997-98 school year and possibly beyond.

The California Heritage Project also has provided the foundation for two other projects, the NEH-funded American Heritage Virtual Archive Project and the University of California EAD Project (UCEAD),[24] the latter funded by UC's Office of the President as the first in a series of UC-wide digital library projects. In addition to building a UC-wide database of finding aids, a key goal of UCEAD was to train archivists at all nine UC campuses to efficiently implement EAD through the use of customized software "toolkits." The American Heritage Project involves a collaboration between Stanford University, the University of Virginia, Duke University, and UC Berkeley; its goal is to demonstrate that EAD can be uniformly applied to diverse existing finding aids for collections documenting American heritage and culture at the four collaborating repositories to enable building a combined virtual archives. The project is exploring the intellectual, political, and technical issues that need to be resolved to provide integrated access to finding aids

[24]The UCEAD Project later was renamed the Online Archive of California.

from multiple institutions.[25] The centerpiece of this project is the development of "an acceptable range of uniform practice" in the application of EAD to existing finding aids. At a meeting in Berkeley in November 1996, representatives from the four collaborating institutions, building on the extensive work of a team of Berkeley technical and archival staff, debated and reached consensus. That consensus was codified in the first draft of the *EAD Retrospective Conversion Guidelines*. Soon thereafter, archivists representing the nine UC campuses met in Los Angeles to launch the UCEAD Project and to further refine the consensus represented by the *Guidelines.* The American Heritage and UCEAD participants, representing twelve university repositories, all agreed to follow these *Guidelines*. These repositories hope that the guidelines will serve as the basis for a discussion leading to a national consensus on "an acceptable range of uniform practice."[26]

The Research Libraries Group recognized that development of EAD training was critical to its community-wide acceptance and use. In the summer of 1996, in collaboration with UC Berkeley, RLG developed the Finding Aid SGML Training (FAST) workshop curriculum. Over the course of the following year, with grants from the Delmas Foundation and the Council on Library Resources, RLG held several workshops in the United Kingdom, Canada, and the United States. Taught by Michael Fox and Kris Kiesling, the FAST workshops have successfully provided initial training to scores of archivists. FAST and other EAD workshops have led to a number of other repositories, large and small, initiating their own finding aid encoding projects. The University of Iowa, University of Vermont, New York Public Library, North Carolina State University, and University of North Carolina, to name a few, all have projects underway. In August 1997 RLG turned the workshop over to the Society of American Archivists at the Society's annual meeting in Chicago, and SAA has now integrated the workshop into its educational curriculum.

RLG and Chadwyck-Healey both are exploring incorporating EAD into their products and services. Following successful development of EAD training, RLG has formed an EAD advisory group to assist in planning and implementing new services. At this stage of planning, RLG intends to provide union access to finding aids worldwide, both those housed on local servers and those deposited on the RLG server by repositories lacking the resources or desire to mount their own findings aids. The advisory group has identified the need for participating repositories to apply EAD uniformly and, in this regard, has decided to use the *EAD Retrospective Conversion Guidelines* to initiate discussions leading to community-wide "best practice" guidelines. In addition, RLG is exploring the feasibility of hosting a retrospective conversion service that would make use of third-party vendors. Chadwyck-Healey is contemplating a similar service and is considering ways to enhance its *ArchivesUSA* product by incorporating EAD-encoded finding aids. In addition to the activities of RLG and Chadwyck-Healey, a number of software vendors have EAD products under development.

[25]Given current technical limits, the project is integrating the finding aids into one centralized database. As technology improves for integrating access to distributed databases (a model much preferred for many practical reasons), the lessons learned from this project will inform migration to the new technology.

[26]Citing a 1980 NHPRC report, Richard Noble reports that commission staff projected that 20,000 repositories and over 700,000 collection descriptions would be included in a national database. See Noble, "The NHPRC Data Base Project," 100. The finding aids in the Berkeley database average twenty-seven pages in length. If this average is representative, then 700,000 finding aids would amount to nearly 19 million pages of text!! It is worth noting again that after only eleven years there are over 475,000 records for archival materials in the RLIN database. Since many of the nation's archival collections have never been processed, arranged, and described, 700,000 may be a conservative estimate.

In addition to the successful transfer of the FAST workshop to SAA, there were several other important developments at the 1997 SAA annual meeting. Jackie Dooley, chair of the SAA Publications Board, reported on discussions with LC concerning the publication of the EAD DTD, tag library, and application guidelines. Kris Kiesling, chair of the EADWG, announced that Meg Sweet of the Public Record Office in the United Kingdom had joined the Working Group and that the Delmas Foundation had funded a meeting of the Working Group in fall 1997 in Washington, D.C. At this meeting, the group reviewed revisions to EAD suggested by the international archival community and, after thorough discussion, decided which changes would be codified prior to EAD's public release as a standard in 1998. The Working Group also reviewed drafts of the tag library, publication of which will coincide with the formal release of the DTD.

Conclusion

Prior to the advent of MARC AMC and *APPM*, the archival community had little motivation to develop descriptive standards. The economic benefits of sharing cataloging that motivated libraries were not available to archivists, whose collections are mostly unique. Nevertheless, archivists wanted to make their materials more accessible, a professional objective they shared with their library colleagues. This desire provided the motivation to explore and eventually embrace MARC AMC and *APPM*, the success of which convinced the archival community of the value and importance of encoding and descriptive content standards. Further, archivists were inspired to want to go beyond summary descriptions and to find a way to provide access to the full, detailed finding aids that constitute the heart of all efforts to make archival collections accessible.

The emergence of the Internet, which has enabled the revolutionary transcending of the spatial and temporal boundaries of our information environment, awakened an abiding but dormant aspiration: to provide comprehensive universal access to the world's primary cultural and historical resources. For the first time in history, it is possible to render the absent present. Not only will archivists be able to better serve those we have traditionally served, but we will also, for the first time, have the means to make our collections accessible to educators and students at all levels and to the general public.

EAD and related standards have initiated the realization of an information future in which serious scholars and the casually curious alike will easily find the cultural treasures they seek. In this emerging future, information seekers will follow clearly marked paths from catalogs to finding aids, and from finding aids to a wealth of information in a multitude of digital and traditional formats. We are embarking on providing not only intellectual access to our collections, but also access to digital facsimiles, at least selectively, of the materials themselves.

While we have not yet fully realized this long-sought goal, and much work remains to be done, it is now possible to begin to envision a future even more promising—one which builds new and unprecedented collaborations between scholars, educators, publishers, archivists, and librarians. Over and above the structured database of catalog records, finding aids, and digital representations of primary source materials, it will be possible to create both private and public information spaces that interpret materials from a wide variety of perspectives and disciplines to serve an equally wide array of cultural needs. Archivists will play an essential role in building the networked digital information environment that promises to transform the intellectual community by admitting new groups of people who, prior to its advent, had never set foot in an archives.