KEYNOTE SPEECH: Digital Libraries in the New Millennium: Partners, Publishers, Potentials, and Pitfalls

David Seaman, Director

Electronic Text Center
University of Virginia
Charlottesville, VA 22903
USA
http://etext.lib.virginia.edu/

As we move into the new millennium it is difficult not to fall into a predictive state of mind. To do this effectively, one needs to look back in order to look forwards in any informed manner. What follows then is a view of the role of the digital library couched in terms of an examination and explication of the part of such a library that I have built at the University of Virginia. The Electronic Text Center, by now a mature service and an ingrained part of our library system, works well as a testing ground for larger digital library issues and definitions. We are variously partners, publishers, researchers into issues and librarians serving a user group which is rapidly diversifying and whose expectations rise constantly. As I look to our achievements and ambitions, and to those of colleagues in other digital libraries, I see the potentials greatly outweighing the pitfalls in the fledgling digital libraries and electronic publishing houses, but both exist in abundance.

Since its inception in 1992, the Electronic Text Center has been an integral part of the University of Virginia Library. From this early start, and from the concentration on long-term, online, standardized data, we have built up considerable expertise in our staff and a driving momentum in our local user community. The Center combines an on-line archive of tens of thousands of SGML-encoded electronic texts and images with a library service that offers hardware and software suitable for the creation and analysis of humanities text (1). Through ongoing training sessions and support of teaching and research projects(2), the Center has built a diverse user community locally, serving thousands of users globally, and providing a model for similar humanities computing enterprises at other institutions. Accounts of our digital library work have appeared in publications as diverse as *The Economist*, *The Chemical and Engineering News*, *The Chronicle of Higher Education*, and *The Australian Financial Review*.

The Electronic Text Center is staffed by a combination of full- and part-time positions, including graduate students drawn from various humanities departments at the University of Virginia. The staff's backgrounds in bibliography, undergraduate teaching, librarianship, textual editing, Special Collections, and graduate research reflects and supports the research and pedagogical needs of our patrons.(3) Indeed, the Etext Center has over the years become an incubator for students who have gone on to make careers for themselves in humanities computing even as the discipline invents itself — we have alumni at the Library of Congress, the University of Georgia, the University of Maryland, Oxford University, the University of Kentucky (fall 1999), and in publishing and business posts that make use of their humanities skills and SGML training.

Defining the Digital library

Since opening in 1992, the Electronic Text Center has pursued twin missions with equal seriousness of purpose:

*to build and maintain an internet-accessible collection of SGML texts and images;
*to build and maintain a user community adept at the creation and use of these materials.
Over the years we have honed the assumptions that lie behind this twin mission; they have worked well for us, and in large measure are generalizable to the notion of a digital library itself(6):

a) Standards are your friend

SGML, MARC, EAD, TEI, TIFF: our life is a flurry of those acronyms that allow us to create data in as standardized a fashion as possible. In this, we are laying the foundation for a permanent digital library collection, and one highly prone to re-structuring and to the re-use of pieces of its data in ways un-imagined at the point of creation. SGML and now XML give us — force us — to describe our data as a hierarchical structure, with metadata and data sections, with divisions containing subdivisions. This notion of the document that both contains the metadata that pertains to it — perhaps the information in the MARC record — along with a data file whose structure is made explicit by tags, gives us highly malleable data, able to change programmatically to another format, and able to deliver pieces of a file — SGML regions — as easily as the file itself. We long ago ceased to think of our data as files that were delivered by a network; rather, we deliver SGML regions — text between SGML tags — that are plucked from one or many files ”on-the-fly” and sent to the user. Such a way of structuring data is unusual to traditional ”file-bound” data managers, but it puts in the hands of the user the data in a way that allows them to build their own compilations, anthologies, and results sets interactively with the data. You want every verse line in English Poetry that mentions ”Sweden”? As long as the SGML is in place you can build your concordance out of thousands of files and create a new document never before conceived of by scholar or librarian: the Anthology of English Poetry about Sweden (it is not a big book, I’m afraid).

b) Online delivery is the dominant delivery

This fundamental aspect of our service has caught attention from our earliest days:
”When Thomas Jefferson built an ’academical village’ in the rolling greenery of his native Virginia, he put a white-porticoed pantheon at its centre and filled it with books. The library’s prominence proclaimed his revolutionary faith in learning and truth. In today’s University of Virginia, a new revolution can be glimpsed....a shared commons in cyberspace, available to anyone prepared to go to a library.” *The Economist*, August 27, 1994. (4)

More and more, our users expect their digital library holdings to come to them, to follow them from office, to classroom, to home, and they chafe at the notion that they have to go to a specific physical place, and a certain set of times, in order to use a piece of digital data — as typically one needs to do when a library holds that data on a CD-ROM. Coming swiftly to us are a growing variety of hand-held devices that will break that online data free from our desktops and our laps and integrate it tightly into our daily routines — cell phones with web browsers are already on the market; GE has in a number of test homes (not mine, sadly) a refrigerator with a touch screen built in, and web access; and at MIT and elsewhere there is much work on wearable computers — ”techstyles” as they are called — that literally will translate the keyboard, cpu and screen into our clothing — literally incorporating the machine into the fabric of our daily lives.

c) Core traditional library skills are also core digital library skills

User training, collection development, and cataloging skills are central to a vibrant electronic collection, and much of our success is a direct result of our ability to integrate the electronic data into the library as a whole, and to draw on the skills that make the print-based library work. The digital library is a library more than it is a computer center. Our work in the Etext

Center is conceived of by us as being more textual than technical — we are emphatically **not** a lab in which books are fed into computers. This point cannot be stressed enough — the core issues of the digital library are user re-training and the management of massive and fluid amounts of metadata and data. The average research library has faced these issues for generations and on a scale that the digital holdings do not yet approach — we track, catalog, and physically circulate 4.5 million print volumes in our library, for example, and the skills that make this work make us work in the new media.

d) Content is king (or queen)

Collections need design, shaping, and building. Our subject librarians allow us to build collections that have coherency; judicious selection serves both our local University users and our Internet visitors equally. Nothing brings people back like a well-selected, reliable, integrated, and cataloged set of data, especially if it is searchable and can be delivered quickly through a common interface across collections.

e) Strength through partnerships

The Etext Center runs on partnerships. We work closely with Special Collections, with local museums, with faculty and student users from near and far, and with other libraries and archives. Some of you may be familiar with our work with publishers — Intelex, Routledge, Accessible Archives, CUP, HarpWeek, Primary Source Media — on individual projects that are delivered as part of our digital library and re-shaped to fit in with a collection from multiple vendors; in doing so, we hope both parties learn something of how to make data that ”plays well with others” in a multi-vendor library setting. You may also be aware of a relationship with Chadwyck-Healey to produce a large product — Early American Fiction 1789-1850 — in which the University of Virginia is paying for the digitization, and owns the product, but has subcontracted to Chadwyck-Healey the sales, product-shaping, and distribution of it as a CD and a web service. Sales, marketing, subscriptions, and fulfillment of orders are all traditional and valued publishing skills, and ones that are difficult for us as a library to duplicate in a cost-effective manner. So far it is proving to be a good combination of talents and interests.

f) User communities need active creation

The value of a library service that provides training in and orientation to a new medium was evident from the beginning — this from early 1994:

”The E-Text Center is both a real place [with] a helpful staff as well as a cornucopia of on-line texts. I’ve dreamed of owning a library of great books, so the mere listing of just some of the e-texts available through the center makes me giddy.” Anne A. Oplinger. ”The Pleasures and Perils of Setting Up House in the Electronic Academical Village.” *UVA Alumni News*. January/ February 1994. (5)

Our frequent and repetitive series of short courses in TEI, HTML, SGML, text-analysis software, scanning techniques, and now XML have contributed to the general vitality of our humanities computing endeavors at Virginia. Well-crafted training sessions and ”walk-in” support provides the primordial soup from which rises both a general literacy in the use of on-line resources, and also a growing suite of individual research and teaching projects.

Faculty and Student Projects

Training is a fundamental element in the building of a user community for electronic resources (see above), and that user community is in turn a fundamental necessity in a vibrant digital library. Faculty and student users create texts, re-purpose existing product, and provide a high degree of critical observation on the various aspects of the digital library in which they invest

themselves. We are delighted to have thousands of users a day from all over the world coming to use our resources, and their feedback helps to improve our mission (on average, we see 50,000 accesses a day, from 10,000 unique host machines a day, representing dozens of countries).

Popular and often award-winning projects include the following past and ongoing undertakings (6), listed here to give some sense of the range of activities that develop speedily once humanities computing services begin to be offered as a serious and integrated part of a library's mission. Dates show when the project began:

FACULTY

- Fall 1998: Witchcraft in Salem Village: the Witchcraft Trials of 1692. Ben Ray, Religious Studies Dept. [will include all the court case transcripts online and freely available — about 1,500 pages]
- Summer 1998: Uncle Tom's Cabin & American Culture: Stephen Railton. University of Virginia English Department.
- Summer 1997: New Religious Movements: Jeffrey K. Hadden, Department of Sociology.
- Spring 1996: Mark Twain in His Times: Stephen Railton. University of Virginia English Department. [Major collections of Twain materials, including rare documents, shaped into a teaching tool]
- 1995: Rita Dove. "Lady Freedom Among Us." [Includes text, images, audio, and video.]
- Spring 1995: Augustus: Images of Power. Mark Morford. Classics Department.
- Fall 1994: American Studies @ The University of Virginia. Alan Howard, English Dept.

STUDENT

- Spring 1999: Searchable database support for The Plymouth Colony Archive Project at the University of Virginia
- Spring 1997: Descriptive Bibliography: an online tutorial. Stephen Ramsay, Department of English.
- Spring 1997: Middlemarch: A Study in Provincial Life: Critical Reception; Publication History; Biography; and Annotated Bibliography. Diana Block, Genevieve Davis, Dorothy Gribbin, Abby Hertzmark, Jennifer Kremer, and Brian Wagner. Department of English.
- Spring 1997: Walt Whitman: Leaves of Grass, The Calamus Revisions. Includes manuscripts, early printed texts, corrected proofs and first editions that pertain to the 1860 "Calamus" cluster of poems. Thomas Lukas, Editor. University of Virginia English Department.
- Fall 1996: Diary and Notes of Louisiana Barraud Cocke. Brendan Coleman, University of Virginia Corcoran Department of History.
- Fall 1996: Harlem: Mecca of the New Negro. A hypermedia edition of the March 1925 Survey Graphic Harlem Number. Matthew G. Kirschenbaum and Catherine D. Tousignant. University of Virginia English Department.
- Spring 1996: The Ladies: A Journal of the Court, Fashion and Society (1872). A journal offering scientifically precise fashion advice and demanding political rights for women. Virginia H. Cope, University of Virginia English Department.
- 1994-1995: British Poetry Archive: Jerome McGann [19th century British poetry].
- Fall 1994: The University of Virginia Fifty Years Ago: 1944-1945 Andrea Nagy. English Department.

Online Journals (7)

- *Studies in Bibliography*: 1948-1997: 50 years of articles on literary, historical and bibliographic matters, from the Bibliographical Society of Virginia.
- *Essays in History* 1990-1998: published annually by the graduate students of the University of Virginia's Corcoran Department of History.
- *IT Journal* : a publication of the Instructional Technology Program at the University of Virginia's Curry School of Education.
- *Journal of Southern Religion* : the first scholarly journal devoted to the study of religion in the American South.
- *The Visual Anthropology Review* [Tables of Contents and information — not full-text]

Online Bibliographies (7)

- *Thomas Jefferson: bibliography of writings about him: 1826-1900*. Frank Shuffelton, editor.
- *Thomas Jefferson: a comprehensive, annotated bibliography of writings about him (1826-1980)* Frank Shuffelton, editor.
- *Thomas Jefferson, 1981-1990 : an annotated bibliography.* Frank Shuffelton, editor. New York : Garland Pub., 1992.
- ASLE: the Association for the Study of Literature and Environment Bibliography, 1993-97.

Online Publications (7)

- Fall 1997: *Shakespearean Prompt-Books of the Seventeenth Century.* Published by the Bibliographical Society of Virginia.
- Fall 1997: Emily Lorraine de Montluzin: *Attributions of Authorship in the Gentleman's Magazine, 1731-1868*
- Spring 1997: *Benedicte Wrensted: An Idaho Photographer in Focus.* The Etext Center and the Visual Anthropology Review are pleased to provide support for this remarkable collection of North American Indian photographs. J. David Sapir, Department of Anthropology.
- *The Online Scholarship Initiative* : UVA Faculty scholarship online. 1995-97.
- *Electronic Theses and Disserations in the Humanities: A Directory of On-Line References and Resources.* Matt Kirschenbaum.
- Timothy D. Pyatt. *Guide to African-American Documentary Resources in North Carolina.* Charlottesville, 1996.
- Michael Plunkett. *Afro-American Sources in Virginia. A Guide to Manuscripts.* Charlottesville, 1995.

Creation, Standards, and Access

Creation

Like any library, the bulk of our SGML electronic text holdings — now numbering about 45,000 titles — are items that we purchase from publishers such as Chadwyck-Healey, Intelex, Accessible Archives, Primary Source Media, OUP, and CUP. However, since we opened in 1992 my staff and I have had TEI text creation as a fundamental daily part of our work. These items, now numbering in the thousands, are mostly in English, but also include Russian, French, Latin, German, and some very well-received online collections of Japanese and Chinese literature — all in searchable online forms.

Our data creation is guided by general library collection development policies, and dictated in large part by the needs of users — faculty and staff who often have a hand in the creation of the text, as we use the exercise as an opportunity to train users as well as a collection-building enterprise. Faculty have added items to the library in order to teach them or to fill a gap in a commercial collection, or to undertake text-based research. Our goal is to help them to create a text that suits both their individual need and is also of appropriate quality to go into our general permanent collections. Indeed, it is precisely the ability to use and search their new text alongside thousands of other items that persuades many user/creators to invest time in such an undertaking.
Notable use has been made of Special Collections, especially (but not uniquely) in the areas of African-American history, the American Civil War, and Mark Twain (8). In recent years we have moved on to larger and more ambitious single data creation projects, in particular the Mellon Foundation sponsored *Early American Fiction (1789-1850)*. This collection draws on our world-class collections of American literature and will include when complete 560 volumes (441 titles) of early American fiction from 81 authors, including works by James Fenimore Cooper, Harriet Beecher Stowe, Nathaniel Hawthorne, Charles Brockden Brown, and Lydia Maria Child., but also by lesser-read novelists such as Delia Salter Bacon, Susan Fenimore Cooper, Nathaniel Parker Willis, and Timothy Flint. Each text exists as a full set of color page-images and a searchable SGML text.(9)

Standards (10)
Data and metadata standards are the bedrock on which we build our long-term digital library. For example, our consistent use of Standard Generalized Markup Lamguages such as TEI and EAD keeps our data usable and malleable, and means that we are able to take advantage of new display and layout possibilities such as XML, as well as new and larger-scale searching and text analysis possibilities. Our long and close partnership with our cataloging department means that our etexts have high-quality metadata, which both eases their conversion to MARC for integration into our general online library catalog, and also allows us to control a rapidly growing collection. We have been among the earliest adopters of both the Text Encoding Initiative SGML tagset for full texts (TEI), and the Encoded Archival Descriptions tagset for archival finding aids (EAD), and have run annual workshops for five years in SGML and digital image creation as part of the UVA summer Rare Book School.

Access:
All online texts in all collections come through a single web interface (of our own creation) that uses the OpenText search engine to search the collections. Having the majority of our electronic texts available on-line affords significant advantages:

On-line texts are freed from the temporal and spacial confines of the library. A user with a computer and a modem does not have to go to the Electronic Text Center to work with most of the texts that it holds, because he or she can access those materials from home, office, or dorm room. Moreover, the electronic text can support many simultaneous users.

In addition to more convenient access to our users, an electronic text collection permits more flexible use. Hypotheses can be tested over massive amounts of data with great speed: a UVa student can use the on-line texts to count and read in context the occurrences of blood images in Macbeth, for example, or search for the earliest recorded use of a word in the English language, or trace an idea or image through the works of hundreds of authors in many languages.

On-line access to the texts also allows us to provide the same search and display ”front-end”

for all our collections. Having been taught to use one database, a user has the knowledge necessary to search all current and future databases, thereby overcoming the frustrations often involved with using CD-ROM products, each of which may have a different interface.

In the seven years that we have been online we have received (and usually replied to) thousands of email, fax, phone and postal messages regarding our materials and services. Along with some less flattering terms, we have been called visionary, cool, overwhelming, and — in one delightfully hyperbolic email — ”the greatest thing since Newtonian physics”. We have provided material from Chaucer and Jefferson to Gene Stratton-Porter and the Native American legends of Iktomi, from classical Chinese and Japanese literature to Boethius and Churchill. We have helped out community colleges with limited textbook budgets, allowed general readers to re-discover long-lost books and find Christmas presents, helped young students improve their grades, made Virginia a more visible college for high school students, and contributed poems to wedding ceremonies.

We have also consciously done what we can with web design and delivery strategies to serve users with disabilities and those on slower or alternative web connections. In short, we have done — in a new medium and for a greatly expanded audience — all the things that a good library, traditionally, has prided itself on doing.

The younger high school students -- who make up a large slice of our 150,000 hits a day (11) — are particularly striking users as they have a touching faith in the Internet’s ability to answer all their questions, and to do it quickly: ”I need the answer before I leave for school in 2 and a half hours”. one hopeful requestor asked; another reported, ”I presented my project today and got an A++! ...Without your help, this couldn’t have been a good project”, another reported. Teachers in the Ukraine find our texts a welcome change from ”old books of Soviet period”, and we have provided material to help prepare students ”for examinations for entry into the Japanese Civil Service” (12).

I’ve always stressed that The University of Virginia’s Electronic Text Center is a local service charged with fostering a local user community, and much of our most informative feedback comes from discussions with an expanding but regular set of UVa faculty and students; however, there is no denying the interest we are developing in serving and learning from our much wider online clientele.

Conclusion

So, a sustainable digital library for us is comprised of standardized data delivered online through a common interface (as opposed to a CD-ROM collection with multiple interfaces); common metadata across collections are already allowing us to achieve cross-database access, and we expect to see this in more and more sophisticated forms. What we invest in is data that is long-term, malleable in its format, interoperable especially at the metadata level, and networkable. Increasingly, stylesheets allow us to provide different renditional views of the same SGML file, with different portions turned on or off from view according to the needs of the user. One view of the file may suppress all scholarly notes, artifactual elements such as deletions on a manuscript, and provide modernized readings of archaic words; another view — this time the same file through a different set of display and rendering instructions — may well display all the scholarly apparatus that the previous user found intrusive.

The next two years are going to see a steady increase in the pervasiveness of XML into our markets, bringing the display flexibility and structural possibilities of XML to a data landscape still too often dominated by non-standard and structureless HTML. As the data becomes more

modular and predictable, we start to see a growth in the number of devices that can deal with it ? pocket electronic organizers and cell phones already on the market can make use of cellular modems to access the web as you walk down the street; the first generation of electronic books allow one to download an ”eBook” into a slab-like device with an LCD screen are already available, and if not entirely useful nonetheless show where the market is heading.

Amidst all the frenzy, excitement and anxiety caused by the arrival of a major new medium, it is heartening to see that parts of the library world has understood and incorporated these skills ahead of the general library patrons and ahead of the commercial publishers, and are firmly on the cutting edge of the world’s use of digital publications.

The digital library succeeds to the extent that it can re-articulate traditional library skills in a new medium — cataloging, collection development, acquisitions, preservation, reference, users services, special collections. Activity in the digital medium allows us to re-articulate the social, pedagogical, and intellectual roles of the library in an academic institution, and to serve as a rich content provider for many thousands of other users worldwide.

Notes

1) The Etext Center’s online collections: http://etext.lib.virginia.edu/uvaonline.html

2) The Etext Center’s services: http://etext.lib.virginia.edu/center.html
The Etext Center’s user projects: http://etext.lib.virginia.edu/projects.html

3) The Etext Center’s staff: http://etext.lib.virginia.edu/staff.html

4) http://etext.lib.virginia.edu/pubs.html#economist

5) http://etext.lib.virginia.edu/pubs.html#alumni

6) The Etext Center’s user projects: http://etext.lib.virginia.edu/projects.html. See also David Seaman, ”The User Community as Responsibility and Resource: Building a Sustainable Digital Library.” *D-Lib Magazine*, July/August 1997. http://www.dlib.org/dlib/july97/07seaman.html

7) The Etext Center’s publications: http://etext.lib.virginia.edu/pubs.html

8)*Mark Twain in His Times*: Stephen Railton. University of Virginia English Department: http://etext.virginia.edu/railton/
Texts by and about African-Americans from the UVA Modern English Collection: http://etext.virginia.edu/subjects/afam.html
Texts About the American Civil War from the UVA Modern English Collection: http://etext.virginia.edu/subjects/civilwar.html

9) *Early American Fiction*: http://etext.lib.virginia.edu/eaf/

10) Standards documentation: http://etext.lib.virginia.edu/standard.html

11) Selected usage statistics: http://etext.virginia.edu/stats/

12) Selected user feedback: http://etext.lib.virginia.edu/etcfeedback.html