Participation in Rare Book School Course: Electronic Formats in a Rare Book Environment

Amigos Fellowship Final Report
October 1997

Submitted by

Nancy Boothe
Rice University

AMIGOS PRESENTATION
NANCY BOOTHE
November 9, 1993

Outline

1. Attendance at RBS course "E-Formats in the Rare Books Environment" with assistance of AMIGOS fellowship: about RBS in general, overview of course.
2. Applicability of course topics to special collections in general.
3. Digitization activities at Fondren (Electronic Studio projects)

   Background: In early 1993, the listserv ExLibris carried a message about the availability by anonymous File Transfer Protocol of the on-line version of LC's Vatican Library exhibit, "Rome Reborn: The Vatican Library & Renaissance Culture." I summoned it, and it turned out to include explanatory text about the exhibit, an outline of the organization, the text of the exhibit catalog, and image files of many of the objects. With appropriate software and hardware you could create and view an electronic surrogate of the exhibit. This combination of textual and visual material, and the marriage of ancient objects/ artifacts with late 20th century technology, was very exciting.

RBS in general:

1. RBS: I had never seriously thought about attending RBS (which was at Columbia for many years before moving to UVA at Charlottesville); I considered the courses too esoteric for me (highly specialized bibliography, cataloging, bookbinding courses). But because I had become interested in electronic applications to special collections material, I applied for acceptance to RBS and to AMIGOS for the fellowship, and here I am. Increasingly detailed information (in both hard-copy and electronically) about RBS, culminating in an indexed, 36-page student handbook with all the information I might conceivably need about RBS and the Charlottesville area: lodging, transportation, weather, food, laundry, banks, special events; just about everything except a reminder to put labels in the clothes we would bring to camp. [Anyone been to RBS?] Info ranged from assurances on our care & feeding (Sunday night supper, breakfast treats before class, coffee breaks, receptions and Fri. afternoon party -- Terry Belanger believes in food for the body as well as the mind --) to a stern admonition that we shouldn't expect to take time off from our intensive 8:30-5 daily schedule to sightsee.

   One of our on-campus housing choices was "The Lawn," two rows of ancient and prestigious cells in front of Jefferson's Rotunda on the university's Central Grounds. During the school year, senior undergraduates compete fiercely for the 52 rooms, which lack bathrooms and air-conditioning, so it was a real privilege for RBS students to have a crack at them. I opted for the local Howard Johnson's.

   Let me read you some of the RBS course description: The premise of the course is that it is the role of the library to collect and make available materials in formats that are suitable for a wide array of applications and on a wide variety of computers. The increasing feasibility of translating flat printed & manuscript library materials into viable, standardized electronic format makes it possible to provide electronic access to collections of diverse formats, which can be made available on a server accessible to a campus or broader community of users. The process of digitizing collections can offer wide-area access and aid in preservation and exhibition. The description went on to say that we would focus on electronic formats of particular relevance to rare books & manuscript work; we would be introduced to E-formats for both conversion of characters to structured searchable text (SGML) and conversion of artifact to image (TIFF and related compression standards). We would discuss standards for text and image and be introduced to various conversion methods. To aid in the process of selecting appropriate formats for digitizing material, we would consider format in relation to collection purposes and anticipated use.

Instructors, all on the UVA staff, were John Price- Wilkin, Information Coordinator, Alderman Lib (primary instructor); David Seaman, Coordinator of Electronic Text Center, Alderman Lib; and Christie Stephenson, Ass't Fine Arts Libn, Fiske Kimball Fine Arts Lib.)

Besides the handbook, I received a 2-page bibliography from JPW with his assurance that although it might look daunting, it would be "a quick read. " Some of it I found in the Rice library, some I had to request through ILL, and yet other items were available only electronically. The quick read turned out to be 6 pounds of paper that I lugged with me because I hadn't waded through it all before I left. Additional course handouts amounted to 3 pounds, so I returned to Houston with 9 pounds of paper. (So much for our electronic, paperless society.)

Eleven students: 8 were special collections librarians/archivists, mostly college & university; 2 systems/ automation people, and an art library cataloger. We met in the library's computer lab. John had created a REBS directory on the UVA gopher, with a file for each class member. We could access our E-mail from the lab terminals, which became as welcome as letters from home to a first-time camper .

We were given an assignment to bring a draft description of a course project creating electronic resources from materials in our collections. The purpose of the projects, which we discussed in our first class session, was to try to evolve a sense of the possible uses of these electronic resources and types of applications needed to facilitate those uses.

Several special collections librarians suggested projects involving digitizing discrete groups of material from their holdings which contained a wide variety of formats, including manuscripts, typescripts, printed material, diaries, architectural plans, sketches, scrapbooks, clippings, maps and ephemera.

Another class member suggested an integrated multi-media system using as its prototype a collection of pamphlets, broadsides, prints and photographs to be available on the home library network, through remote access via modem, and on the Internet.

A project drafted by a LSU librarian was the scanning of different copies of the same rare book in LSU's imaging lab [about which a little more later], storing the images, and converting them to ASCII files to facilitate textual comparison, indexing and keyword searching.

One class member proposed scanning a portion of the microfilmed "churchbook" records of a Baptist historical association for improved readability of the poor-quality microfilm and for wider dissemination.

A university special collections person suggested putting a collection of his university's historical photographic images on his campus network.

In pursuit of a "decision-making matrix" for the creation of E-formats, we then analyzed the types of projects suggested.

- descriptive bibliography
- indexing
- enhancement of quality (as in microfilm)
- instruction
- publication
- exhibition
- preservation (elim. handling)
- textual projects (arrangement, collation, documentary, content/linguistic/geographic/image analysis.
- all involved access (controlled, uncontrolled (as in Internet)

This is a good place to say that throughout the course there was strong and repeated emphasis on the need for meticulous planning -- emphasis on process -- making informed choices, the importance of standards, and the viability of "open systems" systems as hardware-independent as possible. This was probably the greatest value of the course to me. We were exposed to a wide range of digitization methodologies without the intention of becoming technically proficient in anyone. I would say that there was considerably more emphasis on access than on preservation.

Introduction to the InterNet

John Price - Wilkin presented an introduction to the Internet, including definitions and functions of gopher, FTP, veronica, WAIS, World Wide Web, etc. He stressed the rapidity with which changes are occurring, e.g. 6 versions of gopher in a year.

Scanning of texts (OCR)

David Seaman presented a overview of optical character recognition, the process by which a digital image of a page of text is created and then rendered into editable ASCII characters. We discussed issues such as:

- the range of types of text that OCR can reasonably handle
- scanning "difficult texts" such as ornate typefaces, non-traditional character sets
- optimizing sources for the OCR process
- making the best use of the typical variables that an OCR package allows one to set estimating how long a given project will take
- the applicability and limits to training OCR software to recognize new characters. (I found particularly interesting the notion that scanners can be "trained" to recognize recurring images that OCR doesn't recognize; training realistic if there are only a handful of errors or if they are recurring errors of the same type, e.g. capital initial. Presently there is almost no ability to read Fraktur; the most trainable handwriting is Carolingian (14th-century monks learned to write it identically). A scanner that can read mss in general is light years away.

Not a lot of success stories re scanning (OCR); hasn't gotten faster (though hardware may have), but has gotten more accurate and easier to use.

Hand-held vs. flatbed scanners; drop-edge (book saver) flatbed scanner avail; looks like book-edge Xerox. Digital camera evolving; soon may be good enough to use for material you want to scan but for which you can't use flatbed scanner .

Playing with scanning: slash up cheap book and sheet-feed.

Digital Imaging

Christie Stephenson, the specialist in this part of the course, attempted to give us an overall understanding of elements in the decision-making process for capturing and delivering images of materials in the spec. colls. environment, incl:

1. The characteristics of analog and digital input devices and appropriate choice of input device based on characteristics of original and intended use

2. Scanning and image enhancement
    - Basic terminology of scanning
    - Scanning resolution choices
    - Image editing options
    - Issues in image capture
3. Image file formats
    - Importance of standards
    - Advantages and disadvantages of different formats, both proprietary (PICT, BMP) and hardware-independent, such as TIFF, JPEG, GIF, PhotoCD
    - We moved through a guided exploration of several input, scanning and editing devices using slides and photographs we had brought with us. We scanned images at a variety of resolutions and saved them in a variety of file formats. We were able to note and discuss the effects on file size and image quality after this experimentation. The goal of this hands-on session was to expose us to the range of decisions that must be made throughout the process of image acquisition, storage and delivery .

Planning for digization project: factors to consider in selecting appropriate formats for storage and delivery of image files:

- What kind of things are you going to scan?
- Will you need to convert to photographic medium?
- What are physical formats? Do they incl. oversize, 3D materials?
- Will access be organized or unstructured?
- Will use be prescribed or unanticipated?
- Entire collection or part of a whole?
- If diverse materials, need strategies for each type: different parameters for scanning project, based on characteristics of materials, available hardware, staff resources
- Role of preservation in project

Introduction to SGML and TEI

John offered a brief overview of history of formats & applications used in humanities textual computing, and the need for SGML and the Text Encoding Initiative. DEF: Standard Generalized Markup Language: An ISO Standard language for defming markup languages, i.e. sets of markup tags with rules defining when they are applicable, their content, and how they can interrelate. TEI: Text Encoding Initiative: a group working to develop guidelines for the encoding and interchange of machine-readable texts intended for use in literary, linguistic, historical or other textual research; a syntax based on the grammar of the SGML standard.

Premise: JPW:, in an article “Text Files In Libraries: Present Foundations and Future Directions,” (Issue 35 - 9:3 (1991) pointed out the present chaotic diversity of encoding schemes used for such texts makes it difficult to move texts from one software program to others, and researchers who exchange texts with others lose valuable time deciphering the texts and converting into their local encoding scheme. However, he says, this sort of environment, where we require the process of research to take place within the library, and where software impedes comparison rather than facilitating it, does not need to prevail.

The answer, he and others feel, is the creation of E-text centers in university libraries. The goals of the E-text center in UVA’s Aldennan Library, established September 1992, were

1. to provide a software and hardware platform adequate for serving a broad spectrum of textual analysis needs;

2. to facilitate remote, multi-user access; and
3. to begin a local collection of text in a standardized format suitable for a number of applications. They purchase "raw text", e.g. from UMLibText (inc. Modern English & other texts), Oxford Text Archive, Toronto Old English Corpus, and from other univ's computing services. At UVA, convert to a format using SGML-style tagging.

I think the espousal of this position and the pre-eminence of UVA's E-text center was the basis for the course's emphasis on SGML. We viewed examples of text before and after markup (examples), went through an SGML tutorial, and practiced encoding sample documents. This was the most technically intensive part of the course and the one I envision using the least.

Philosophically, I wonder whether it is the role of the library to provide encoded texts for their users? The premise of our instructors was that it is. One forum where this question has been aired recently is the online Discussion Group on Electronic Text Centers. One position (Peter Graham of Rutgers) is that the role of libraries is to make available what they acquire (e.g. text in electronic form) and not to take on the intellectual responsibility for its content. Another point of view is that libraries should develop the "technical, instructional, and consultative skills to support E-texts at the level which humanities centers have to date," at least partly to offset the risk of the increasing marginalization of libraries and their relegation to the "niche " of print media. Another compared the traditions of bibliography and textual criticism associated with rare book rooms (and the sometimes overlapping roles of rare book librarian and bibliographer) to the tagging of electronic texts for the purpose of aiding scholarship.

Library applications of topics presented in course

Text conversion and dissemination

- Electronic Text Center, UVA (& elsewhere)
- Louisiana State university's Electronic Imaging Laboratory: scanning, digitizing, indexing and placing on CD-ROM B.F. French's Historical Collections of Louisiana (1846-53); project eventually to include original manuscripts, photos, slides and drawings.(searchable text file)

Hypermedia projects

- Library of Congress American Memory project: variety of media (photos, movies, sound recordings, documents, pamphlets, interviews) into e-format for access at local work stations; will go on CD's (digital) and videodiscs (analog video). (searchable text)
- Library of Congress exhibits on line (ftp, etc.): Vatican, 1492, Dead Sea Scrolls: introd. text, brochure text, exhibit texts, outline (organization of exhibit), text of related bibliographic sources; and images (in GIF)
- Virginia Commonwealth university's MultiCultural database, documenting cultural diversity in Richmond and central Virginia area, includes minutes, correp., photos, memoranda, newsletters, clippings.
- Users can create hypertext in E-text centers (e.g., UVA) by pulling material from several electronic databases. (OED, Modern English texts, images of Jefferson & UVA, etc).
- Special collections guides on the Internet

Increasingly we are seeing finding aids to manuscript collections and archives in university gophers. At UVA, Special Collections Librarian Michael Plunkett worked with John Price- Wilkin to mount inventories of selected manuscript guides on the UVA gopher, choosing collections for which computer-generated guides already existed. Searchable using WAIS.

Other universities which have finding aids on gopher include Johns Hopkins, Yale, University of Tennessee at Knoxville, Wheaton University, Ohio State University, Trent University, SUNY/Oswego, and the University of California at San Diego, and UT/Austin. (Some may not yet be searchable online.) Oregon State University has gone a bit further by including its University & Records Management Handbook and University Records Retention Schedule online. Also described is a group of historic university photos endearingly called "Harriet's Collection after OSU's first archivist. Just last week, it was announced that the Western Historical Manuscript Collection at the University of Missouri-St. Louis (Anne's old stomping ground) had started placing collection guides on gopher.

Dana Noonan's "Special Collections on the Internet," which she updated last month and disseminated on Network-News, is a valuable guide on finding holdings related to one's topics in the over 700 library catalogs now available on the Internet. Quote: She suggests starting with printed sources, particularly the Guide to Special Collections in the OCLC Database (show copy); many research libraries special holdings appear in their OPACS (and thus in the Internet) as well as in OCLC (Give some southwest examples.) One of my handouts is selected pages of this.

Digitization projects at Rice

"Electronic Study" is the umbrella term for ongoing and new initiatives connecting faculty and students with each other and with online resources on the campus and beyond; promotes the use of network technologies. Projects include Owlnet, an educational network for Engineering and Computer Science begun at Rice in the 1980's; RiceInfo, the university's campus-wide information network, a joint project of Information Systems and Fondren Library which also serves as a gateway to Internet.

Probably the most glamorous electronic application being developed is the Virtual Notebook System" a network-based, multimedia program developed at Baylor College of Medicine, where it was used as a collaborative, medical laboratory notebook, with researchers all entering their research findings into one electronic notebook. A series of demonstration projects were constructed over this past summer to introduce VNS into the humanities and social, as well as science, curriculum at Rice. A virtual notebook can be thought of as the electronic equivalent of a cut-and-paste hardcopy notebook, which might contain a student's course notes, illustrations, instructor's lecture outline, bibliography, book reviews, and many, many more.

Runs on X-windows

Currently, 8 Rice courses are using virtual notebooks: 2 art history, 3 architecture, and one each of political science, biology and math. Collaborative: a page available for feedback from students. Students can work remotely on group assignments. On all notebooks, the "home page" has instructions in using VNS.

The art history notebooks include a guided lecture, a slide viewlist, handouts, and suggested reading lists. Slides brought up with text. Architecture course has an image study guide to help students prepare for midterm and fmal exams. Slides organized as in lecture. On all notebooks, the "home page" has instructions in using VNS.

The differential equation notebook has outside program to do graphical representation of their work. Problem solutions are available and student can ask for help on homework assignment.

Can be run on any UNIX lab facility (in soc. sci., hum., and arch.) Also Rice Advanced Visualization Lab in arch).

Stress that this is evolutionary; things developing, changing all the time.

Strongest element still collaborative nature. Also ability to link with outside systems, draw in resources and capabilities it lacks. So far can't search.