Digital Library Structure and Software

Cavan McCarthy
Louisiana State University, USA

INTRODUCTION

Digital libraries (DL) can be characterized as the “high end” of the Internet, digital systems which offer significant quantities of organized, selected materials of the type traditionally found in libraries, such as books, journal articles, photographs and similar documents (Schwartz, 2000). They normally offer quality resources based on the collections of well-known institutions, such as major libraries, archives, historical and cultural associations (Love & Feather, 1998). The field of digital libraries is now firmly established as an area of study, with textbooks (Arms, 2000; Chowdhury & Chowdhury, 2003; Lesk, 1997); electronic journals from the US (D-Lib Magazine: http://www.dlib.org/) and the UK (Ariadne: http://www.ariadne.ac.uk/); even encyclopedia articles (McCarthy, 2004).

BACKGROUND

Digital libraries require appropriate presentation and careful logical organization to make them easily accessible, but arrangements typical of Web systems are inadequate for them. The classic Web structure, where random links can be created between any pair of pages, is not appropriate to highly organized data. The other classic arrangement is the tree or directory structure often found in computerized systems, where the user starts from a “trunk” or “root directory” and goes to a branch, then a subdivision of that branch. This is effective for individual images, but is inadequate for navigating sequential pages, as in a digital library system presenting lengthy texts. Before discussing the different software solutions available, it is useful to review the principle types of digital material currently offered by digital libraries.

DIGITAL LIBRARY MATERIALS

At this time digital library resources can be divided into three categories: images, texts and other resources:

Images

Image access is used for individual visual resources, such as photographs, posters, drawings, etc. The classic procedure uses a series of three types of image. Scanning produces a high-quality archive image, which is then used to generate an access image, for general public use. Finally, a small thumbnail image is produced, for quick reference (Boss, 2001; Lee, 2001). In more detail:

Archive Image

A high-quality image, scanned directly from the original, destined for long-term preservation. Normally an uncompressed TIF (Tagged Image File Format) image is used here; TIFs offer the highest quality images and a resolution of 600 dpi (dots per inch) is standard. As scanning is an expensive operation, which exposes original materials to possible damage, the archive image will be carefully preserved. It must always exist at the system level, but is not necessarily available to the end-user. TIF files occupy significant server space and imply lengthy download times. Another factor is that some DL will want to sell their own hard-copy prints of quality images.

Access Image or Working Image

A quality image, adequate for consultation and serious study by digital library users. This is normally a high-quality JPG (Joint Photographic Experts Group) image, generated from the archival TIF. JPG files are widely used on the Internet and offer quality spatial and color reproduction and a high compression ratio. For DL purposes JPG images will often be generated at a resolution of 300 dpi; a size of 640x480 pixels is also common.

Thumbnail Image

A small reference image, which gives the user a general idea of the Access image, before downloading that image. Typically a medium to low quality JPG, generated from the Access image, but about one-tenth of its size, and commonly produced at a resolution of 72 dpi. GIF format (Graphic Interchange Format) can also be used for thumbnails (Arizona, 2000; Western, 2003).

Text

Multi-page text documents, such as books, or journal articles require special procedures. Numerous options are possible and the principle alternatives for input, simple text presentations and pagination will be examined in turn; the earliest procedures will be discussed first.

Text Input

Manual keyboarding was originally adopted by Project Gutenberg, the first significant text-oriented digital library (http://promo.net/pg/), founded in 1971. This is a laborious process which severely limits productivity, and is now rarely used.

OCR (Optical Character Recognition) software is now routinely used to scan text into digital libraries. OmniPage Pro (http://www.scansoft.com/omnipage/) or the Russian software ABBYY (http://www.abbyy.com/) are frequently cited in the digital library context. OCR text requires careful revision, because even 99.99% accuracy means that there will be one mistake every couple of pages, but only a person fully conversant with the literature will be able to identify errors at this level. Many digital library texts are older books whose ornate type faces or soiled pages can generate additional OCR errors.

Simple Text Presentations

“Plain-vanilla” ASCII texts were the original basis of Project Gutenberg (http://promo.net/pg/). These TXT files can be used by a variety of computing platforms, but are suited to a word processing, rather than an Internet environment.

HTML text is now firmly established for textual work in digital libraries, including Project Gutenberg. All Internet users are familiar with HTML, files need not be much larger than TXT files, it is easy to present crisp black text on white background, text size can be quickly adjusted on computer screens, HTML can be indexed, readers can easily take extracts from the text, and, within the constraints of ethics and plagiarism, manipulate them and insert them in other documents.

The entire text, without page divisions, is normally delivered by Project Gutenberg as an ASCII or HTML file. Entire chapters or lengthy blocks of text are supplied by other digital libraries, such as Bartleby.com (http://www.bartleby.com/), another pioneering system, which has been offering free copies of classic books since 1963.

Disadvantages of HTML are that it does not necessarily communicate the original text in an integral manner: the “look and feel” may be different, transcription errors may occur and it may not accurately reproduce a combination of text and images.

Paginated Text

Pagination is typical of the traditional book, and most sophisticated digital library software now creates easily-navigated page-by-page sequences. The reader wants above all to go to the next page, also to go back a page when necessary, using appropriate icons or buttons.

JPG images reproduce individual pages exactly, without scanning or transcription errors, making them suitable for research. JPGs are widely used on the Internet and are easy to generate, download, and print; they are also appropriate for mixed text and images. But small size text, older typefaces and brown, mottled or foxed paper may reduce legibility, while readers cannot easily take extracts from a JPG document.

Adobe Acrobat’s PDF (Portable Document File) format offers advantages similar to JPGs, such as quality reproduction of the original in relatively small files, excellent enlargement of the text and easy combination of text with images. Content producers can adjust the security settings within Acrobat to constrain the end user’s interaction with the document. The digital library needs to purchase Adobe Acrobat (http://www.adobe.com/products/acrobat/main.html) to create PDFs, but end-users can easily download the free Adobe Reader.

Digital libraries offering access to texts therefore have two principle possibilities, text, typically delivered in HTML, or page images, which typically use JPGs. In fact, these are not alternatives, they are complementary approaches. The image offers a reliable reproduction of the text, while HTML may be easier for the end-user. Intellectual property and copyright issues may also impact the decision to present either HTML text or images. Even if only images are made available to the end-user, it will be necessary to create a text version in order to generate an index. Word-by-word indexing is a major advantage offered by digital libraries, which cannot be matched by traditional libraries.

For examples of sophisticated paginated access, browse the 8,500-volume Making of America collection from the University of Michigan at http://www.hti.umich.edu/cgi/t/text/text-idx?tpl=browse.tpl&c=moa&cc=moa. This uses DLXS software, discussed in more detail below. The Digital