probably realizes that they should not trust the information in this public-access periodic table, it seems probable that any number of new citizens of the network will rediscover this inaccurate resource in the coming months and years; hopefully, they will quickly realize that it is unreliable, but there are no guarantees. This problem is closely related to a number of currently unclear issues having to do with liability for incorrect software or databases.[64] While liability for failures to provide correctly functioning software and accurate electronic databases are likely to provide fertile ground for litigation in future, many of the issues in this area are focused around some sort of implied quality or appropriateness of a product in the context of a sale or licensing agreement: it is particularly unclear how these issues apply to information resources that are provided on the "use at your own risk" basis that characterizes many (particularly not institutionally supported) "free" Internet information resources today. [65]

10. Access to Electronic Artifacts: The Problems of Versions, Presentation and Standards

The Nature of Electronic Works

When we consider the contents of digital libraries, we encounter a bewildering menagerie of different artifacts: interactive services, computer programs, text files, images, digital movies, networked hypertext information bases, real time sensor feeds, virtual realities. The taxonomy for these artifacts has yet to be firmly established; indeed, one of the greatest sources of difficulties in developing a consensus vision of the electronic libraries we hope to make available through the NREN. There is some basic conception of a "work", however—not necessarily in the legal sense of intellectual property law, but rather an electronic "object" or "document" (in the broad sense), as distinct from a service. Such works can be viewed, copied or referenced; they may be electronic texts, or multimedia compound documents. The are the analogs and generalizations of books, journal articles, sound recordings, photographs, movies or paintings.

The printed page or image can be apprehended with no more technology than the human eye; other works are recorded on storage artifacts such as videotape or phonograph records which require technology to convert them to a form that can be apprehended by the human senses. Yet these storage formats and technologies are fairly well standardized (at least in comparison to the digital world), and, at least to some level of approximation we believe that the content of such a work can be moved from one storage medium to another (say, from 35 mm film to videotape or from LP recording to CD audio disk), perhaps with some degradation of quality but without

---

64 There is reasonably well established liability for provision of incorrect information in commercial settings, such as credit bureaus; the situation for individuals or even institutions that make information available for free (with no promises of accuracy or maintenance) is much less clear. See, for example, [Samuelson 1993] for a further discussion of liability issues.

65 Resources that are institutionally supported are likely to be of better quality than those that are support by the actions of an individual. But it is also important to recognize that there are a number of resources that are offered by third parties, particularly information access providers such as Dialog, that also have major problems with completeness and accuracy.

loosing the “essence” of this content. This relatively high level of standardization is facilitated by the fact that methods of presentation are well established and fairly simple-one listens to sound and views movies. The experience is not interactive. But, when we consider the new electronic “works,” it is clear that they can only be apprehended by the human senses and the human brain interacting with a computer system that includes software and various input and output devices. The experience of these works is complex and interactive; a work can be viewed or experienced in many different ways. Further, other intuitive measures of a work are lost; for example, browsing a printed work gives the browser a sense of the amount of information that the work contains. It is unclear how to measure the amount of information that is contained in a multimedia database.

There is enormous variation among the capabilities of these computer mediators, and too few (or too many) standards. Material may be converted and transferred or displayed in many forms; these transformations can cause major changes in both the presentation and the intellectual content and thus threaten the integrity of the work. Worse, in many cases the content of the work is inextricably bound with the access methods used to view (and navigate within) the work. It is impossible to separate content from presentation of viewing. This inability to isolate intellectual content calls into question the long term and broad based accessibility of works.

Presentation and Content Standards

Standards are needed that permit intellectual content to be encoded into a work in a way that is independent of any specific access software; this is needed to ensure that material can be viewed from any appropriately configured platform, and ensures that the work can outlive specific hardware or software technologies. This is both a preservation issue-ensuring that the work will be available for the foreseeable future-and a portability issue, ensuring that the work will be available from multiple hardware platforms today or in the near future. Today, few useful standards exist, and, for various reasons, those that do are not widely adopted as vehicles for distributing electronic information. Worse: it is becoming increasingly clear that there is a serious lack of consensus as to where intellectual content stops and presentation begins,[66] and the extent to which information providers such as publishers are prepared to distribute content as opposed to specific presentations of content. This controversy has had a considerable impact on the development and acceptance of relevant standards, and most standards currently seeing broad use tend to encode presentation along with content rather than separating the two.

66 This ambiguity between form and content has very significant implications fOr copyright. A number of information providers seem to be taking the position that by simply reformatting or marking up out of copyright works they can return the specific representation of the work to copyright control. This is not necessarily a bad thing, as the availability of out of copyright works in marked up forms such as the encoding specified by the Text Encoding Initiative program adds substantial value, and the copyright protection protects the investment of the companies doing the markup, but it is a source of considerable uncertainty and confusion.

A great deal of electronic information today comes packaged with access mechanisms. In CD-ROM publishing, access software is almost always part of the disk.[67] In other environments one finds information packaged in formats such as Hypercard stacks for the Apple Macintosh.[68] Increasingly, we are seeing the use of programs that provide access to books and journals—Superbook from Bellcore, the RightPages system from ATT Bell Labs and similar systems. Such integration of retrieval apparatus with content raises serious concerns about the ability of libraries to provide access to this information in the long term and across multiple platforms. At the same time, other information providers, such as the Government Publishing Office (GPO) and the Bureau of the Census are creating other problems by issuing data, such as the Tiger files, without any access software, which raises serious questions about the ability of libraries to meet the objectives that are mandated by program such as the Federal Depository Library legislation.

Purchasers of electronic information such as universities and libraries are particularly eager to obtain information in presentation-independent formats such as SGML that allow the information to be reformatted for a wide range of display devices, and even easily processed by computer programs that perform various manipulations or indexing on the information. By contrast, many publishers have expressed a great deal of concern about the loss of control over the presentation of their works; they are worried about both quality and integrity questions. Thus, publishers are in many cases much more comfortable distributing information in formats like bitmapped images or PostScript files that allow reproduction of the information only in a specific presentation (that is, the information can only be replicated either in print or in a screen display with exactly the layout, fonts, and similar properties that were defined by the publisher when the information was created). Another factor in representation choice is that formats such as bitmaps, while they preserve the integrity of the print page, require a much more sophisticated support infrastructure of storage servers, networks and display devices than simple, unformatted text, and thus bar a large number of current network users from access to the material. SGML, while less demanding of storage capacity and network bandwidth, requires sophisticated viewing systems which today serve as a barrier to very broad access.

Even the transfer of files in presentation-specific formats is problematic. Standards for bit-mapped images are still immature; while the Internet Engineering Task Force recently came up with a basic standard for monochrome bitmapped images of pages, [Katz & Cohen, 1992] there is still no agreement on the broader structure that should

---

67 This explains, for example, the great difficulties that libraries and other organizations have encountered in networking CD-ROM databases. Most CD-ROM software is designed to run in the PC environment which does not currently integrate easily or well with commonly used wide area networking technology (i.e. TCP/IP and related protocols). Yet the content of the CD-ROM is so closely integrated with the access software that it is not feasible for purchasers to write their own access software that is more compatible with the network environments in use within the purchasing institutions, and, in fact, the information publishers regard data about how their information is formatted on the disk that would permit the writing of alternative access and presentation software to be a trade secret in many cases.

68 I am not aware of any software in common use that permits browsing of Hypercard stacks on other hardware platforms such as the PC, although there are of course programs that provide the ability to construct and browse similar databases on the PC.

be used for sets of pages[69]. Reproducing such bitmapped image files exactly is still tricky due to variations in the resolution of display and printing devices, which may require interpolation or decimation of the image. The difficulties of printing PostScript files created on one system and then transferred to another system are well known, and include not only header file incompatibilities and PostScript interpreter problems but also problems with fonts.[70]

SGML is frequently suggested as a good prospect for a content standard that can also incorporate some presentation information. However, SGML is really a standard which defines a language for defining document schemas (called Document Type Definitions, or DTDs) and for marking up documents according to a given DTD. While there are numerous attempts to use SGML as a basis for developing industry or application specific document markup standards (for example, the publishing industry has developed ANSI/NISO Z39.59, sometimes called the "electronic manuscript format" which is aimed at the transfer of manuscripts from author to publisher and for use by publishers during the editorial process; the Air Transport Association has developed DTDs for applications such as aircraft maintenance manuals; and the Text Encoding Initiative is developing standards for deep markup to support textual analysis processes of various scholarly disciplines in the humanities and social sciences), it is unclear to what extent these will be accepted.

Currently, most authoring systems do not support SGML (though there is some modest evidence that this is slowly improving). Most documents are either authored using word processors—Microsoft Word, WordPerfect and many others-or using various markup systems such as Troff and Tex. There are converters that move from one of these formats to another, but usually with some considerable loss of information for complex documents. Further, languages like Tex and Troff have many variants and enhancements.

Effectively, if one looks at how documents are being distributed on the network today, there is typically a canonical version of the document that provides as much content markup as possible-typically right out of the authoring system that created it. Users of the same authoring system can use this version. Then there are typically two derivative versions-one in pure ASCII text suitable for viewing on lowest-common-denominator terminals—and one in a PostScript or other print-ready format that (with a good deal of

---

69 The IETF standard allows multiple page images to be transferred as a group, but access to a specific page within the group is awkward. Some publishers are distributing collections of pages as sets of files, with one page per file, using the IETF format. Other groups are looking at higher-level structures that can be used to transfer groups of pages.

70 Fonts represent a particularly subtle problem. While a given font is not protected, certain representations of font families can be copyrighted, as I understand it. Today, in order to facilitate reproducibility, many PostScript files are shipped with their font definitions (which make the files huge), violating the copyrights of the fonts' rightsholders. There are a series of technologies being deployed at present, such as Adobe's SuperATM and Acrobat systems which permit documents to be shipped without fonts. The receiving workstation can use some parameter information to "regenerate" or interpolate a font that is supposed to be similar to the font actually used in the document, thus avoiding not only the copyright problem but also the very real practical problem of receiving a document that employs fonts one does not have on one's local machine. However, this very convenient substitution is performed at the expense of degrading the integrity of the document, and, again, most users may not really even be aware of what is happening, other than that the document looks a little strange.

luck) can be used to produce a printed version of the document. Most of the presentation, and often a great deal of the content, is lost in the ASCII document; the print-image document may or may not be usable, but if it is it preserves presentation integrity at the cost of making most of the intellectual content inaccessible.

While image information does not face the same dichotomy between content and representation a number of the same themes reappear. There is an enormous proliferation of formats for various types of digital images—TIFF (various dialects), GIF, PICT, etc.—which can be intermixed with various compression schemes (CCITT Group 3 or 4, JPEG, etc.). The formats themselves are in some sense more of a nuisance than anything else, and there is software available that converts from one format to another without much, if any loss of information. The compression standards are another matter entirely; JPEG, for example, includes "lossy" compression modes in which the size of the compressed image can be traded off against accuracy of reproduction, with more compact images offering a less detailed reproduction of the original image. Accuracy of reproduction of images—particularly color images-depends of course on having a sufficiently high quality display and/or printing device. But there are also more subtle issues: different monitors, for example, display the color palette differently, and a painting will look quite different when displayed (at identical resolution) from one monitor to another. This is a substantial concern for many color imaging applications, particularly in the fine arts.

There are on the order of twenty different standards for the storage of digital sound at varying degrees of quality and compression; size of the stored sound object depends on sampling rate, dynamic range and the compression scheme used (and whether that scheme is lossy or lossless), among other factors. Many of these schemes are platform specific.[71] More general multimedia standards—for example, for video or other compound objects—are still in their infancy, with a number of platform-specific standards such as Quicktime for the Apple line and Video for Windows coming into use.[72]

As one views these standards in the context of digital library collections one can see a conflict between integrity and access emerging. While it seems likely that over the next few years the industry will continue to adopt more platform-independent standards and that at least some publishers will move towards standards for electronic text that provide more content information and are less presentation specific, it is improbable that at any time in the near future we will find that the user community has installed a base of access devices with homogeneous, high-quality capabilities for reproducing

71 Part of the problem here is that hardware for sound reproduction is not universally available on various platforms (it is now standard within some vendors' product lines, but not across vendors) and the hardware used to play back sound is also not well standardized yet.

72 Quicktime is now available on Windows platforms, at least for viewing movies, and it seems likely that over time most of what come to be the better established digital movie standards will be ported across multiple platforms. This will become easier as the processing to support material that is formatted according to these standards is more readily done entirely in software with acceptable performance and reproduction quality. Right now, these formats push the capabilities of the basic hardware on most common platforms, and often benefit greatly from specialized hardware assistance. It also remains to be seen whether translators from one format to another become readily available for movies as they have for the popular image formats.

sound, displaying color images or viewing digital movies. It is also important to recognize that the capabilities of a user workstation are not the only issue in full quality reproduction of electronic information; network bandwidth is also a factor. The larger the object the longer it takes to transfer it. A high quality digital image or sound recording will take longer to move to the user's workstation for viewing than a lower quality one. In some cases this may translate into a cost issue; if the user is paying connect time for some service, or paying for network bandwidth on a usage-sensitive basis, he or she may be unwilling to pay for a full quality image, or may wish to browse lower-quality simulacra before selecting objects to move in full resolution. For some services, bandwidth limitations may be absolute in the sense that using a full-quality service will make it too slow to be usable. Consider users connecting from home; in most cases they are limited to rather slow connections (9600 bits/second in most cases today, even with relatively expensive and modern modems; perhaps in the not too distant future ISDN at 64 or 128 KBits/second will become a generally available and affordable reality); at these speeds, many users find it more satisfactory to communicate with systems by emulating older character-based terminals rather than running available graphical interfaces based on the X Window system technology. The issue is not that they cannot support the X Window system on their workstations; rather, it is that they don't have the bandwidth on their network connections to run it effectively. It seems likely that libraries, as institutions, will obtain high-bandwidth links to the Internet much more quickly than most end users. and that the issue of available bandwidth to the end user will be a critical factor in determining when information can be delivered directly to the end user in electronic form; it is likely that there will be a long transitional period during which libraries will have to act as staging sites for information on its way to the end user, or to some printing facility in cases where the end user does not want to move the information all the way to his or her workstation prior to printing it.

Thus, information servers such as libraries will face the need to decide, as they design their systems, how much downward conversion they are willing to do in order to permit viewing of material in a degraded format, and how they will make the user of that material aware that he or she is indeed receiving a degraded presentation. Some degraded versions of content will obviate the need for specialized software viewers, thus adding value. Also, while the libraries or other information providers may make the decisions about the spectrum of capabilities available in access systems, rights holders—publishers and authors—will clearly, through license agreements, have a voice in the extent to which libraries will be allowed to apply these capabilities to deliver degraded-quality representations of works to users who cannot view the "canonical" full quality version of the work.

The Problems of Structured Information

There is a natural tendency to view the contents of digital libraries as primarily information intended for direct human apprehension—text, sound, movies and the like. In fact, it is clear that a major class of electronic information resources will consist of structured databases—not databases of abstracting and indexing records or of fulltext, but of genetic sequences, weather observations, commodities prices, chemical compounds and their properties, menus, plane schedules, biographical summaries, and thousands of other groups of information. This information may be viewed by humans