Converting Hard Copy Documents for Electronic Dissemination

Forrest Hoffman
Environmental Sciences Division
Oak Ridge National Laboratory*
P.O. Box 2008
Oak Ridge, Tennessee 37831–6036

* Managed by Lockheed Martin Energy Systems, Inc., under contract DE–AC05–84OR21400 with the U.S. Department of Energy.

The hypermedia capabilities of today's computer systems allow for easier and much-improved information management and dissmenination. The growth of the Internet and the evolution of information-based standards has driven many organizations to provide on-line information within the organization and to the public as well. Hypertext and hypermedia systems will be explained and discussed and the process of conversion of a simple document to an on-line hypermedia format suitable for distribution on the Internet's World Wide Web will be presented.

INTRODUCTION

Since the advent of computer systems, the goal of a paperless office, and even a paperless society, has been pursued. While the normal paper flow in an organization is far from totally automated, particularly for items requiring signatures or authorizations, electronic information dissemination is becoming an almost simple task. The reasons for providing on-line documents are many and include faster and easier access for everyone, elimination of printing costs, reduction of wasted shelf and desk space, and the security of having a centrally-located, always up-to-date document. New computer software even provides the user with the ability to annotate documents and to have bookmarks so that the old scribbled-in and dog-eared manual can be replaced without loosing this "customizability." Moreover, new hypermedia capabilities mean that documents can be read in a non-linear fashion and can include color figures and photographs, audio, and even animation sequences, capabilities which exceed those of paper.

The proliferation of network-based information servers, coupled with the growth of the Internet, has enticed academic, governmental, and even commercial organizations to provide increasing numbers of documents and data bases in electronic form via the network, not just to internal staff, but to the public as well. Much of this information, which includes everything from mundane company procedures to spiffy marketing brochures, was previously published only in hard copy. Converting existing documents to electronic form and producing only electronic versions of new documents poses some interesting challenges to the maintainer or author.

CONVERTING HARD COPY DOCUMENTS

In some organizations, hard copy documents are not retained or maintained electronically. This makes conversion to an on-line product difficult since the documents must be typed in or scanned in. While scanning documents can save a significant amount of time, such documents must either be stored as images, which means that full-text searches are not possible, or they must be converted to text using optical character recognition (OCR) software, which usually introduces its own errors. In any case, the documents must be diligently proofread as if starting from initial document creation. After the text is stored electronically, the job of converting to a useful electronic form only then begins.

McGrew and McDaniel in *On-line Text Management* state, "...you may find that the styles you use to present information in a printed format do not lend themselves to the best presentation on-line" (p. 164). Spacing between list items and before and after headings, which would consume valuable "screen real estate" if preserved from paper documents, is cited as an example of formatting differences between on-line and printed documents. While such formatting issues are of concern, McGrew and McDaniel understate the differences between on-line and hard copy publishing.

```
LINEAR DISCRETE SYSTEMS

A linear discrete system, like that shown in Figure 1, has a transfer
function g(k), accepts an input signal u(k) and provides an output
signal y(k) where

                  oo
                 ___
                 \
          y(k) = /__ g(i) u(k-i)
                 i=0


This is called the convolution-summation representation of a general
linear discrete system (Cadzow, p. 92).

The output signal y(k) of a linear discrete system to a Kronecker delta
input signal, u(k) = (k), is equal to the time-reversed transfer
function of that system.

     Example

     In this animated example, the black box represents a linear
     discrete system. The response function g(k) of the system is
     shown in the box.  As a sharply-peaked input signal is fed
     into the the black box, the output y(k) is seen to be the
     time-reversed response function shown in the box.


REFERENCES

Cadzow, James A., 1973, Discrete-Time Systems: An Introduction with
Interdisciplinary Applications, Englewood Cliffs, New Jersey:
Prentice-Hall, Inc., 440 pp.
```

Figure 1: A typical on-line document.

Figure 1 shows a typical on-line document. This document has been converted to simple text and can be searched and displayed on a terminal screen. However, the drawing described in the on-line document is not available and the equation has been converted to a poor text representation. In addition, the example cannot be shown and the significance of the point being made will be lost on most readers. This document would benefit from conversion into a hypermedia format. This format would not only retain the figure and equation from the original paper document, but would allow for the animated example described in the text. The first step involved in the creation of a hypermedia document is the conversion to hypertext.

HYPERTEXT

Hypertext has been around for a long time and has been applied successfully to educational applications (Jonassen, NATO), but on-line hypertext, particularly that associated with the World Wide Web over the Internet, is the hottest thing going. Hypertext is "the concept of perusing text in a non-linear manner through the use of software links between document elements, as well as through sequential and associative methods" (McGrew and McDaniel, p. 87). This merely means that words or phrases or other parts of a document are linked to related parts of the same or other documents in such a way that the user can move around in or between documents very easily.

Software tools are available for converting word processing files into some kind of hypertext format; however, human intervention is often required to structure the hypertext in a logical and application-dependent way. While some elaborate software packages have been developed in an attempt to eliminate this need for human intervention, these systems have many limitations and are, in general, too expensive for most organizations.

There are many approaches to structuring hypertext. Conceptual structures link document elements with related content, such as explanations which are covered previously or elsewhere in a document. Task related structures are organized to facilitate completion of a task, such as in engine or component assembly. Knowledge related and problem related structures are more hierarchical in design, similar to knowledge-based expert systems. Knowledge related structures are particularly useful for pedagogical applications (Jonassen, pp. 48-54). Different applications call for different structures, and not every document is designed to be read in a non-linear fashion. As a result, not every document will benefit from conversion to hypertext. Nevertheless, on-line availability of linear documents, as well as hypertext documents, is usually desirable.

The document in Figure 1 could be enhanced by linking all occurrences of the words "linear discrete system" to a formal definition which is presumably contained in a related document. The reference to the figure in the first sentence of the document, could be linked to the area of the document which contains the figure. This would enable the user to find the figure quickly by merely selecting the words "Figure 1" in the first sentence. Likewise, the citation made in the second sentence could be linked to the actual bibliographic citation at the end of the document. Additionally, text could be enhanced for emphasis.

HYPERMEDIA

Hypermedia is a super-set of hypertext; it includes hypertext and multimedia objects like graphics, photographs, audio, and video. The World Wide Web (WWW) on the Internet is a hypermedia system which uses the HyperText Markup Language (HTML) (Balasubramanian; Berners-Lee; Connolly; Grobe; Musciano; NCSA; Tilton), based on the Standard Generalized Markup Language (SGML) (Bryan; Smith). Browsers, client software which runs on a user's local machine, such as Mosaic from the National Center for Supercomputing Applications (NCSA), allow the user to access WWW hypermedia from anywhere on the Internet. A nearly exponentially increasing amount of traffic on the Internet results from the use of the WWW by users browsing the interconnected web of hypermedia documents. Converting documents to a hypermedia system like that used on the WWW usually results in much improved documents.

Graphics

In HTML, in-lined graphics, *i.e.*, pictures contained directly in documents, are generally in the Graphical Interchange format (GIF) for 8-bit color images or the X11 bitmap format for 1-bit monochrome images. Such graphics, which include everything from line drawings to business graphics and color photographs, can be scanned in using common page scanners, and can be converted, cropped, scaled, and normalized for use in HTML documents. Other kinds of graphics, including other file formats or PostScript documents, can be included in HTML files as external objects. When converting existing documents to HTML, original photographs are usually sought for scanning. Since much of the graphics used in documents today is computer generated, it is often straightforward to save them in the correct format when they are created. This allows for simple on-line maintenance of images.

The text equation in the document in Figure 1 could be replaced by a graphical image of the real equation. Similarly, the drawing described in the document could be included just as it was in the original hard copy document.

Audio

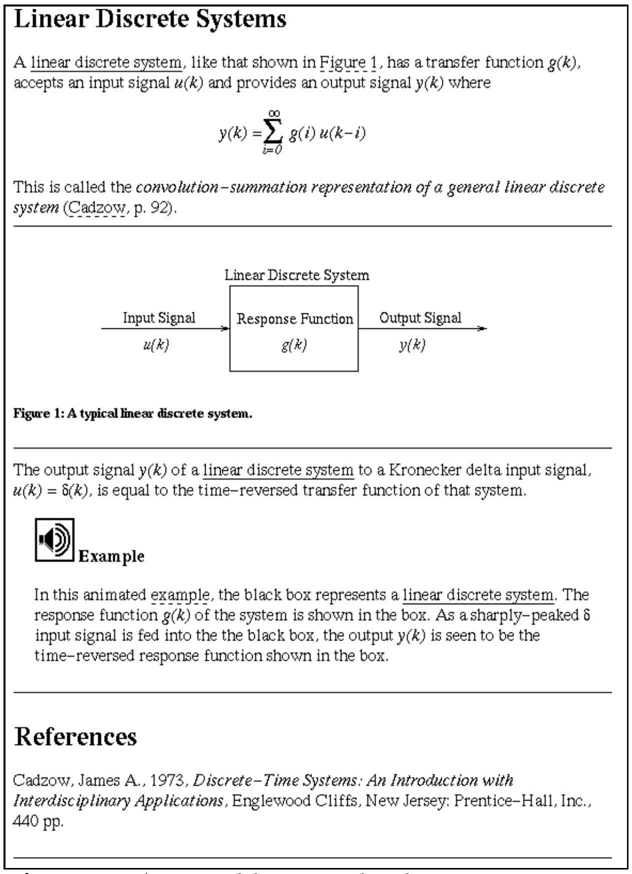
Audio is being used on the Internet for radio programs and conferencing. Audio files are usually in the ULAW format, but other formats are used as well. Audio is used in documents for reading text to the user (often accompanied by background music), for providing information which is only in audio form or best presented in audio form (like music or talk shows), or providing additional information which has not been keyed in but is available in audio form. Generally, audio is used to enhance a document by allowing a user to listen to information instead of just reading it. Audio is also used in conjunction with video and animation sequences. Adding audio to existing documents is not often done but may, depending on the application, be useful in helping users retain information longer since some people are better at retaining auditory information than visual information.

The typical reader would probably grasp the example, contained in the document in Figure 1, more quickly and easily if the text were read to him or her. With audio capabilities, this type of document enhancement can be realized.

Video

Many different video formats are being used to distribute short movies and animation over the Internet. The Moving Pictures Experts Group (MPEG), a body of developers working under the direction of the International Standards Organization (ISO), is still working on what is expected to be the standard for video, which includes both image frames and audio. Movies can be useful for all kinds of applications including assembly manuals, training documents, marketing documents, and much more. Adding movies can really enhance some documents by putting the subject matter in motion.

Figure 2 shows the document from Figure 1 after being converted to hypermedia. Hypertext links are available as described above; some pieces of text are emphasized by changes in fonts; the equation and the drawing are shown as graphical images; the example is read aloud when the user selects the speaker icon; and when the underlined word "example" is selected, an animation sequence is played for the user.



Linear Discrete Systems

A linear discrete system, like that shown in Figure 1, has a transfer function $g(k)$, accepts an input signal $u(k)$ and provides an output signal $y(k)$ where

$$y(k) = \sum_{i=0}^{\infty} g(i)\, u(k-i)$$

This is called the *convolution-summation representation of a general linear discrete system* (Cadzow, p. 92).

Figure 1: A typical linear discrete system.

The output signal $y(k)$ of a linear discrete system to a Kronecker delta input signal, $u(k) = \delta(k)$, is equal to the time-reversed transfer function of that system.

Example

In this animated example, the black box represents a linear discrete system. The response function $g(k)$ of the system is shown in the box. As a sharply-peaked δ input signal is fed into the the black box, the output $y(k)$ is seen to be the time-reversed response function shown in the box.

References

Cadzow, James A., 1973, *Discrete-Time Systems: An Introduction with Interdisciplinary Applications*, Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 440 pp.

Figure 2: A typical hypermedia document.

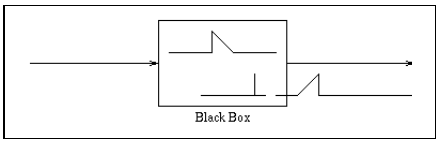


Figure 3: A single frame of an animation sequence.

A single frame from the animation sequence described in the document is shown in Figure 3. When the reader views the animation, he can easily see that the incoming Kronecker δ signal does, in fact, convolve with the response function of the system to yield the time-reversed response function. Reading, hearing, and seeing the process enables the reader to quickly and easily understand it.

CONCLUSION

As document handling capabilities of computer software have improved over the years, many of the benefits of having on-line information have been realized. Distribution of information over the Internet is easily done even with complex hypermedia documents which can include graphics, audio, and movies. Evolving standards and industry pressures will continue to drive organizations toward on-line information management and reduction or elimination of paper-dependent operations.

It has been shown that a short and relatively simple document can be improved using hypermedia. Due to the audio and video capabilities of hypermedia systems, information can be conveyed in multiple ways. This usually results in faster and better understanding of the information on the part of the reader.

REFERENCES

Balasubramanian, V., 1994, State of the art review of hypermedia issues and applications, URL: `http://www.csi.uottawa.ca/~dduchier/misc/hypertext_review/`

Berners-Lee, Tim, 1994, Style guide for online hypertext, URL: `http://info.cern.ch/hypertext/WWW/Provider/Style/Overview.html`

Bryan, Martin, 1988, SGML: An author's guide to the Standard Generalized Markup Language, Reading, Mass: Addison-Wesley Publishing Company, 364 pp.

Connolly, Daniel W., 1994, HTML design notebook, URL: `http://www.hal.com/~connolly/drafts/html-design.html`

Grobe, Michael, 1994, HTML quick reference, URL: `http://kuhttp.cc.ukans.edu/lynx_help/HTML_quick.html`

Jonassen, David H., 1989, Hypertext/hypermedia, Englewood Cliffs, New Jersey: Educational Technology Publications, Inc., 91 pp.

McGrew, P. C. and McDaniel, W. D., 1989, On-line text management: Hypertext and other techniques, New York: McGraw-Hill Book Company, 242 pp.

Musciano, Chuck, 1994, Introduction to HTML, URL: `http://melmac.harris-atd.com/about_html.html`

North American Treaty Organization (NATO) Advanced Research Workshop on Designing Hypertext/Hypermedia for Learning, 1990, Designing hypermedia for learning, Jonassen, David H. and Mandl, Heinz, Eds., NATO ASI Series, Computer and Systems Sciences, Vol. 67, Berlin: Springer Verlag, 457 pp.

National Center for Supercomputing Applications (NCSA), 1994, A beginner's guide to HTML, URL: `http://www.ncsa.uiuc.edu/General/Internet/WWW/HTMLPrimer.html`

Smith, Joan M., 1992, SGML and related standards: Document description and processing languages, New York: Ellis Horwood Ltd., 151 pp.

Tilton, James "Eric," 1994, Composing good HTML, URL: `http://www.willamette.edu/html-composition/strict-html.html`

BIOSKETCH

Forrest is a computing analyst and staff member of the Environmental Sciences Division at the Oak Ridge National Laboratory. His professional interests include computational physics, expert systems, modeling and scientific visualization.