Electronic documents and the World Wide Web

Darren S. Foltinek, Robert R. Stewart, and Don C. Lawton

ABSTRACT

This paper provides an overview of document information structure, describes the World Wide Web project, and then discusses how the Web technologies were used to implement an interactive electronic document.

INTRODUCTION

In any area of research, effective communication of results is always a challenge. The ability to independently verify published results, while very important, is currently quite difficult, especially in the field of exploration geophysics. Traditional means of publishing do not include any raw data, nor the software that was used to process, produce or visualize the data. An interactive document would come with both data and software, enabling the reader to easily reproduce results and run their own experiments.

In writing this paper, we have used some terminology that may be unfamiliar to the reader. Please see the glossary at the end of the paper for clarification.

DOCUMENT INFORMATION STRUCTURES

In a traditional document such as a book or technical paper, information is presented in a linear, or sequential, order. The person reading the document usually starts at the beginning and proceeds through the information in the order in which the author indended. There are several advantages with a linear document; it can be easier to read and author because it is the traditional, and thus more comfortable, method. Since the order of presentation is determined by the author, the reader can be 'lead by the hand' through the material. It is presented in an order that the author feels is the most appropriate. The counterpoint to this advantage is that without knowing the level of background knowledge of the reader, it may be difficult to define a 'most appropriate' order of presentation.

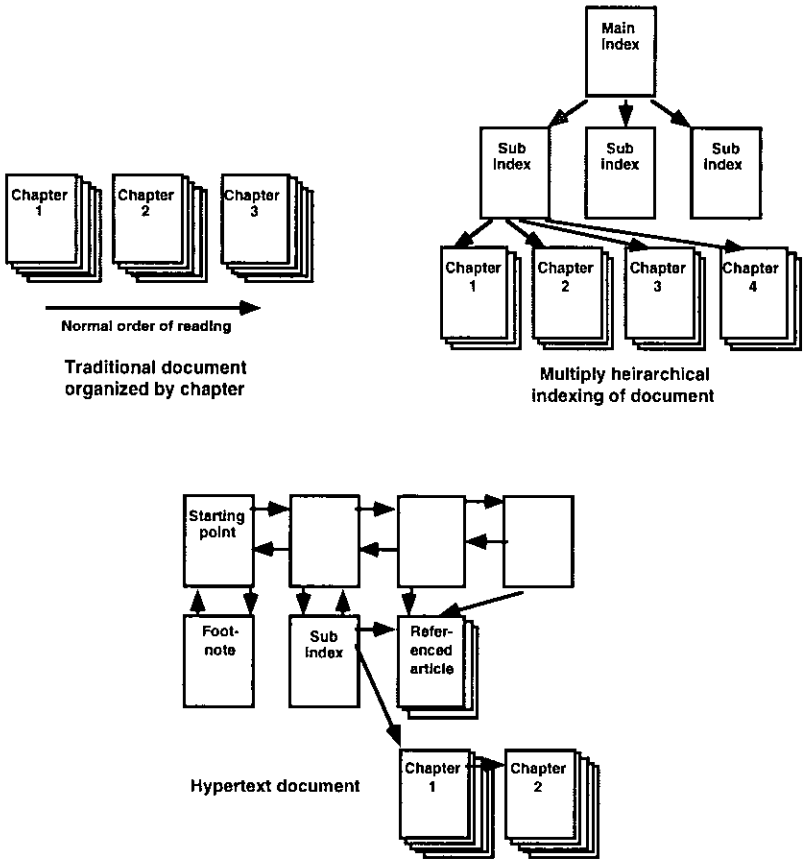

FIG. 1. Document information structures

When a document is electronic, the presentation possibilities are much greater. The information within the document may be accessed through a heirarchy of menues, through hypertext links, by query, as a traditional linear document, or any combination of methods. When a document is authored so that its contents can be accessed in a user-controlled manner, several advantages become evident. The reader now has control over the order in which the document is read. They can choose to read what they think are the most important parts first, and explore details and related topics later. Information is easier to find due to the random access nature of the document. For the author, it is easy to expand or restructure the document while writing it, since the information is broken into small, interlinked components.

Authoring Hypertext Documents

The first and most important step in authoring a hypertext document is to structure the information you wish to present in a modular manner. Hypertext documents can, of course, be similar in structure to a traditional book, organized by chapter and paragraph. A more radical approach to structuring a hypertext document would be to write a single page on each of the concepts you wish to present and then place hypertext links between these ideas. A compromise, however, between these two information structure extremes is more likely to produce a document that is easy to read and therefore more useful to its intended audience.

After the overall information structure of the document has been determined, the authoring techniques used to create a hypertext document are very similar to those used

in a traditional document. Word processors are used to enter the text, and drawing and visualization packages are used to create the figures.

THE WORLD WIDE WEB

Web Summary

This description of the World Wide Web (WWW) has been extracted from the World Wide Web Summary (Berners-Lee, 1993). In fact, it was found and extracted in only a few minutes by accessing the Web from the same computer on which this paper is being written, a small example of the power of this technology.

> "The WWW project merges the techniques of networked information and hypertext to make an easy but powerful global information system. The project represents any information accessible over the network as part of a seamless hypertext information space.
>
> It is currently the most advanced information system deployed on the Internet, and embraces within its data model most information in previous networked information systems.
>
> The WWW browsers can access many existing data systems via existing protocols (FTP, NNTP) or via HTTP and a gateway."

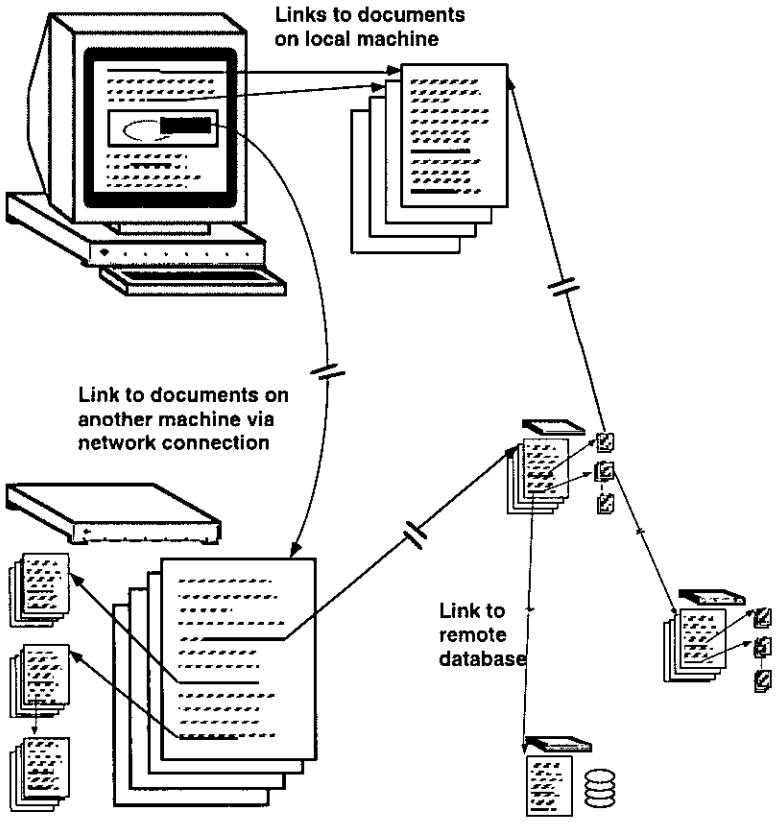

FIG. 2. The World Wide Web: a global multimedia hypertext system.

History of the Web

The Web project was first proposed in March of 1989 by Tim Berners-Lee at CERN (European Laboratory for Particle Physics). By May of 1991 the WWW was running on central CERN machines, and by January 1993 there were 50 known hypertext (HTTP) servers running on the Internet. Browsers for Unix, Mac and DOS machines had also been developed.

CERN maintains a Virtual Library containing over 80 subjects, each subject containing lists of subtopics and/or lists of servers. There is a searchable index to this subject list containing over 12 000 entries. The Geology and Geophysics Department at the University of Calgary maintains the Virtual Library list of Earth science servers which now has about 60 entries, including university departments and government organizations. The CREWES Project maintains the subject list for Geophysics, with links to about 30 Web servers. The Web has experienced explosive growth, an example of which is the growth in usage of the CREWES server, shown in figure 3. Estimates of the growth rate of Web usage are about 3000% per year (Gray, 1994), based on the network traffic in 1994. As the Web has grown, so too has it become commercialized. At last count there were almost 900 commercial Web servers in existence. These statistics are all current as of November, 1994.

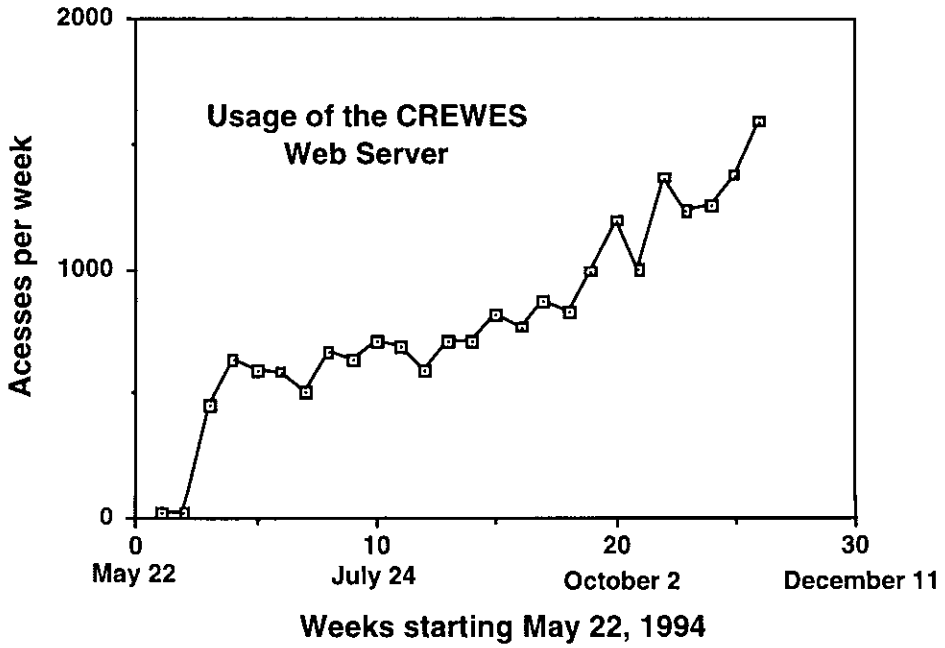

FIG. 3. Usage of the CREWES Web server.

Web Technology Capabilities

Presently, the technology of the Web allows interactive, multimedia hypertext documents to be published on the Internet, where they are available to anyone with an Internet connection. From a single user interface (browser program), it is possible to access all of the previous Internet information providers. This allows simple point and click access to data and file archives on FTP servers, Usenet news on NNTP servers, the earlier network hypertext system known as Gopher, as well as the hypermedia HTTP servers. Software tools for writing and distributing documents are all available in the public domain. Browser programs (document viewers) and HTTP servers are

available for most common operating systems, including MS-Windows, Apple Macintosh and X-Windows.

Web Browsers

The browser program is the user's interface to the Web. These programs are available for a wide range of hardware and software platforms, and may either be text or graphics based. When a text-based browser accesses a Web document containing graphics, it simple ignores the graphic element.

The most popular Web browser is called Mosaic. This program, written by the U.S. National Center for Supercomputer Applications (NCSA), is fully graphics-capable, runs on the major computer operating systems, and is freely available on the Internet.

Publishing on the Web

To publish a document on the Web, it is necessary to make the document available to an HTTP server and then to establish links to your document so that others may find out about it and access it. If your site already has an HTTP server running it is a simple procedure to allow it to access your document. Setting up the server is beyond the scope of this paper.

Once the document is on an HTTP server it is accesible to anyone on the Web, but unless it is advertised, no one will ever know about it. This advertising is usually done by adding links to your document to Web subject lists or archive sites. There currently exist lists covering many different subjects, including Earth Sciences and Geophysics.

INTERACTIVE BOOK PROJECT

Goals

The overall goal of the CREWES interactive book project is to develop a means of disseminating information such that the reader can interact with the document. By allowing the reader to interact with the document, the speed of learning and depth of understanding should increased, since the learning process is no longer passive. Within the CREWES Project, we produce an annual Research Report, run courses, and deliver software for our sponsors. Each of these deliverables would be improved by being produced in an electronic, interactive format.

Another goal is to improve the reproducibility of research results. This is possible within the framework of an electronic document because of the increased level of integration between documentation, theory and implementation.

Implementation of Prototype

User interaction with a document can take place on several levels.

- The document can be annotated by the user, similar to a reader scribbling notes in the margin.

- The figures can be regenerated by the user using different display parameters or data.
- The document itself can be modified, including the text, figures and data.
- The methods or algorithms used may be modified.

The level of user interaction considered here to be a minimum was that the figures should be able to be regenerated by the user. This implied modification of the display parameters, and possibly the data itself. Annotation of the document by the user would also be highly desirable.

There were several issues that needed to be resolved before a prototype was implemented. A development platform was selected by evaluating the different toolkits available based on functionality, compatibility within our working environment, flexibility, and cost. The problem of how to integrate the many different tools that researchers and authors use is one that will never be fully resolved, since these tools are constantly changing. The size of geophysical datasets will always cause problems in the areas of storage space, processing and rendering time and data transmission times.

After evaluating several different implementation platforms, it became clear that the amount of work required to implement an interactive electronic document increased with the desired level of interactivity. For the greatest level of interactivity, the electronic book would essentially be a custom application. A compromise on the level of interactivity would be needed in order to have resonable development times.

The Web technology (HTML documents and HTTP information servers) was chosen as the development platform for several reasons. This technology is capable of implementing the required level of interaction. The platform is non-propietory, multi-platform, and in the public domain, thus minimizing the costs and maximizing the ease of distribution of a document. The most unique capability of the Web, however, is that it is one of the only, and by far the most widespread, **networked** multimedia hypertext system in existance. Its explosive growth rate is indicitive of its widespread acceptance, and it is therefore likely to be a stable development environment. Because the Web document environment is very modular in nature (see figure 4) adding elements or modifying a document is very easy.

Creation of the interactive Web document involves implementing each of the modules in Figure 4.

- HTML page with user input, control buttons, menus etc. is created.
- control program is written to interface between the server and the data processing and rendering applications used to create the figures. Control program also generates a new HTML page with the new figures.
- the applications used to render the document figures must either be custom-written, modified, or have their output translated into the Web standard graphic format (GIF).

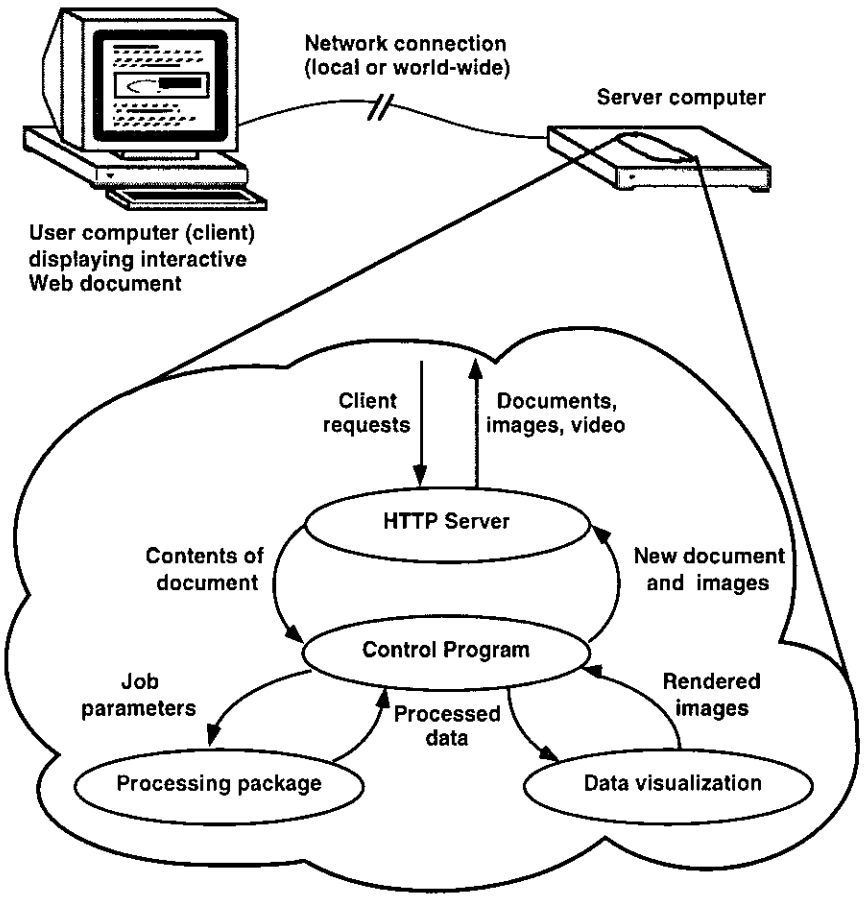

FIG. 4. Data flow in a Web interactive document

An important item to note is that this scheme may require a fair amount of custom code to be written for each document. Specifically, there has to be a control program written for each figure in the document. This control program interfaces between the HTTP server and either custom written or commercial data processing and visualization packages.

P-S Foldmap example

As a prototype, we decided to convert a Fortran program to calculate converted wave fold distribution (3C-3D Design) into an interactive document. This program was originally written as a stand-alone DOS application (Lawton, 1993). In its interactive book form, it is an integrated document containing the paper describing the theory, the original program's user documentation, a user interface (parameter entry form), the fold-calculation program itself, code to render the output, and all of the source code.

The document is broken down into four modules (see figure 4).

- hypertext documents containing the theory, user manual, parameter entry form, and source code.
- the 3C-3D Design program, which takes a parameter file as input and outputs a data file to be rendered.
- the foldmap rendering program, which takes the 3C-3D Design output data file and renders it into a graphic image.

- the control program, which takes the users parameters, generates a 3C-3D Design parameter file, executes 3C-3D Design, calls the rendering program, and recreates the HTML document containing the rendered image and the parameter entry form.

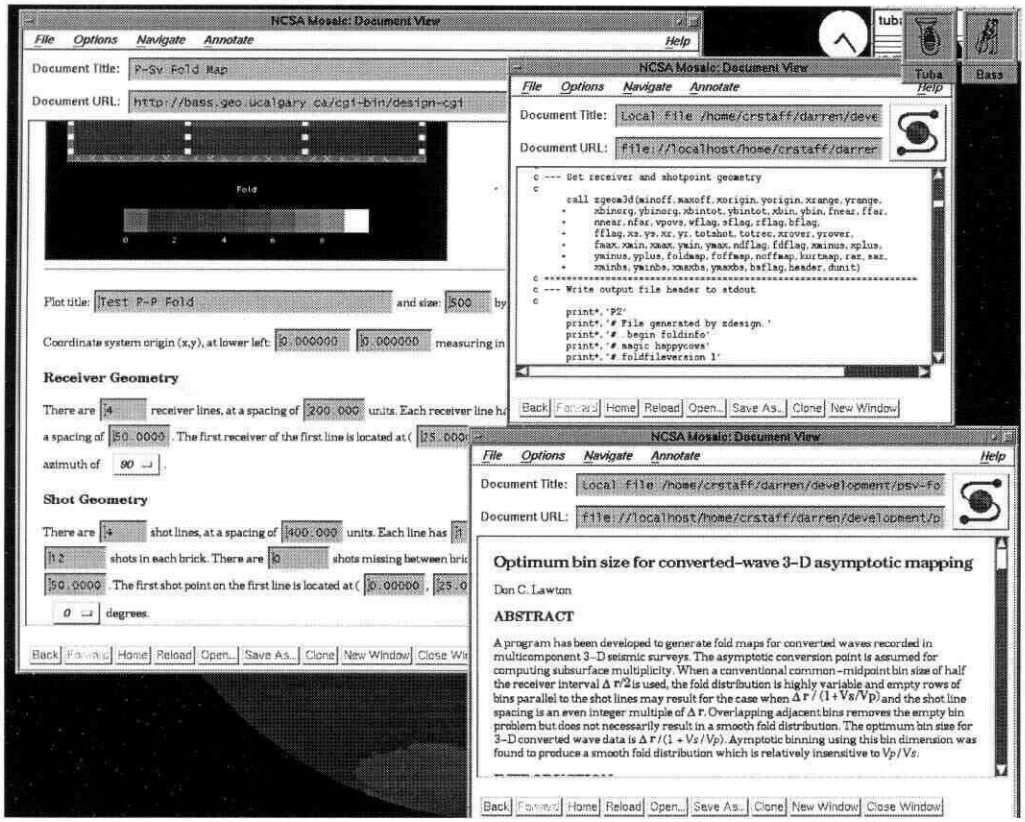

FIG. 5. Screen shot of Mosaic views of the 3C-3D Design interactive document. Note that there are three windows displaying the paper, an interactive figure, and the source code.

Future work

The first interactive document has been produced, and the technique shown to be feasible. We now need to gauge the benefits to the end user in terms of increased ease of understanding and speed of learning. If these benefits are substantial, then other, larger projects may be pursued, with perhaps the ultimate goal being the production of the entire CREWES Research Report in an interactive format. Other technical problems need further investigation, including the integration of commercial software packages and large seismic data sets into this interactive environment.

An interactive approach to seismic methods

We are currently studying the feasibility of putting together a fully interactive document covering most of modern seismology. This material covered would include rock properties, numerical and physical modelling, borehole methods, 2-D, 3-D, and 3-C surface seismic, and case histories. Integration of commercial software and databases will be an important part of this document. We believe that interactive documents will play a increasingly important role in both education and publishing, especially in rapidly changing technical fields.

CONCLUSION

The advantages of a hypertext document information structure have been described. CREWES has contributed to both the content and the maintenance of the global multimedia hypertext system known as the World Wide Web. The technologies behind the Web have been used to implement a networked interactive hypermedia document.

GLOSSARY

Document - a collection of organized information.

Electronic document - a document that is stored and viewed on a computer.

Hypertext document - an electronic document that contains embedded links to other documents.

Multimedia document - an electronic document containing forms of information other than words and pictures, for example video and audio.

Hypermedia - Multimedia hypertext. A multimedia document where the hypertext links can be text or graphics elements.

Interactive document - an electronic document that may be modified by the reader.

Network information server - A program on the net that provides access to data or document archives.

HTTP - HyperText Transfer Protocol - the communications protocol used on the Internet to transfer Web documents from server to client.

HTML - HyperText Markup Language - The page description language used to write hypertext documents on the Web.

REFERENCES

Berners-Lee, Tim, 1994, The World Wide Web Summary, http:// info.cern.ch/ hypertext/ WWW/ Summary.html

Gray, Matthew, 1994, Wow, The Web is Big, http://www.mit.edu:8001/ afs/ sipb/ user/ mkgray/ ht/ wow-its-big.html

Lawton, D.C., 1993, Optimum bin size for converted-wave 3-D asymptotic mapping: CREWES Project, Annual Research Report, Volume 5, 28.1-28.16