Becoming An Information Provider On The World Wide Web

Brian Kelly <B.Kelly@leeds.ac.uk>

Abstract

This paper describes the tools which are being evaluated at the University of Leeds for use by information providers on the World Wide Web. The paper also gives a introduction to the World Wide Web's client/server architecture and the Hypertext Markup Language (HTML). Information is provided on further sources of information which will assist information providers and trainers of information providers.

The paper is intended for new information providers on the World Wide Web and for people who are involved in their training.

I. Introduction

The University of Leeds is one of the largest universities in the UK, with a student population well in excess of 17,000. The University Computing Service provides IT facilities for a large user community, with over 20,000 registered users on centrally managed systems. The majority of users make use of an extensive Novell PC network. In addition a powerful Unix server and over 150 centrally-managed Unix workstations are also available.

The University of Leeds was one of the first UK universities to become involved in the World Wide Web (referred to as WWW in this paper). By May 1993, when the University held an Open Day, the departments of Chemistry, Computer Studies, the Computer Based Learning Unit, University Administration and the Computing Service were providing information services on WWW.

The first information providers at Leeds were experienced computer users. They included academic and technical staff who had responsibilities for the system's management of departmental Unix systems. In order to set-up their information service they had to install the server software as well as WWW client software and associated viewers.

By the end of 1993 increasing use was being made of WWW. Its importance was apparent from its mention in the University's IT Strategy Document [1]. In November 1993 the Computing Service published a special edition of its Newsletter [2] which had the theme of online information services in general and WWW in particular. The Newsletter attracted interest from a number of departments who wished to publish on WWW.

II. Promoting WWW

The key to the development of a successful WWW service is motivation of the information providers who will be responsible for the ongoing maintenance of the information service.

The Computing Service has been involved in promoting and marketing WWW since May 1993. The Computing Service's contribution to the University's Open Day included:

- A poster display illustrating the uses of WWW by departments in the University.
- Access to the Open Day programme on WWW.

The poster display has subsequently been used at a number of conferences. An electronic version of the poster display is available on WWW at the URL `http://www.leeds.ac.uk/ucs/posters/WWW_display.html`.

III. WWW Client/Server Architecture

Steve Jones in his opening presentation "*An Overview of CWISs, What The Workshop Is About, What The Future May Hold*" [3] summarised a debate held on the UNITE Mailbase list by saying "*there are a great many people (the majority) around who still use plain ASCII terminals (or emulators) for their work.*". He concluded by saying "*If any conclusion can be drawn from these examples, it is that we must not let ourselves be overtaken by the new tools ("toys" if you like) to the detriment of our real task. ... Yes, it is interesting to play with X-Mosaic (if we have the technology) ... "*

These sentiments illustrate a misunderstanding of the client/server architecture of WWW. Information providers need to be aware that X-Mosaic (or NCSA Mosaic For X as the developers at NCSA prefer it to be called) is just one of a range of client browsers which can be used to access WWW.

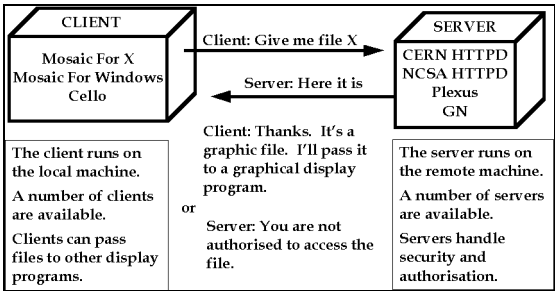

Figure 1 WWW Client/Server Architecture.

Figure 1 illustrates WWW's client/server architecture. WWW client software, such as Lynx (shown in Figure 2), enables text information held on WWW to be accessed using vt100 terminals.

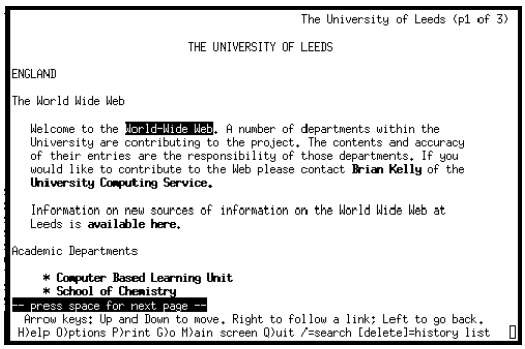

Figure 2 The Lynx Browser.

Graphical browsers such as Cello and the NCSA Mosaic browsers for X (shown in Figure 3), Microsoft Windows and Apple Macintosh provide access to graphical information, as well as displaying text in a variety of fonts, styles and sizes.

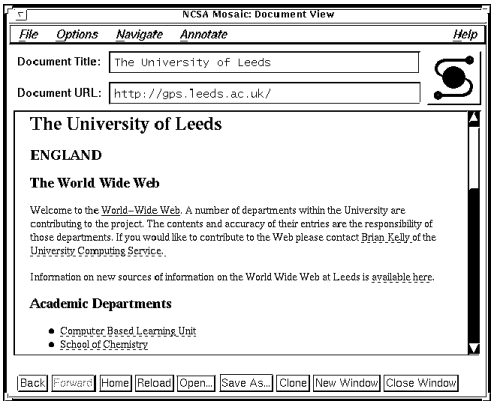

Figure 3 The NCSA Mosaic For X Browser.

It is important that potential information providers who are currently still using command-line systems are not discouraged from using WWW because of a mistaken belief that it is only accessible from powerful workstations which are capable of running graphical user interfaces.

IV. HTML

Native documents on WWW are written in the Hypertext Markup Language (HTML). An example of an HTML document is shown in Figure 4.

```
<TITLE>The World Wide Web</TITLE>
<H1>About The World Wide Web</H1>
<P>The World Wide Web is a
<EM>distributed multimedia
hypertext</EM> system.</P>
```

Figure 4 A Simple HTML Document.

Structural elements in the document are identified by start tags and end tags. In Figure 4 the `<TITLE>` tag is used to give the title of the document. The `<TITLE>` is normally displayed by the browser software. The `<H1>` tag defines the first level heading, which is normally displayed differently from the body text. The `<EM>` tag is used to emphasise text.

Although many browsers will handle the document shown in Figure 4 correctly tags such as `<HTML>`, `<HEAD>` and `<BODY>` should be included in document. Use of these tags will enable indexing and auditing software (known as *spiders* or *robots*) to work more efficiently.

```
<HTML>
<HEAD>
<TITLE>The World Wide Web</TITLE>
</HEAD>
<BODY>
<H1>About The World Wide Web</H1>
<P>Information about WWW is available
<A HREF=http://info.cern.ch/hypertext/
WWW/TheProject.html> at CERN</A>.</P>
</BODY>
</HTML>
```

Figure 5 A Simple HTML Document.

Figure 5 illustrates the use of these tags, together with the anchor tag `<A>` and `</A>`. This is used to create the hypertext links to other files. In this example the phrase `at CERN` will be highlighted. Selecting this highlighted phrase will cause the client to request the file `TheProject.html` from the directory `/hypertext/WWW` using the `http` protocol from the server at `info.cern.ch` .

V. Authoring Tools

The first generation of information providers at Leeds created HTML documents by hand (by using a text editor such as vi or Emacs) or by developing macros in their preferred editor or word processor to convert documents to HTML format.

The next generation of information providers will expect simple, easy-to-use authoring tools, typically with a graphical user interface. Authoring tools are becoming available, some of which are extensions to word processing packages while others are self-contained applications.

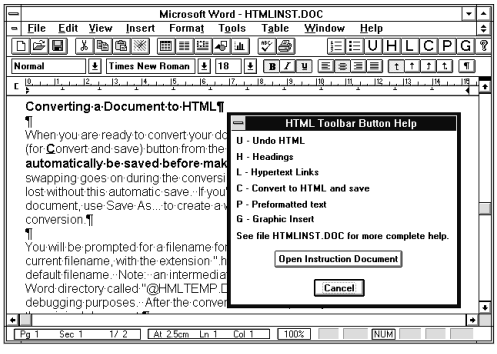

Figure 6 Word For Windows Authoring Tool.

Authoring tools which are currently available for Word For Windows include:

- Georgia Technical Research Institute GT_HTML.DOT template [4] which is illustrated in Figure 6.
- Chinese University of Hong Kong CU_HTML.DOT template [5].
- The HTML Edit tools [6].

Other authoring tools available for Microsoft Windows include:

- The HTML HyperEdit Toolbook application [7].
- The HTML Assistant application [8] which is illustrated in Figure 7.

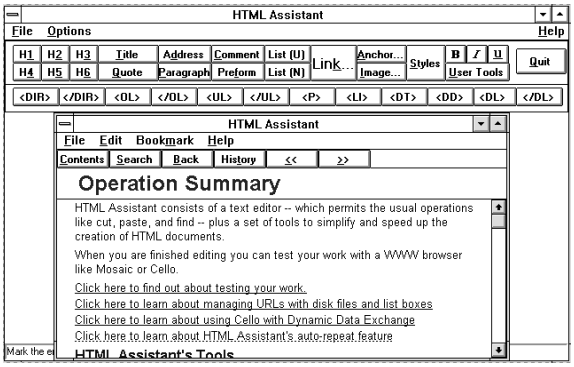

Figure 7 HTML Assistant Authoring Tool.

At the time of writing most of the authoring tools are at an early stage of development. It is too early to identify which is the best tool (or indeed tools, if a number of authoring tools will be required to suit the requirements of a range of information providers, with different backgrounds and expertise). However it is clear that criticisms regarding the difficulties in creating HTML documents are no longer true.

V.A. Quality Of Authoring Tools

Information providers who make use of authoring tools need to be aware of the ongoing support for the tools and the quality of the HTML produced by them. A mail message to the author may help in discovering if the author intends to provide long-term support for the software. Validating the quality of the output from the program may prove more difficult, especially given the lack of easy-to-use validation tools.

VI. Document Conversion Tools

The authoring tools mentioned in the previous section are suited for creating new HTML documents. Many information providers will wish to convert existing documents into HTML format. Although a number of the authoring tools can import existing documents, they normally require structural elements in the document (such as headings) to be marked up within the authoring tool. This can be time-consuming.

Document conversion tools will, however, normally make use of existing markup in the native word processing format - such as converting the `Heading 1` style in a Word For Windows document to the `<H1>` and `</H1>` HTML tags.

VI.A. RTFTOHTML

The Computing Service has converted over 100 of its documents using Chris Hector's RTFTOHTML conversion program [9]. The software can convert styles used in word processing packages such as Word for Windows and Wordperfect, to HTML tags. The conversion table which maps styles to HTML tags can be configured to reflect the styles used locally. In addition styles can be flagged to be ignored, which enables parts of a word processed document to contain text which will be available in the printed version of the document, but omitted in the HTML version. RTFTOHTML also generates the appropriate markup for images in the original document.

VI.B. LaTeXtohtml

Nikos Drakos of the Computer Based Learning Unit in the University of Leeds developed the LaTeXtohtml conversion program [10]. This conversion program is widely used around the world, providing a range of useful features, including automatically handling inline images and equations, generation of navigational aids and footnotes and tables of contents.

VI.C. Rainbow

Electronic Book Technologies have developed an SGML document type known as Rainbow which is designed to represent documents which have been created using word processor and desktop publishing software [11]. Rainbow-Maker software is freely available by anonymous FTP from `ftp.ebt.com` in the directory `/pub/nv/dtd/rainbow`. Implementations of Rainbow-Maker can convert documents in Word, Wordperfect, Frame and Interleaf formats.

Software to convert documents from the Rainbow DTD (Document Type Definition) to the HTML DTD is not currently available. This will, no doubt, change in the near future. Rainbow may be the most appropriate tool for converting documents which have been created using a desktop publishing package to HTML format.

VII. Graphical Tools

HTML documents can contain images in GIF format which can be displayed using a suitable browser as part of the document. In addition, images held in JPEG format can be viewed using an external viewer. A number of tools are available for converting graphical images to GIF and JPEG formats. The xv viewer program can be used to convert image formats. It is illustrated in Figure 8.

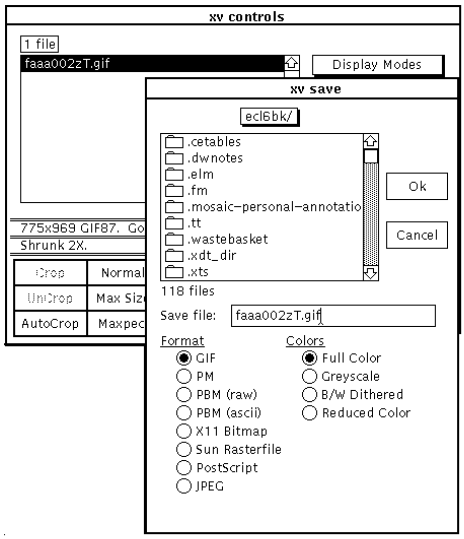

Figure 8 Converting Graphical Images.

As well as converting files to the appropriate format, conversion software should also be used to reduce the colour depth of a graphical image to an appropriate level: a simple line drawing, for example, should have a colour depth of 2.

VIII. File Store Structure

For reasons of security the httpd (hypertext transport protocol daemon) should be restricted to a limited area of the file store on the WWW server. In Figure 9 httpd can only access files which are located (or linked) beneath the directory `/apps/WWW`.

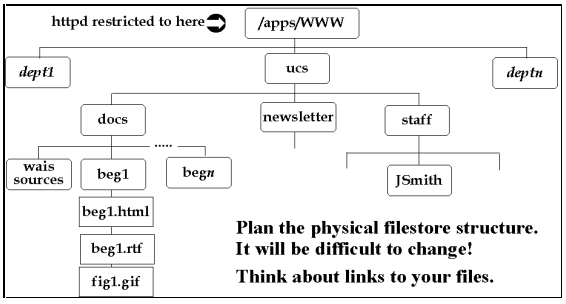

Figure 9 File Store Structure.

The structure of the file store should be carefully planned, so that the WWW service can grow in a systematic manner. Figure 9 illustrates part of the file store structure used by the Computing Service.

The directory `docs` contains Computing Service documentation. A separate sub-directory, which has the name of the code of the document (e.g. directory `beg1` contains the files associated with the *Getting Started On The Novell PC Network* [BEG 1]) is provided for each document. The directory will contain the HTML file, associated images, together with Postscript and RTF versions of the files. A subdirectory called `waissources` contains the files created by the WAIS indexing software.

IX. WAIS

Searching for information held in files on WWW can be achieved using WAIS (Wide Area Information Server). The Computing Service uses the freeWAIS software to index its documentation and Newsletters (the freeWAIS software is available by anonymous FTP from `ftp.cnidr.org` in the directory `/pub/NIDR.tools`).

Once the files have been indexed, a link is created in the directory used by the waisserver software. A variety of clients can be used to access the WAIS database. Later versions of the NCSA Mosaic browsers provide support for WAIS. In addition dedicated WAIS clients can also be used.

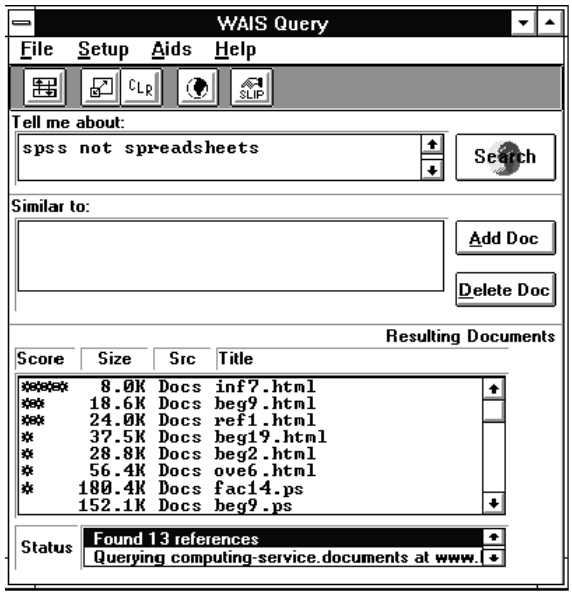

Figure 10 Accessing A WAIS Database.

The Computing Service has indexed its documents and newsletters. Figure 10 illustrates how a WAIS client can be used to search for documents containing the string of characters *spss* but not *spreadsheets*.

The WAIS sources for the documents and newsletter databases are stored in sub-directories beneath the documents and newsletter directory. Links to these sub-directories are made from the

directory used by the WAIS server software. Since additional WAIS databases must have links to the same directory it is important that a consistent naming convention is used for WAIS directories. For example, since many departments produce documents, the Computing Service has called its WAIS database *computing-service-documents*.

X. Futher Sources Of Information

Information providers should read the latest information about WWW on the Web itself. The Usenet News groups `comp.infosystems.www` and `comp.infosystems.wais` provide a means of discussing topics related to WWW and WAIS.

WWW developers may wish to subscribe to the `www-talk` list by sending the message **add www-talk** *name* to the email address `listserv@info.cern.ch`.

The author has produced a series of slides which may assist trainers of information providers. These are available in Postscript format at the URL http://www.leeds.ac.uk/ucs/people/BKelly/aberdeen_ohps.ps.

References

[1] University IT Strategy Internal University of Leeds document.

[2] Computing Service Newsletter, University of Leeds Issue 209 November 1993.

[3] Proceeding Of A Joint Workshop On CWIS Development And Management, University of Newcastle, April 15-16th 1993.

[4] Posting by Jeffery L. Grover in the comp.infosystems.www Usenet group, 21 October 1993. See also the documentation which is available at the URL http://www.gatech.edu/word_html/release.htm.

[5] Posting by Anton Lam in the comp.infosystems.www Usenet group, 11 March 1994. See also the documentation at the URL http://www.cuhk.hk/csc/cu_html/cu_html.htm.

[6] Posting by Jorma Hartikka in the comp.infosystems.www Usenet group, 21 March 1994.

[7] Posting by Jonathon Magid in the comp.infosystems.www Usenet group, 30 December 1993. See also the documentation which is available at the URL ftp://ftp.info.curtin.edu/pub/internet/mswindows/hyperedit.

[8] Further information is available at the URL ftp://ftp.cs.dal.edu/htmlasst.

[9] Chris Hector, Cray Research. Further information available at the URL ftp://ftp.cray.com/src/WWWstuff/RTF/rtftohtml_overview.html.

[10] Nikos Drakos, "A LaTex to HTML Translator", University of Leeds. Available at the URL http://cbl.leeds.ac.uk/nikos/tex2html/doc/latex2html/latex2html.html.

[11] Message from Marc Andreessen to the www-talk mailing list, 10 December 1993.

Author Information

Brian Kelly is the Principal Information Officer in the Computing Service, University of Leeds. He received a B.Sc. in Civil Engineering at the University of Leeds in 1977. After working in the Department of Naval Architecture in the University of Newcastle he began his career in computing as an Analyst/Programmer at Sunderland Borough Council, followed by Humberside County Council. In 1984 he became a Systems Programmer at Loughborough University of Technology. In January 1990 he moved to Liverpool University, working as the Information Officer in the Computer Laboratory. Since May 1991 he has been employed as the Information Officer in the Computing Service at the University of Leeds.

Brian Kelly first became interested in the World Wide Web following a demonstration at Leeds University organised by the University's Information Exchange Special Interest Group in December 1992. He was responsible for installing the programme for the University's Open Day (held on 8th May 1993) on the World Wide Web and for producing a poster display on the University's involvment with the World Wide Web. Since then he has given presentations about the World Wide Web at workshops at Manchester Metropolitan University, University of Kent, University of Oxford and Aberdeen University.