Digital's World-Wide Web Server: A Case Study

First International Conference on the World-Wide Web
Geneva, Switzerland
May 1994

Russ Jones
Internet Program Manager
Corporate Communications
Digital Equipment Corporation
rjones@pa.dec.com

Digital Equipment Corporation was one of the first large corporations to embrace the World-Wide Web as a basis for customer support, global electronic marketing, online interactive product demonstrations, and electronic commerce. This paper uses Digital as a case study to help other organizations that wish to participate in this evolving electronic community.

This paper will address the organization, structure, evolution, and management of the information contained in the Web server. The original design goals of the Web server will be mapped against the actual experiences gained from the week-to-week management of the environment. The paper will also identify common pitfalls to be avoided when deploying a large, commercial-class World-Wide Web server.

Introduction

Corporations do not exist in a vacuum. They interact with customers, prospects and partners on everything from joint development to business transactions. First done historically face-to-face, then by hand script, then by telegraph, then by telephone -- it is now being done electronically using computers. And it is being done with the building blocks of the electronic communications age -- words, pictures and sounds.

Just as the telegraph or the car extended the radius of the marketplace served by a business, electronic communications are today extending the market reach of organizations on a global scale. The World-Wide Web is the enabler. It provides an open, vendor-neutral environment for organizations, big and small, to participate in the world of global electronic communications.

Built on top of the Internet, the World-Wide Web offers most of the benefits of the so-called "information superhighway" -- but offers them today with protocols, client/server technology, and a global telecommunications infrastructure that is already in place.

The World-Wide Web presents the first tangible example of the world of the future. A world that is just a mouse-click away for information consumers. A world offering low barriers-to-entry for information providers.

Growing rapidly, the World-Wide Web provides the mechanisms for corporations to establish an Internet *presence*. This paper examines just how Digital Equipment Corporation established its Internet presence via the World-Wide Web. It examines the initial goals of the program, how the goals effected the Web server design, and then describes the lessons that were learned. It also what we would have, and in some cases already have, done differently since the launch of Digital's World-Wide Web server.

World-Wide Web Background

The introduction of the World-Wide Web into Digital Equipment Corporation occurred on May 1st, 1993 when Mosaic was brought into the research labs for the first time. Although well-versed in the concepts of hypermedia information systems, Digital was for the first time able to readily see what the World-Wide Web architecture brought to the Internet in terms of information access:

For information consumers:

- A simple point-and-click interface that would provide wide-area information access to non-technical users.
- A simple, unified naming scheme (URL) that would provide transparent access to other Internet information resources (Newsgroups, Telnet, FTP, Gopher, etc.).
- The increased usability associated with hypertext and the ability to follow threads of interest across departmental, organizational and national boundaries.

For information providers:

- The ability to integrate graphics and pictures with richly-formatted textual information.
- Customizable gateways that could easily provide backwards compatibility to legacy information systems (Notes, RDBMS, etc.).
- The ability to focus on Web server development while leveraging ongoing public-domain Web browser development being done elsewhere.

For Digital, the crowning benefit of the World-Wide Web architecture was twofold; the breadth of freely-available Web tools (browsers, servers, etc.) and the diversity of platform support they offered.

The World-Wide Web was recognized as a true multivendor environment that embraced all flavors of UNIX, Macintosh, Microsoft Windows and OpenVMS. This is exactly the environment found inside of Digital and the environment that most of our customers are dealing with every day.

World-Wide Web Server Goals

While the proliferation of Web servers inside of Digital was immediate, with over 60 Web servers running on our internal TCP/IP network with 40,000+ nodes, attention was placed on development plans for an external Web server that would be accessible for our customers, prospects and partners.

Time to market

There was a real sense inside Digital that a global electronic community was being built on top of the World-Wide Web.

Most of the initial Web servers, although excellent, were found in the university space and in the research labs. There was no other major corporation from the business community yet online.

We knew from working closely with the Web community that other organizations (many of them our existing customers and partners) were going through the same internal digestion process with the World-Wide Web. They knew it was exciting, they knew it had implications, but they were looking for others to lead the way.

We saw a clear opportunity to move quickly to extend the information services that we were already providing to our customers on the Internet. The sooner we got online, the sooner we would be able to help customers wanting to do the same thing.

Ease of use

The World-Wide Web provided the architectural underpinnings to do very innovative design work and advance the state-of-the-market in electronic customer support and electronic marketing. The hypermedia characteristics of the World-Wide Web provided the perfect environment to deliver quick and direct access to information, as well as provide alternative views, or navigation routes, into the same base of core information.

Time-to-Market Impact On Design Issues

To meet the time-to-market requirements it was decided early on that the initial Web server would be a facade, or umbrella, across a well structured FTP archive. Digital had spend the previous year designing and populating an extensive collection of product, service and technical information for FTP access.

The initial thoughts were to focus first on providing friendly, hyperized access access to the existing documents in this FTP Archive, with subsequent focus on conversion to HTML for full hypertext interaction.

The groundwork for this approach was already in place. We had established strict guidelines for information placed in Digital FTP archive with regards to directory indexes, file names, data formats, and most importantly, document abstracts.

The FTP archive was divided into directories according to how Digital classifies information. For example, technical overviews, performance reports, whitepapers, infosheets, technology backgrounders, etc. Each document was available as an ASCII text file and/or as a PostScript file. If the PostScript file was available, it was also made available in multiple compressed file formats using the UNIX-standard *compress* utility and a PK-ZIP compatible compression utility. These alternative PostScript versions were intended to give information consumers the ability to more readily retrieve PostScript documentation to their local system with the information already in the compressed form of their choice.

Each logical document also had a companion abstract that described the formal document name, the document publication date, the document length (in printed pages), and a paragraph summary of the document.

For example, a whitepaper called *Understanding RAID Technology* might be represented in the FTP archive as:

```
/whitepaper/raid-technology.abs
/whitepaper/raid-technology.txt
/whitepaper/raid-technology.ps
/whitepaper/raid-technology.ps.Z
/whitepaper/raid-technology.zip
```

These five files, although consistent and well indexed, just represented a single document. Taken as a whole, Digital's FTP archive was a collection of 3000+ logical documents spread across 12 directories.

All of the initial design ideas around the World-Wide Web server were based on the assumption that Digital's FTP archive, as a library of documents, was already in place, accessible, and clean for external Internet consumption with no copyright, ownership or privacy issues.

Representing the Document with HTML

Using the previous logical document as an example, it was decided that the way to represent a logical document was in a more meaningful citation form:

Understanding Raid Technology
Whitepaper -- Abstract, Text, PostScript
July 1993, 8 Pages

In the first and third lines the document name, publication date and length could be captured from the .abs file. The second line represents the document type, with hyperlinks to the underlying available data formats.

Using this document presentation technique, users could click to view the ASCII text version directly or could retrieve the PostScript version with the Web browser and have it feed directly into a PostScript previewer. It was decided that when nothing but a PostScript version of the document was available, the citation should still have the hypertext link labeled *PostScript* so the user could make retrieval decisions based on document type.

Of course, most Web browsers also allow users to retrieve a copy of the file to the local disk for subsequent printing if needed. Users could also click to look at the abstract before viewing the actual document.

Furthermore a series of these document citations could be arranged vertically on a HTML page to represent the availability of documents according to any number of different index, query and sort criteria.

Design Issues For Ease of Use

One of the powerful capabilities of a hypermedia information environment such as the World-Wide Web is the ability to present multiple views into the same logical set of documents. The FTP archive was organized primarily by how Digital views information internally with topics organized by document category (or document type) and content spread across multiple directories. Obviously the multiple entry-points into a World-Wide Web server would facilitate multiple faces or orientations on the same set of information.

The initial design called for the representation of five different views into the same core set of documents. The different views would provide users with the ability to navigate by:

- Alphabetical document title

  This would give users the ability to go to a specific document by first clicking on the leading letter of the document title and then browsing an alphabetical list of document citations. It would be independent of the underlying directory and document structure.

- Document type

  This would give us the opportunity to develop a series of graphical icons that would correspond to the different document types (technical overviews, performance reports, etc.). These could be used to add an intuitive feel to the Web server for users browsing by document type.

- Document subject

  This would provide users with an orientation that would view the Web server by terminology and information categories that were more representative of industry terms.

  For example, if someone wanted to learn about multimedia, they could go to the multimedia subject index and browse through document citations without having to know beforehand the names of Digital's multimedia products. It would be independent of the underlying directory and document structure.

- Chronological order

  This would actually be reverse chronological order so that users familiar with the Web server could see what was new on a day-by-day basis. It would be independent of the underlying directory and document structure.

- Random word search

  This would facilitate the search for documents when the document title, document subject area, or document type were not obvious or known to the user. The goal was to have the ability to search for random documents within the constraints of the major document-type categories, but also be able to search for documents across document-type boundaries.

This cross-server searching capability, in particular, was viewed as one of the key potential strengths of the Web server. It would give users the ability to search for a technology keyword such as *DCE* (Distributed Computing Environment) and get back a document list that would span the technical software description for all DCE products, DCE part numbers and pricing, historical DCE-related announcements, or any whitepapers on the Web server related to DCE technology.

World-Wide Web Design Issues

In addition to these major design goals we knew there were other design questions that would have to be answered:

- What balance should be struck between the use of graphics and text?
- How extensively should graphical navigation aids such as icons be used?
- What type of consistent look-and-feel, or style guideline, should be developed?

The Use of Graphics

Certainly one of the key attractions of the World-Wide Web environment is the ability to integrate graphics and text. The use of supporting graphics (diagrams, sketches, photos, etc.) can add clarity to concepts that might otherwise be hard to grasp with words only. However, there is a price paid by both the information consumer and the information provider. The information consumer experiences the frustration of waiting for the graphic to be retrieved and delivered over the Internet. The information provider risks alienating the user from further use of the server if graphics are used too heavily.

It was decided that the use of graphics would fit very well with product photos and diagrams where clear value could be derived from the graphic. As a corollary to the graphics issue, it was decided that the use of ISMAP (the functionality that allows hyperlinks to be embedded in graphics) would also be appropriate if the graphic had meaningful content.

It was also decided that the Digital logo should be used at the top of all of the Web server to add Digital's identity to each page and to clarify for users randomly walking the World-Wide Web that they had entered into Digital's Web server.

It was felt that over time the home page as the main entry point into the Web server would give way to multiple supported entry points. Furthermore, we knew that downstream we would be encouraging other partners, customers and electronic magazines to embed hyperlinks to the hypertext pages of their Web servers. We wanted to assure that given the Webs ability to span organizations our identity was apparent at every server entry point (i.e. every page.)

Graphical Navigation Aids

Icon size and style was evaluated to determine the feasibility of icons as navigation aids. The initial icons were 1" by 2" in size, resembled the 3D Motif buttons, and included an assortment of international symbols. They also contained descriptive text phrases that added further clarity to the icon.

These were quickly discarded as the load time for a page of 12 icons (organized three columns wide and four rows deep) was barely acceptable for local Web browsers hitting against local Web servers at Ethernet speeds.

The icons were then downsized considerably and the descriptive text phrases were removed. This resulted in an 85% reduction in the number of bytes that needed to be retrieved to represent a page of navigational icons. The descriptive text phrase were then added back into the page as HTML text. This accomplished two things. It made the display of the icons acceptable and it forced us to think through the use of descriptive text versus icons for navigation. The icon-only approach has the added benefit of better addressing the global World-Wide Web user base, but penalizes users with non-graphical Web browsers.

In the end, the graphic on the icons were created using Adobe Illustrator, and rendered in Adobe PhotoShop using antialiasing. The icon background color was matched with the color of the Digital logo at the top of each page to add visual unity to the overall Web server.

Consistent Look-and-Feel

Visual look-and-feel consistency was deemed to be extremely important, as it would add usability to the Web server and also help to project a consistent image of the information service and Digital as a company. It was decided that there would be a minimal style for each page of HTML (regardless of purpose) and several additional styles for major areas of the Web server based on function..

The minimal HTML Style Guideline that was established outlined several key page components that were to be carried over onto each page. Drawing on the Web etiquette and style guidelines developed at CERN, the Digital HTML style includes from top to bottom:

- A meaningful document title that could stand alone in a hotlist or as a bookmark.
- A consistently sized Digital logo
- A Level 1 or Level 2 Header that matched the document title as closely as made sense
- An optional subheading in italics
- Two horizontal rules to break the HTML page into a logical top, a middle section with content, and a logical bottom
- Optional navigation and/or search hyperlinks
- A page creation date
- The initials of the document author

The two horizontal rules, used to frame the document content, would create the visual metaphor of a page and separate the document control information from the content.

In addition to this Digital HTML Style Guideline, major look-and-feel designs were established for different types of logical documents. In particular, incremental style guidelines were developed for:

- Navigational indexes
- Product information sheets
- Electronic Newsletters

A consistent navigation metaphor was also establish for the overall Web server. This navigation metaphor would provide users, regardless of where they were in the Web server, with the ability to return to the server home page, follow any hyperlinks that may be on the current page, backup using either the *back* command or through *window history*, or go to one of three different search engines.

The search engines represented a consistent capability to navigate on a tangent based upon random searches against all document's content, all document citations, or all HTML document titles.

Web Server Implementation

Once the major design decisions had be made, implementation of the Web server was fairly straight-forward. The initial time-to-market constraints dictated that the majority of the initial Web server was simply an easy-to-use facade on top of the FTP archive. This, combined with the state of HTML authoring tools in the summer of 1993, drove an implementation plan that would rely heavily on automatic generation of HTML.

Auto-generation of indexes

A *perl* script was developed that would traverse the directories of the FTP archive and build and in-memory model of all the logical documents on the FTP archive. The in-memory model would contain:

- Document title (from the abstract)
- Document type (from the directory name)
- Available document forms (from those available)
- Document creation date (from the abstract)
- Document age (file modification date)
- Number of pages
- Associated keywords

Essentially this would be all of the information needed to repeatedly walk the model and automatically generate multiple views of the logical documents cut by document categories, title, subject mater (using associated keywords), and document chronology.

From this very basic model a precedent was established that the Web server would be regenerated on a nightly basis as a reflection of the current contents of FTP archive.

txt2html

The next step was to develop a simple ASCII text to HTML (*txt2html)* converter that was specific to the format of many of the ASCII text documents in the FTP archive, and in particular, specific to the format of the document abstracts in the FTP archive.

This converter would generate HTML that complied with Digital's HTML style guidelines (inserting the Digital logo in the appropriate place.) It would also automatically identify headers, bulleted list, sublist, and pre-formatted tables.

Once implemented, this converter was used to convert simple ASCII text to HTML with 90% accuracy on random ASCII text documents, but was honed to do 100% correct ASCII text to HTML conversion for logical document abstracts found in the FTP archive.

Auto-hyperizing Document Abstracts

The *txt2html* converter was then coupled into the Web regeneration process, such that prior to regeneration of the logical document citations, all of the logical document abstracts are converted to HTML. In addition the the conversion of the abstract content, the HTML abstracts were also embellished with hyperlinks that gives users the ability to retrieve and review the abstract, and if interested, directly link to the underlying documents in either ASCII text or PostScript format.

Manual vs Automatic Hyperization

Many ASCII text documents proved much too complicated for the *txt2html* converter, but had obvious hyperlinking opportunities. This was partially a reflection of the natural order of the information found in Digital's server, and partly a result of specific reference in many documents to other, more detailed, documents.

The information food chain found on the FTP archive was such that if you examined any one subject area you could follow an obvious path to more detailed information. For example, a software product is described in a press release (with three paragraphs of detail), described in an information sheet (with three pages of detail), and then finally described in a formal, legal product description (with 20 pages of detail.)

This conceptual relationship would prove beneficial in establishing manual hyperlinks. This relationship was also exploited in the area of electronic newsletters. Digital distributes many newsletters to customers and prospects over the Internet. These newsletters typically take the form of a main newsletter body (with the highlights) and supporting documents with more information. These were obvious areas to manually establish hyperlinks. As mentioned earlier, an electronic newsletter HTML style was established as part of the style guidelines to add consistency and ease-of-use for users that would regularly be browsing HTML-formatted newsletters.

Also relevant were specific pointers to other documents by part number or document name. These were natural opportunities where hyperlinks could be automatically established between documents. For example, software product description, or SPD's, are standardized with a 12.34.56 numbering scheme. These SPD number can easily be recognized in and hyperlinks automatically established.

Pseudo-Hyperization

Based on this notion of manual vs automatic hyperization, Web regeneration was further extended to perform pseudo-hyperization of many ASCII text documents. Pseudo-hyperization was defined as the process of taking an ASCII text document, as is, and encapsulating the document with HTML:

- The top of the document would contain a valid document title, the Digital logo, and a Header level 1 title.
- The original ASCII text content would be wrapped with <PRE>...</PRE> HTML tagging phrase.
- The bottom of the document would contain standard navigation options and document time-stamp information.

This approach does not convert the original ASCII text document to HTML because of document complexity. Nor does it provide for a visually appealing display of the document content.

However, this approach does conform the original ASCII text document to Digital's HTML Style Guideline (which projects the Digital identity

when the user views the document), it does provides a clean and accurate document title (if a user decides to add the document to their hotlist or list of bookmarks), and it does provide the standard navigation control that was required throughout the Web server.

Another very important consideration was that even though the bulk of the document content was still pre-formatted as ASCII text, the document was no longer an ASCII text document. It was an HTML document. As such, it could be run through an automatic hyperization filter that would look for embedded references to other documents. If the referenced documents were available on the Web server, a hyperlink would automatically be established between the reference and the document.

Hand-Hyperization

Even with the use of the *txt2html* converter, it was felt there were still clear areas that warranted hand-hyperization. It was decided that information available in large quantities, volatile in nature, or is being continually updated did not provide the pay back that would be needed to justify the time-sink associated with manual hyperization.

However, information that was static and easy to convert to HTML via the *txt2html* converter could easily be justified. The conversion would only have to be done once, it would be 90% accurate, and any problems could be cleaned up while manual hyperlinks were added based on human insight into the overall breadth of information bound on both the FTP archive and the Web server.

Electronic newsletters fit into this category, as did announcements associated with new products or services. By hand-linking where appropriate, this approach would, over time, build a substantial base of hyperized documents.

As many users are hungry for information, or "what's new", this approach would also yield the highest pay back by at least hyperlinking the documents that are most often accessed. This, in turn, would further address the ease-of-use requirement.

The World-Wide Web As The Unifying Environment

Orthogonal to the development of Digital's Web server, other Internet information services had already been established by Digital and in many cases by customers. There were also other initiatives inside the company that would bring additional Internet services online in support of customers, prospects and partners.

Digital's user community had already established multiple newsgroups. Customers had already established private mailing list to resolve common system management problem. Digital had also recently added public-access Alpha AXP demonstration systems to the Internet for customer evaluation. Work was also under way to bring the Electronic Connection (which is Digital online order placement and order tracking service) onto the Internet via *telnet* access.

The Web provided the integration glue that would tie together not only simple point-and-click access to product and service information, but also provide users with visibility to support newsgroups they may have been unaware of. It would also be used to profile newsletters and describe the subscription process. It would also be used to highlight and navigate users to the online Alpha AXP demonstration system where they could kick the tires on actual products over the Internet, and then if they were satisfied, they could navigate to the Electronic Connection for ordering support.

Results And Lessons Learned

Digital's World-Wide Web server was officially launched on October 1, 1993. The HTML facade on top of the FTP archive was represented by 1500+ HTML files containing 32,000 links. It can be accessed using the following URL:

http://www.digital.com/home.html

It was registered with the NCSA *What's New* list and the CERN *World-Wide Web Servers* list. In addition, its availability was also announced into the main Digital newsgroups and followed up with articles in all of Digital's electronic newsletters.

Launching publicly five months after the World-Wide Web was first brought into the corporation, we were able to meet our main time-to-market goal. In the process we became the first major computer company to join the Web, and we hoped, sent a clear signal to the Web community that the corporate world understood the inherent benefits of the World-Wide Web and would embrace their vision of an electronic, hypermedia-based global village.

From direct user feedback we were able to judge our ease-of-use goal as largely met. Many provided insight on further services and refinements they would like to see, but most of the anecdotal feedback was positive.

The launch was also timed to coincide with the first issue of O'Reilly and Associates *Global Network Navigator (GNN),* the first magazine on the World-Wide Web. This innovative electronic magazine allowed us to sprinkle electronic ads across the magazine by the use of a small icon at the bottom of featured articles. By clicking on this Digital icon a reader could go to a Digital advertisement. From the advertisement, a reader could click and go to our World-Wide Web server. This type of advertising approach fit right in with the norms of the Internet -- non-intrusive, but available to those that want it.

Entry Points And Log Analysis

As we continued to promote the World-Wide Web server to more and more audiences, this provided our first lesson on things we should have structured differently. Although acknowledging all pages as potential entry-points into the Web server, we were essentially promoting a single home page as the entry point. When it came time to analyze our HTTPD server logs, we quickly realized that we could cut usage statistics every conceivable way except the one we cared about the most -- which World-Wide Web server the visitor linked from.

This was particularly annoying because we could not tell if a user had seen the home page URL in a newsgroup and just opened it, had clicked on it from the NCSA *What's New* page, or the CERN *World-Wide Web Servers* listing, or the GNN advertisement. GNN could tell us how many people read our advertisement on their Web server. We could tell how many people visited our home page. But we could not correlate the two. The lesson learned was that if you care about incoming access statistics by source, then independent, but identical, home pages are required to help map the correlation. Even this in not perfect and has its flaws, but it can be somewhat instructive to help determine general access trends.

FTP and HTTP Directory Mapping

Another area that should have been addressed up front was the mapping between the *ftp* and *http* server directories. We initially configured our *httpd* directory structure to lay on top of the same directory structure used by the *ftp* directory archive. In the initial design of the Web server we relied heavily on relative URL's to get to the non-HTML documents on the FTP archive. What started as a nice feature became more clouded when we started trying to make sense of *http* and *ftp* access logs to try to determine access tool trends.

It became unworkable when we decided to separate the *http* server and the *ftp* server onto two different systems so that we could better keep up with usage demands being made on both servers by an increasing number of users. At this point we backed up and decided that all URL's that specify *http* as the access method should be relative within the *http* directory structure, but URL's that linked to ASCII-text files and PostScript files on the FTP archive should be absolute URL's using the *file* access method.

CNAME

As a corollary problem to the downstream separation of the *ftp* and *http* servers, it became obvious that we had made an initial mistake by using the absolute host name of the Web server (gatekeeper.dec.com) and not establishing a CNAME for our Web server. Prior to separating the two servers onto separate systems we established *www.digital.com* as a CNAME and then had to go back to everyplace we had registered our Web server or knew we had customers pointing to our Web server.

Content Ownership

Another problem area perhaps less visible, but certainly no less important, is in the ownership of information that is available from a Web server.

People have been distributing textual information on the Internet for years and the techniques for identifying and declaring ownership, copyright and reproduction rights are well-accepted. Also it is fairly obvious that if an organization produces a textual document, and claims copyright on it, they have the right to place it on their own Web server.

In the complicated world of electronic photo ownership and copyrights, this is not always the case. The ownership of photographs that are used by corporations, even of their own products, are not always retained by the corporation. Photographers will charge for their work based upon who retains ownership of the photographs. If the photographer retains ownership, they will have sliding scale that might permits limited hardcopy usage (but not electronic usage) or perhaps limits usage on a country to country basis.

Electronic publishing technology makes the capture of photographs technically easy. The World-Wide Web makes the electronic distribution of photographs on a global basis easy. It is important for the organization that is deploying a Web server to make sure they actually ownership or permission rights for global, electronic-distribution to any photographs that are part of their Web server.

Summary

For companies that want to establish an initial electronic presence on the so-called, "information superhighway", the World-Wide Web offers the infrastructure, the tools, and the global reach to do just that, but offers it today.

The technology is powerful enough to deliver electronic catalogs with product descriptions, pictures, and ordering information -- all on demand -- to a user's screen.

The World-Wide Web does not demographically reach into the home yet with any significant numbers, but is growing wildly in the business community. If the consumers of your product or service are not found it the home, but are found in high rise buildings and corporate campuses around the in the world, then the infrastructure is in place.

Our approach to getting online quickly with an easy-to-use interface has largely succeeded. In the coming months we will be working to convert more of the underlying logical documents into a more visually-rich HTML format. We will also be working to increasing the breadth and depth of information and services we make available to customers, prospects and partners via the World-Wide Web.

Digital's entry into the World-Wide Web has proven successful. We are seeing the initial results in terms of more satisfied users, lower cost-of-sales, and increased revenue from customers through electronic channels.

Acknowledgments

I would like to thank the following people who either directly or indirectly played a role in the development of Digital's World-Wide Web server; Bob Heiney, Erik Goetze, Richard Schedler, Glenn Trewitt, Ralph James, and Lucia Springer.