The Internet, the Web, and the Chaos

Neville Holmes, University of Tasmania

The public and, I suspect, many computing professionals discern at best only a vague distinction between the Internet and the World Wide Web. In many popular writings, authors use the simple phrases "the net" and "the web" interchangeably, which is understandable as these terms are synonymous in ordinary speech. Admirably, the *International Herald Tribune* uses an initial capital in *the Net* and *the Web*, but professionals should go further and always refer to *the Internet*.

As an observer, I can dissect the digital networking world into usefully distinct major components, although the boundaries between them may be blurred. If we do not make such dissections, this world will seem hopelessly chaotic both to the profession and the public. The profession at large will not understand what we do, and the public at large will not properly appreciate it. The chaos will continue.

THE INTERNET

Today, the communicating world moves most of its digital data around via the Internet. Its creators designed the Internet, once simply a network connecting digital computers, to be resistant to disruption by all manner of disasters—including, for example, nuclear warfare.

The Internet has two major layers: the physical, in which the signals travel, and the logical, in which the data travels. In turn, each layer has two aspects: the nodes and the links between the nodes.

Our vocabulary hides the structure in the great digital "out there."

The physical layer

The physical Internet has many different kinds of links. Signals are transmitted electronically over wires, an increasingly unpopular medium, and photonically through fibers, cables, and space. Electronic mail in transit, or in storage, is not usually electronic, but rather photonic in transmission and magnetic in storage. Engineers use many different technologies and standards for signaling, and these technologies and standards come and go.

The physical Internet also has many different kinds of nodes. The mainframe computers that were nodes of the original Internet are now relatively few. Routers perform the buffering and forwarding of signals at intermediate nodes, while digital computers of various kinds dispatch and receive signals at terminal nodes.

Servers—straightforward digital computers—personal computers, handheld computers, and personal digital assistants now can all be Internet nodes. So also can the bewildering variety of mobile phones or cell phones. Further, we face the very real prospect that a huge assortment of automatic data-gathering equipment will be connected to the Internet to, for example, handle security surveillance and the identification and location of prospective customers.

Such a variety of technologies, standards, and equipment can seem chaotic. But professional engineers in the field will try to subdue this chaos and to make the Internet and its terminal nodes simple and reliable. Yet, one in every seven mobile phones sold in the UK last year was faulty (news.bbc.co.uk/1/hi/technology/4745205.stm), which shows that this objective is not always met.

The logical layer

Signals carry message data along Internet links in complex packages. The messages have a sequence of wrappings that carry transmission-management data for use at the nodes. One of these wrappings contains a numerical value that specifies the package's terminal node destination—its Internet Protocol address.

The Internet functions as a single entity because all data traveling across the network must be packaged with an IP address, and these addresses define the Internet's logical structure. The logical Internet is chaotic in two ways: addressing and packaging.

The size of the IP address puts a nat-

Continued on page 106

Continued from page 108

ural limit on the number of nodes that can be in the Internet at any time. Under the current IP version 4, that size is too small. The accepted version 6 would solve this problem, but its use is being smothered largely by commercial interests for short-term financial benefit, and by lack of international support. The result is that IP addresses are no longer fixed in their identification because of the tricks that must be employed to let everybody who wants to use the Internet do so. This in turn causes many further problems.

Developers intended the use of data packaging to make the Internet robust, and they succeeded. However, the Internet's success has led to a burgeoning of both the amount of data and the kinds of traffic on it.

In general, an increase in the amount of traffic can be handled simply by investment in the network's equipment. However, for the transmission of video data of one kind or another, it's not just a question of the transmission equipment's raw capacity, it's also a question of transmitting the huge amounts of data quickly and smoothly. There is reason to doubt whether the Internet can handle video data satisfactorily in packages.

THE WEB

The Internet, in the simplified way I have described it, is to the World Wide Web what an engine and transmission are to a car: The Internet makes the Web possible. It plays host to many different systems, such as e-mail and instant messaging, but these appear to be on their way to being subsumed within the Web ("Instant Messaging: A New Target for Hackers," *Computer*, 2005 July, pp. 20-23).

The original Web, which forms the foundation for the current Web, functioned as a logical network of links joining files. The files, stored at Internet nodes, held text encoded with the Hypertext Markup Language to specify the text's structure and nature. Uniform resource locators provided the links, which pointed to specific HTML files or to points within them.

The reasoning behind such links, as devised by Vannevar Bush—although his links had the advantage of being bidirectional—focused on finding details of specific ideas in a document if needed simply by following a chain of links, a generalization of formal documents' citations and bibliographies.

Search engines have been remarkably successful programs because they hide the chaos.

A URL has up to three parts that contain addresses: one for the Internet node at which the file is stored, one for the file's location within the node, and one for the HTML-labeled spot within the file where the specific details can be found.

The node address within the URL is not the IP address but a symbolic address that special Domain Name System servers scattered around the Internet can translate into an IP address. This node address consists of a sequence of names, vaguely and unsatisfactorily related to trademarks, which are used in marketing commercial products and warfare. URLs, often with just a node address, are now used as pointers to many different entities beyond mere text files.

Browsers

To access the Web, users need a browser, first to fetch and display a text file and any files associated with it, and second to follow a link within the current file to another file. The Web is useless without a browser, and only the success of various browsers has let it achieve its current popularity. But the browsers have brought chaos with them.

The Web no longer seems to be a worldwide collection of linked text files because browsers have gone far beyond merely wandering the Web. The browser provides an interface to most of the Internet's services and also has become an interface to the computer running the browser. Vendors, hackers, phishers, and anyone else who can persuade or trick a browser user into interfacing with a computer can run their programs on that user's machine.

There would be less chaos if browsers could only load passive files and if another kind of program, a *seller* let's call it, could interface with files that provide active and potentially dangerous services. The seller program's developers could focus on security and, since most of these services are commercial, on identification and payment. There could well be a market for simple handheld devices specialized to run sellers.

Search engines

The Web's success has meant that a huge variety and stupendous number of text files have become available on it. To make these files more accessible amid the chaos, many users turn to *search engines*.

The search engine has two components. The *crawler program* continually extracts text files from the Web and constructs an inverted file from their content and URLs. The *responder* uses textual queries to select entries from the inverted file and constructs a list of URLs that match the query, then sends the list with some relevant context to the browser that sent the query.

Search engines have been remarkably successful programs because they hide the chaos. Even so, they only provide access to a small fraction of the Web's content and use somewhat arbitrary methods to select and rank what they process—making their activities more of a raid than a search.

Clearly we need some way to formally separate the Web's wheat from its chaff, something like a reviewing procedure to qualify Web files for inclusion in inverted files for searching. The people responsible for the reviewing in any area could also be responsible for making sure all the qualified files are stably and permanently stored.

Addressing the Web

Web addresses, or URLs, are themselves textual and based on the Latin alphabet. This raises important issues for those who use other writing systems—a population that is rapidly increasing its Web usage.

The Latin alphabet's predominance is unlikely to continue, particularly when the Chinese realize that their wonderful writing system needs encoding by radical rather than by character to save it from chaos. Segregation of Web content by writing system, if not by language, seems inevitable and will be much easier to implement if anticipated.

A primary requirement will be meeting the need to separate DNS server complexes for each writing system. This challenge seems closer to hand given that the US government has opted to retain its overall control of the present DNS complex (news.bbc.co.uk/1/hi/technology/4640441.stm).

After all, the vast majority of Internet messages, whether selling or searching, use only one writing system. Most machinery and applications are specific to a particular writing system, and trying to mix systems usually proves difficult and unsuccessful. Everyday keyboards in English-speaking countries can't even properly handle other languages that use the same alphabet.

THE POPULACE

Another global network runs alongside the Internet and the Web: the social network. Introducing new technology to so-called backward societies can bring chaos (www.unitedearth.com.au/HNHinterview.html), but it can also disrupt developed societies. At one time, shopping provided a rich social activity populated by real personalities like grocers, greengrocers, butchers, bakers, chemists, newsagents, and clothiers. The social component of today's supermarkets is minuscule, and seller programs will remove it altogether.

Functions like management drift further and further from personal contact to digital mediation, using e-mail, spreadsheets, and PowerPoint via BlackBerries and their like. I have heard modern management called a parallel world, and I suspect that middle management may soon move to outsourcing on its way to becoming completely automated.

The biggest threat I see is that entertainment and marketing will take over the Web as they have taken over television and radio, media that at first seemed to offer great potential benefit to personal society in areas such as education and information.

Perhaps the best way to get a truly social benefit from the Web will be to support and extend efforts like Wikipedia (www.wikipedia.org) and Project Gutenberg (www.gutenberg.org), and move toward a stable and organized Accessible Authoritive Archive (AAA), as the British Broadcasting Corporation has tried to do against determined opposition (www.guardian.co.uk/online/story/0,3605,1522351,00.html).

Chaos might seem too strong a word to describe the modern Internet and its Web, but even if it's only an awful mess, the computing profession must take some responsibility for that mess. To clean it up, we must start by cleaning up the terminology, then proceed by imposing a clear structure on the Web.

A good first step might be for our Computer Society to press, perhaps through the International Federation for Information Processing (www.ifip.or.at), for professional computing societies to pool their publicly accessible publications in a common AAA complete with inverted file. Open documents from the computing industry could also be brought in, such as those at www.research.ibm.com/journal/rd/. ■

Neville Holmes is an honorary research associate at the University of Tasmania's School of Computing. Contact him at neville.holmes@utas.edu.au. Details of citations in this essay and links to further material are at www.comp.utas.edu.au/users/nholmes/prfsn.