2

The Internet

Project manager's responsibilities

- When setting up a website make sure the domain name has been chosen and registered correctly
- Confirm that the server can serve the types of assets (graphics, audio, etc.) included in the web pages
- If you are non-technical, know enough about the mechanics of the Internet to help communicate with the technical members of your team and your clients
- Act as intermediary representing any concerns and issues of the technical personnel to the client
- Be aware of the trends and risks to help guide your decisions

The Internet: Why? What? When? Who?

Networks are almost as old as computing itself. Before personal computers became widespread, most computers were operated using remote terminals connected to mainframe or mini-computers. A typical mainframe, occupying a whole room and typified by the huge machines beloved of movie makers in the 1960s, featured numerous cabinets, each with a different function: processors, memory, disk units, tape drives and so on. These parts of the computer had to be networked together.

A mainframe from the 1960s also featured a number of terminals which could be used simultaneously. Initially they were teletypes – basically electric typewriters – and then they became screens and keyboards. This combination became known as a terminal. In some cases the terminals were in different buildings to the computer, although the connection was always on a one-to-one basis with a long wire linking each terminal to the computer.

J.C.R Licklider at MIT is credited with first envisaging what he called the 'Galactic Network' concept – interconnected computers through which anyone could quickly access data and programs from anywhere in the world – in a series of memos written in August 1962. Within three months of writing up his concept, Licklider became the first head of the computer research program at DARPA (Defense Advanced Research Projects Agency) and his vision undoubtedly influenced his colleagues and successors in the organization. By the early 1970s, DARPA's predecessor, ARPA, had been behind a program to link together research computers around the USA, in order to make better use of what was then a scarce and expensive resource. This network was also designed to be connected up in such a way that losing a single connection (or even several connections) would not close things down because another route could be found between any two machines. This coincidentally made the network less vulnerable to deliberate sabotage or attack.

ARPANET – now wound down – was the result of this initiative and it was an internet: an *inter*connected *net*work. The protocols that were eventually used for traffic on the network, allowing files to move around the network intertwined with each other and to control how a file was routed from one machine to another, were the basis of the TCP/IP (Transmission Control Protocol/Internet Protocol) system used today. ARPANET's ghost lingers however, and occasionally Internet log files will show visits from the .arpa top level domain. These machines are used to map IP (Internet Protocol) addresses back to domain names, which you might want to do to identify visitors to your website in log files.

The National Science Foundation took over funding of the Internet backbone in 1988 and replaced the ARPANET with NSFNET. In the mid-1990s the backbone changed again, and is now a commercial operation run by commercial telecommunications companies (telcos) spreading into most of the countries of the world.

So who owns and controls this Internet? Undoubtedly different organizations 'own' the infrastructure on which the Internet exists. But if one of them disappeared the traffic would find another way to travel because the Internet is multi-connected. The Internet is not even like the world's telephone network, because the users don't even need to know which country their target is in. The domain name is all you need and the Internet systems do the rest. So although some organizations have responsibility for the numbers and names used on the Internet, the Internet itself has no centre, no administration and no owner.

Connecting a computer to the Internet

The Internet is 'out there' just like the worldwide telephone network. To make use of it you connect to it and use your computer to 'talk' (using the TCP/IP protocol) to other computers on the network. You have to have a unique IP address, which is the Internet's equivalent of the phone number. But how do you connect? Before we look at the role of the Internet service provider (ISP) and the use of domain names and other such paraphernalia we should look at the physical practicalities. Basically there are two kinds of Internet connection: dial-up and 'always-on'.

A dial-up connection means just that: you connect to the Internet by dialling into it using a telephone or ISDN line – Integrated Services Digital Network. If it's a phone line, a modem is used at both ends of the connection, and this translates the digital computer data into an analogue signal so it can go through the telephone system. Modem stands for modulate-demodulate. A fax machine works in the same way. But modems have limits on speed and even though ingenuity has lifted modem speeds up to around 50 kilobits per second we are probably getting to the limits of such methods. Modems are very susceptible to 'noise' and other imperfections on the line. In addition to this, telephone companies are usually using sophisticated techniques to maximize their bandwidth and this may conflict with what the modem (or indeed a fax) may need to work at its best. A digital ISDN dial-up connection will be more reliable and the process of establishing the Internet connection will be faster, but per channel you still only get 56 or 64 kilobits. It is possible to link channels together to increase speed but this usually counts as more than one call and is charged accordingly.

Increasingly people are connected to the Internet through a continuous link, rather than dialling up. Most companies with networks of their own will have a link over a private data connection. But cable modems, satellite and DSL (digital subscriber line or loop) are more recent options. Now we are getting into broadband territory with data rates in megabits available at – and this is the important part – quite low cost. We are talking about tens of dollars or pounds a month for this kind of connection and, because it is always on, there are no call charges. Unfortunately the lower costs mean that the telcos could see themselves losing money as a result of people

moving away from dial-up and, when you add regulatory restrictions into the mix, you can see that the rollout of consumer-oriented broadband at low cost is not going to have a smooth ride everywhere.

Satellite broadband works by having a connection from the Internet to you through a direct satellite link to your computer via your satellite receiver dish. You would use a phone call or something similar to provide the connection from you to the Internet so then you'd receive the data back from the Internet via the fast satellite link. This can work virtually anywhere and it will eventually be possible to dispense with the non-satellite part of the loop.

Cable modems are provided along with television and telephony services by many cable operators. Of course the cable has to go past your door, but in many parts of the world it does. Even though the potential bandwidth is high you will be sharing that with everyone else on your cable.

The most common form of DSL is ADSL (A for asymmetrical) and works by piggybacking the broadband data channel onto an ordinary copper telephone line. Within a few kilometres of the exchange you can get data rates up to two megabits from the Internet and half a megabit back. This doesn't interfere with the telephone so you can still make calls. Your ADSL operator, who may either be your telco or an intermediary buying in bulk from your telco, may share each connection between several people. This is what is called the contention ratio and if the ratio is 10 then there are up to nine other users competing with you for the two megabits.

There are other ways of connecting, for example you might pay for a permanent wired connection (known as a leased line), especially if you are a medium or large company in an urban area. There are wireless connections using the services of a mobile phone operator (see Chapter 4) or even neighbourhood wireless networks using wireless LAN (local area network) cards and base stations. One important thing to remember about all of these systems (except the neighbourhood wireless one) is that they will be provided commercially by an operator. This means that whatever is technically possible will always be filtered through business and marketing decisions and the service offering will change on a regular basis. So if you have an Internet connection you need to continuously review it. Neighbourhood wireless LAN can work by neighbours informally grouping together to share a single broadband connection, but this is actually illegal in some parts of the world since it can be interpreted as setting up a telecommunications network and therefore needs to be licensed.

Now we're wired up, let's look at the Internet connection from a systems viewpoint.

Connecting a computer to the Internet requires that the machine has a unique identifying number which 'places' it on the Internet, called its IP address. These addresses, which are numbers made up of four one-byte numbers separated by dots (such as 172.32.23.1), are allocated by a small group of organizations of which the main three are the American Registry

of Internet Numbers (ARIN), The Asia-Pacific Network Information Center (APNIC) and Réseaux IP Européens (RIPE) who 'carve up' the world's IP addresses between them and sometimes delegate for individual countries. So you get the numbers from ARIN, APNIC or RIPE as appropriate. The possible number combinations used at present for IP addresses are large but finite and the system will eventually have to be updated.

You will probably connect to the Internet through an Internet service provider (ISP) who will themselves have a fast connection (fat pipe) to the Internet at large – known as the Internet Backbone – and they will give you a set of IP addresses for use on your network. In an organization which is already on the Internet, the IT manager will probably allocate IP addresses to individual machines from the list that the company has been given.

If you were Imaginaryco's IT manager and you wanted to put Imaginaryco on the Internet – and the .com domain was where you wanted to put it – then you'd contact a Domain Name Registrar for the .com domain (or one of their resellers on the many websites offering domain name services), see if the domain imaginaryco.com was available and, if it was, pay to register it. If you were in the UK you might prefer to go for the domain imaginaryco.co.uk or perhaps imaginaryco.fr if you were in France, because there are top level domains (TLDs) for countries as well. Historically US organizations have used the generic top level domains such as .com, .org and .net. This reflects the start of the Internet in the USA. The TLDs .edu, .mil and .gov are reserved for the US education, military and government although non-US organizations can register names in the others, and often do if they want to appear to be international. Incidentally, there is a .int TLD, but this is reserved for bodies set up by international treaty such as the European Union.

Different top level domains have different registrars. The .com domain actually has several registrars. Once the domain imaginaryco.com has been registered the computers on the Imaginaryco network can be named.

When you register a domain name you need to check that you are recorded as owner of the domain. For .com domains a check on www.whois.net will tell you. It has been known for some domain name registration resellers to leave their names as owners instead of changing it to the buyer's. This can cause a problem if you want to move the domain between ISPs because it appears that you are not the real owner.

All the allocation of IP addresses, domain names and Internet protocols is carried out under the auspices of the Internet Corporation for Assigned Names and Numbers (ICANN). This is a non-profit corporation that was formed to assume responsibility for the IP address space allocation and domain names, among other things. ICANN have an ultimate authority over certain aspects of the Internet and it is they, for example, who are responsible for creating new top level domain names to go along with the current ones. This process started in 2001 with the addition of the .biz and .info domains, and more will follow.

To sort out the domain name for a particular computer you look 'outwards' towards the Internet. If you work for Imaginaryco and their Internet domain is called imaginaryco.com and you want to add a machine called 'Bilbo' to their network you would ask Imaginaryco's IT manager to arrange this. The IT manager would set up the DNS (domain name server – of which more later) that effectively told the world at which IP address the machine bilbo.imaginaryco.com lived.

As long as your machine is the only one in imaginaryco.com to have the name then you can call it what you like. A computer used as a web server is usually called www for World Wide Web (which makes sense) but it doesn't have to be. You will sometimes see websites which have domain names that don't begin www. The World Wide Web uses a system called a Uniform Resource Locator (URL) which (in an absolute URL) consists of the domain name for the relevant machine together with the protocol being used. In our example this would be http:// which defines the protocol used (others might be ftp:// or mailto:) and www.imaginaryco.com which is our machine name. The URL can include more information like a path down to an actual web page, the file extension and occasionally there is an extra number which denotes which computer software 'port' is requested. The default port for a web server is 80 but others could be used. So the full URL might look like http://www.imaginaryco.com/sales/prices.html:8080 which says that this web server is on a different port to the normal 80. Unless you want to run more than one web server on a particular machine (or you are an IT specialist) you shouldn't need to worry about port numbers.

As an aside, you find interesting names on computers on the Internet. The IT people at ebay.com seem to be Star Trek: Deep Space Nine fans since they have machines named Garak, Kira and Keiko on their network. I know this because I looked at the route e-mails took from them to me and they passed through those machines. You need to remember that any Internet machine name might be seen by other people. (I am not really a Hobbit freak: Bilbo was just an example.)

So now you have your machine set up, connected to the Internet and ready to roll at its brand new domain name. If it's a web server you will have tried to find a domain name that best represents the website. This could be the company name as in the Imaginaryco example but you might have a more generic website name. You can probably guess what you'll see when you go to www.carp.org without me telling you.

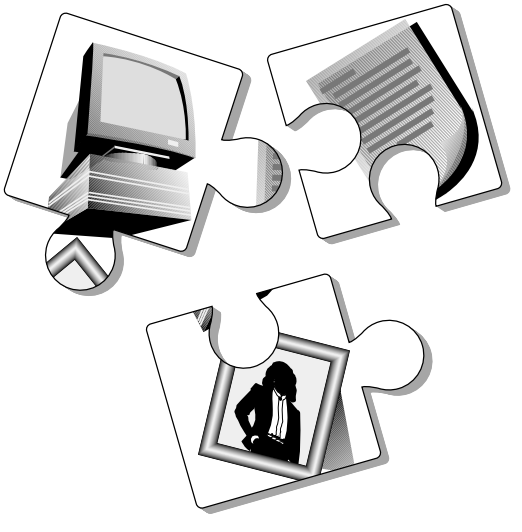
Setting up a website

Let's look at what is involved in setting up the website itself. This chapter won't discuss how to produce the pages but it will cover the basics of getting a server and setting up the files to make up the site.

There are three choices for setting up the site. You can use a computer of your own on your own premises (as outlined above), you can co-locate your own computer at the premises of an ISP (or someone else with a good Internet connection) or you can rent space on an ISP's server. As far as your visitors are concerned these options can look exactly the same so the decisions about where to put the site are managerial and technical issues you, as project manager, have to address.

A web server is actually a computer program which accepts requests for web pages from the Internet (using a protocol called Hypertext Transfer Protocol or HTTP) and serves pages, graphics, Flash animations, videos and so on back as a result. It will also deal with requests to run other programs on the server for more specialized tasks. A web server can run on almost any computer but many of the world's websites are hosted on computers using the Unix operating system and many of those use a web server called Apache. One reason for this is that Apache is free, but it is still a very reliable and powerful web server. A common server for PCs is Microsoft's IIS (Internet Information Server) and the best-known server for Macs is WebStar.

There are others. Internet Product Watch listed 96 different servers and server-related software at the time of writing.



Web Site Construction Kit.

A static website – one which consists basically of linked HTML (Hypertext Markup Language) pages – sits on the server like a bit of a hard disk directory structure. That's because that is exactly how it is set up. The server will define which directory is the root of the website and everything on the site sits below that. The server also defines the default filename. This is the one you get if you don't type an actual filename (ending in .html for example) in the URL. This is usually index.html or default.html (or .htm) but some web servers will allow you to specify any filename as the default.

Bear in mind that if you don't have a file in a directory with the default name then visitors may see a complete listing of files in that directory instead. Visitors will also see the directory structure and file path when they select a page so you shouldn't name directories and files in a way that might cause embarrassment to you or your client.

The filenames on the website can be the same as the ones on your computer but you should consider a couple of issues.

If you have a PC or Mac then your filing system will treat upper or lower case in the file path, directory and filenames as being identical. PCs and Macs are case-insensitive and treat ABC as being identical to abc. The Unix filing system used on many web servers is case-sensitive (ABC is different to abc) and you will risk broken links if you don't take great care about this because if the filing system is case-sensitive then the URLs on that server will be case-sensitive as well. It is a good idea to always use lower case in the paths on a website and you can do this on your hard disk version as well. The case of filenames used in links is one of the easiest ways to break a site when moving it from your own machine to its final home on a server.

Some characters are not 'legal' in web requests and the most commonly used one is a space. The Internet should translate spaces into the characters %20 (because 20 is the ASCII code for a space) but it is better to avoid spaces altogether. To be safe stick to a–z, 0–9 and an underscore character '_' which you can use to simulate a space if you want a filename to be easier to read. Don't put an underscore in a URL you publish in print as most users will make a mistake with it. Other characters like hyphens are also legitimate and you will find many domain names with hyphens in them. However, some characters like =, +, $ and ? are used in passing variables to other programs running on the server, / and : are file path delimiters and the tilde '~' is used to refer to home page space in Unix so these can cause trouble. If in doubt stick to the basics.

A small website can be built using static HTML but as the site grows it will become increasingly difficult to manage. A large dynamic site (like a news one for instance) uses a database to build pages on-the-fly for visitors. The database holds the text, images and other resources. The journalists use a form in which they enter the text of a story. They might choose a page layout from a number of templates and they might specify an image or video that goes with the story. The database program would then build the page for the story each time the story was requested by the server and send it to the viewer. This is one aspect of a content management system. To the

visitor this looks exactly the same as if the page was created statically with HTML.

The standard which connects the web server to an external program is called the Common Gateway Interface and so these programs are sometimes called CGIs. The language Perl is a popular mechanism for programming CGIs and is freely supported on most computer platforms. Since it is closely linked with Unix, Perl had a head start on many servers but its popularity has waned.

Server Side Includes (SSI) and systems like PHP, JSP and Microsoft's Active Server Pages (ASP) are now more commonly used to generate dynamic pages because they are more tightly integrated in the server software. This avoids an overhead where a program like Perl has to start up before being able to work. (The monthly Security Space survey of web servers, which looks at three million servers on the World Wide Web, determined in mid-2001 that the Apache web server was in use on about half of the sites and in those PHP was the most popular extension.) In these cases the page template appears to be an HTML file but with special tags. As the server-side scripting language parses the page, it replaces the tags with dynamic content as required. This can be as simple as calling up commonly used sub-pages of static HTML, like a menu bar, so that this code doesn't have to be included in every page. At this level, dynamic pages make site maintenance and updating easier. If necessary they can still link to another program, such as a relational database, to call up content. While web pages made of static HTML will end in extensions .htm or .html you can get a clue as to what other systems are being used from other file extensions in URLs, such as asp, dll, php, jsp and so on.

If the pages include JavaScript or some other client-side (browser) code together with HTML and server-side inclusions then what seems to be a single web page can be dynamically built from three or more sources simultaneously. We'll look at the client–server relationship in websites more closely in Chapter 5 of this book, *Platform parameters*.

When a website is being built, pages will need to link to each other. This can be done by including the full URL as the link address but it is usually better to show URLs relative to the page with the link. This means that when the site is moved elsewhere (as it inevitably will be) the internal links will still work. Obviously links outside the site still need the full (or absolute) URL. Less obviously, links in pages built by CGI programs might also need absolute rather than relative links. This is because the browser works out the whole path for a relative link based on where it thinks the current page is. If it thinks the current page is in a CGI directory then the links are unlikely to work correctly. From time to time it is possible for a browser to make mistakes with its paths, especially if it is running out of memory. This issue is dealt with more fully in Chapter 9, *Integration*, in this book.

There is also a more generic term URI which stands for Uniform Resource Identifier. A URL is a special case of a URI and since the difference between the two is rather esoteric, for most of us the term will be URL.

The web server needs to be configured to serve different kinds of content in the appropriate way. A browser doesn't necessarily use the file extension to determine how to handle something: it should also look at the Multipurpose Internet Mail Extension (MIME) type which is included in the header of the file the server sends back. Browsers are notoriously inconsistent in this and to be safe the extension and the MIME type must match. Incorrect MIME types can result in pages that won't display properly and are especially problematic with database-driven sites where the database has to specify the MIME types as it builds the page.

When a new kind of file is added to a web page it is important that the web server knows how to describe it in this header. This is part of the configuration of the server. If not, the file will probably be treated as text and if the file is actually an MP3 music file then you can see that it just won't work and the browser will display a page of gibberish.

MIME was originally specified to handle encoding of attachments to e-mails, as you might guess from the name. Some MIME examples: image/gif, video/quicktime, and text/html.

Other server configurations can be basic ones like the default file name described above or it can be setting different pages to be displayed depending on the language the browser wants to use. This in turn would have been set in the user's preferences for the browser and, by default, is usually set for English. Finally, the server logging will need to be configured so that the information you want will be recorded for later analysis.

Scalability

Websites can be victims of their own success and if your site is going to receive thousands of hits it needs a different technical design approach to that of a simple server. This extends to every aspect of the system: the link to the Internet backbone, the number of computers used, the amount of disk storage, the server software and the middleware or gateway (CGI) techniques.

This goes beyond 'mere' load balancing between machines, where a number of separate computers work together sharing the load on the website. The system will need to take into account the overhead involved in serving pages and executing the searches (or whatever) required to build a dynamic page. Server software will need to be able to execute many transactions simultaneously, ideally as threads of the one server process rather than by spawning many processes – because it's faster and more responsive. The links to any commercial transaction system need to be efficient and the programs may need to be written using optimized fast code using, say, Java rather than interpreted Perl scripts. At a much simpler level, image download times can contribute to scaling difficulties.

The difficulty with scalability is that you can get swamped by a sudden wave of traffic to a website. New media developers coming into the world of

Stress Testing.

television can find the speed of response and sheer numbers involved very surprising. But if you think about it, if millions of people are seeing a plug for a web URL at the end of a top-rated show then a significant number of them will attempt to go to the site within a very short period of time and this can make many web servers fall over.

If there's any risk of this then the server should be stress tested. This can be done by writing a suitable program which impersonates one or more web browsers and users doing things like looking at pages, entering passwords, clicking on links and doing their online shopping. Run many of these on many machines and you can give the site a hard test. There are companies who will stress test websites with prices into hundreds of thousands of dollars: but they would argue that this is better value than seeing your mission-critical shopping site go down under a heavy load. For smaller solutions, vendors like Microsoft (who have a free Microsoft Web Application Stress Tool for IIS) provide software so you can do it yourself and there is a Java-based tester called Jmeter provided by the Apache web server project. Clearly this kind of testing should be done before the site goes live for real.

These issues of site size, type of page and scalability affect all the decisions that your programmers need to make about how to organize and develop the applications for a client. It is questions around these issues that your technical analyst or equivalent would need to raise as an extra part of the

analysis for a scoping questionnaire when getting to grips with a new project for a new client (see Book 1 Chapter 3, *Scoping*).

DNS and other initials

Back in 1965, when researchers connected computers in Massachusetts and California together over a telephone line for the first time, identifying the machines was not an issue. But today we have millions of machines on the network. To recap, IP addresses are used to uniquely identify a machine and these take the form known as dotted-quads, like 123.231.12.23, and if the user of the machine wants to access another, the first machine has to know what the IP address of the other one is. IP addresses are the telephone numbers of the network.

Clearly these IP address numbers are not very friendly and so machines are named instead. A distributed database called the Domain Name System (DNS) holds a lookup table that maps IP addresses to domain names so you can find the IP address when you start from the domain name. A complete machine domain name – one that uniquely identifies it – is known as the fully qualified domain name (FQDN). So, .com is a domain but mailhost.imaginaryco.com, which identifies a mail server computer at the imaginary company Imaginaryco, is an FQDN. But for simplicity let's just call it the domain name.

If you make a request for another machine you will either know its IP address or (more likely) you will know its domain name: you don't need to know where it is physically. (In fact, at the time of writing there is only one server on the whole Internet which will tell you where a domain name is physically/geographically.)

Each computer on the Internet (with its own IP address) will have a gateway to the Internet. It will also know another machine's IP address where domain names can be translated into IP addresses. This is the address of the domain name server (also sometimes known as the DNS). So if someone wants to connect to mailhost.imaginaryco.com what happens, in simplistic terms, is that their computer will first ask its local domain name server to supply the IP address that corresponds to mailhost.imaginaryco.com. The server will probably do this by first asking the name server that holds information for .com for the coordinates of the name server for imaginaryco. When it is told this it will ask imaginaryco's name server for the address of mailhost. The DNS then passes mailhost's IP address to the computer that made the request in the first place and it will be able to make its own connection to, in this case, send some mail. The Internet is so large that it is not practical for every DNS to know where everything is, and also ruggedness criteria mean that there is always more than one DNS available to each user and more than one server that has the data for each domain. In fact there will be a large number of servers which have information about the larger domains like .com and they will be synchronized

and updated regularly. It is the time taken for this information to reach all the servers that limits the speed with which a domain can be set up or moved.

If you have a network connected to the Internet you can have your own DNS machine which will store information about the machines on your network and also about other computers that your network regularly contacts. You could also delegate your DNS to your ISP, and many smaller networks will do this.

We know where you live!

The 'fountain of all knowledge' about a domain is known as the Start of Authority (SOA) and this is the DNS computer which has the authoritative record of the IP addresses for the domain name. The authoritative record for a particular computer is called its A record. So in the case of our Imaginaryco example, the SOA might be called dns.imaginaryco.com and this computer would know about every machine on Imaginaryco's network connected to the Internet.

However, the A record that says where a domain name is can point outside the local network and this can be useful when a domain is moved and its IP has to change. This happens more often than you might think: servers move from one location to another or the management of a website changes. To move the domain from one network to another the relevant DNS records have to be changed for both networks. This can take time as the changed information is passed around the DNS computers. A useful shortcut is to modify the A record on the original SOA DNS machine to reroute the connections until the new SOA information works its way around the Internet. One other thing that the SOA record contains is the MX record. The MX identifies the computer that will accept mail for the domain. In the case of imaginaryco.com, the MX record would point to something like mailhost.imaginaryco.com.

There are special reserved blocks of numbers which are never allowed to be on the 'real' Internet: these can be used on your own internal network. They are:

10.0.0.0	to	10.255.255.255
172.16.0.0	to	172.31.255.255
192.168.0.0	to	192.168.255.255

and you can use numbers in these groups without reference to anyone else outside your own closed network. You don't have to use these numbers if you have been allocated enough IP addresses by ICANN but using them will give more flexibility. Incidentally, a closed network which uses all the protocols and programs of the Internet but is not necessarily connected to it is known as an intranet.

So can your closed network with these reserved IP addresses connect to the Internet? It can by going through a gateway router that uses Network Address Translation (NAT) and which will hide your network's addresses from the world outside. The Internet only sees the router, not the network behind it. One advantage of this technique is that whole networks can be hidden behind a single IP address, which saves using up numbers. (Another way this is done is by dynamically allocating IP addresses to machines when they connect to the network and this is usually done by dial-up ISPs and on many private networks.) Using NAT is one way of implementing a firewall, which is a computer which acts as a gatekeeper to keep intruders out of your network.

When one machine connects with another on the Internet it will usually be asking for a transaction using a particular protocol and/or at a particular port. In this case a port is an abstract software concept to help with the management of Internet traffic. The port number is one of a standard group including things like HTTP (for web pages) on Port 80 or a DNS enquiry on Port 53 and these are called 'Well Known Ports'. A firewall or NAT router can be programmed to route requests for specific ports to particular machines on the closed network. It could, for example, route requests for Port 80 to a machine that operates as a web server.

Every machine on the Internet has to have a unique IP address, although not every one has a name. Of course, if it has a name then you can pin it down to a company or at least an ISP. Most ISPs (but not all) allocate casual users an IP address when they log on, so it is not possible to pin down the individual machine in this case. If the surfer is routed through another computer then it is the IP of this machine that you will see.

You should note that this is not necessarily the case with all Internet transactions. My e-mailer, in common with many others, notes the IP address of the machine on which the mail was written, even if it is a reserved IP address on a private network.

As we've introduced the function of a firewall, you will find that a firewall can also decide whether to admit connections by checking for unusual use of ports or protocols and by blocking or allowing access depending on the IP address of the computer asking for the connection.

… or do we?

How much can the identity of an individual machine be pinned down – even what country it is in? This is a subject of more than passing interest since, increasingly, Web content is being subjected to boundaries. Some web pages may be 'illegal' to view in some places – such as Nazi memorabilia – and some organizations may wish to limit access to sites geographically.

If the web address is www.societeimaginaire.fr then that should be clear. If they are not in France then they want to be. Some countries are not so nationalistic about their top level domains. Italy's .it is used elsewhere because 'it' is a useful English word – consider a site called 'book.it' and

you'll see the attraction, and the Tuvalu domain .tv is marketed at television companies. These top level domain names can be worth a lot of money and could be the next big thing for small countries since stamp collecting. But, of course, where is a .net or a .com?

You might think that the IP address would be traceable territorially, and this should be the case, although you would have to have a database relating IP to countries for it to work. In this case, if a proxy server in a different country was used, this would fool the system. Why would you need a database? This is because the allocation of number groups to countries is not simple.

Let's use, as an example, a set of what are called Class B numbers. Here we are looking at a grouping of IP addresses representing medium-sized networks. My own ISP is Demon Internet in the UK and they have a Class B allocation starting at 158.152.0.0 which, by going up to 158.152.255.255 allows them about 65 thousand addresses. You might think that the other Class B networks starting with 158 would be geographically similar. No, they're not. They are probably allocated in the order they were asked for. So you find such geographically diverse networks as Moscow State University, Hong Kong Baptist College, the Australian Prime Minister and Cabinet and the USAMC Logistics Support all sharing IP addresses starting 158.

To add to the problem, it is possible that different subgroups of IP addresses allocated to an international organization may be in different countries. Where are the networks of the European Commission, for example: in Brussels, in Luxembourg, in Strasbourg?

You can follow the path from your machine to a distant one by using a utility called TraceRoute. This is a standard part of Unix but for other machines you will need to get a small program which does it. When you run TraceRoute you see all the hops that the connection makes as it makes its way around the world.

Information on where a Domain Name or IP address is can be found by querying what are called WHOIS servers and the registries who allocate the IP addresses. This is probably not totally foolproof since it is technically possible for a machine to come into the Internet via an anonymous mirror or even using internal connections in a multinational organization, but the chances are that if you really need to know where your web visitor is, you can find out with a high level of confidence.

Even though many people browse the Web through proxies or firewalls and these machines' IP addresses are the one you will see, they are probably in the same country as the real user's machine.

There's more to the Internet than the World Wide Web

The Web has been described as the 'killer application' for the Internet, but there are other things you can do with it, of course.

E-mail

The biggest use of the Internet is for e-mail. There was e-mail before the Internet but the Internet has forced a standard so that anyone can exchange mails with anyone else. Some transactions on the Internet, like file transfer and web browsing need a full point-to-point connection between the machines at either end so that a dialogue can take place. Mail, which works using a relay system, does not need this.

Any e-mail you send will initially be a transfer between your machine and your mail server. You will send the mail to the server and the server will work out what to do next. It will look at the e-mail address – of the form user@machine.subdomain.domain – and will find out from the domain name server which machine handles mail for the domain in question. The mail will then be sent to that machine as a point-to-point transfer. In some cases the mail will be relayed more than once but eventually it should reach the mail server for the destination. The recipient will then either download the mail when they log on to the server (known as Post Office Protocol or POP mail) or the final point-to-point destination will be the recipient's machine (using Simple Mail Transfer Protocol or SMTP). A recent alternative to POP is IMAP (Internet Message Access Protocol) which allows users to use their electronic mailboxes seamlessly from any location and some mail clients can be configured to work with either standard.

An e-mail address can be an alias, which means that any mail received for that address is automatically routed to another address (or addresses). In this way mail to sales@imaginaryco.com can be routed to whichever person is handling sales at that time. The alias can be any e-mail address and doesn't have to be a local one.

There is no set way of tracking a mail. Usually, as it is relayed along, each server will store it briefly until the next scheduled time for a transfer and will then send it. If, for some reason, it can't be transferred the server will wait until the next transfer time. Sometimes the server will send a message to the sender of the mail saying that the mail is being held. Rarely, the system will 'time out'. I have found a particular problem with some university mail servers being turned off during vacations!

Because of this relaying of mail, the system is not instant, although it can work within minutes if the mail servers being used relay mail frequently. Some are set up to relay on demand, but not all. Another mail relay issue is the relaying of mail for unauthorized users. In the golden age a mail server would relay mail for anyone who asked, but this was so abused for spam (bulk unsolicited e-mailing) that now mail servers will usually only relay mail for people in the same domain.

If you set up a mail server you have to take care that it cannot be used for spam. If you do not do this it is likely to contravene the agreement you have for Internet connection through your ISP. If you only set up websites then you might think this doesn't apply to you but don't forget that a form on a web page can generate an e-mail. You will need to check to make sure that someone impersonating your form cannot send spam. (If you didn't know, the use of the word spam is in homage to a Monty Python comedy sketch about a restaurant whose speciality was the said tinned meat product that was part of every dish.)

In theory, e-mails can only contain text and, on almost all computers, text is represented by a standard called ASCII (American Standard Code for Information Interchange) which maps alphabetical and other characters to numbers. Strictly speaking e-mails can only be 7-bit ASCII (numbers up to 127) and although 8-bit ASCII (numbers up to 255) can be transferred, some servers may lose that eighth (top) bit which can change some of the more unusual (as far as English is concerned) characters. It is possible to attach a binary file to an e-mail but in order for it to pass through the mail system the mail client will encode the file into ACSII using one of a number of standard formats such as Base64, BinHex or UU-encoding. MIME types were specified to help this process. If this seems archaic then it is sobering to think that there was something of a battle to get ASCII to include lower-case characters, never mind anything 'foreign'! It is increasingly popular to produce an e-mail message as if it were a web page, in which case graphics and other assets are referenced using links in HTML and the HTML itself is, of course, text.

FTP – file transfer

A file transfer uses File Transfer Protocol (FTP) and, like mail, the computers at each end need to be able to handle the protocol, usually with special programs. FTP is important for websites because their files will probably be loaded onto the web server using FTP.

Unlike mail, a file transfer has to be between two machines connected point-to-point via the Internet. The sender makes the FTP connection and the recipient says 'OK send me something'. The file is then transferred, with checking done continuously to make sure the transfer is correct. (This is similar to the x-modem protocols used before the Internet.) Along with this function is the ability to do simple file system work such as view directories and even delete files.

A file can be transferred as a text file or as a binary file. If the file is binary then it is transferred without changes, but text files are usually checked to make sure that the carriage returns (CR) and/or line feeds (LF) are appropriate for the machines involved and, if necessary, changed because they can cause problems. If a binary file is incorrectly sent as a text file then there is a strong risk that it will become corrupted. If in doubt, transfer as binary. (End of paragraph/line is CR+LF on a PC, CR only on a Mac and LF only on a Unix box.)

If you have a Windows or Unix machine then FTP is built in. On Windows it is available from the Run dialogue box. On the Mac, Fetch is a common FTP client with, as you would expect, a mouse-driven interface.

Remote control – TelNet

Working with a remote computer can be done using a protocol called TelNet. There are WYSIWYG visual systems for remote control such as Timbuktu from Netopia (originally from Farallon) and Virtual Network Computer from AT&T Labs in Cambridge (UK). With TelNet your machine operates the distant machine as a text terminal and this is often used to allow website 'owners' to work with their websites on a remote Unix or NT box. You can't TelNet into a Mac web server (unless it is System X, which is Unix-based) in order to control it because the Mac has no command line interface, so Timbuktu or VNC would be more appropriate if you wanted to remotely control a Mac. Depending on how the remote computer is configured, a TelNet connection can do anything a local user can do. This is, of course, a throwback to the 'bad old days' of teletype terminals and mainframes.

If you have a Windows or Unix machine then a TelNet client is built in. On Windows it is available from the Run dialog box. For the Mac there is a free client from NCSA called NCSA TelNet, now produced under the name 'Better TelNet' and MacTelNet.

Chat

There is an Internet standard for chat, called the Internet Relay Chat protocol or IRC, but some other standards have caught on including ICQ, AOL Instant Messenger (AIM), Yahoo! Messenger and MSN Messenger. All

these systems use servers through which chat messages are routed and on which users are registered. IRC can also work between two machines as long as each has an IRC client.

In some respects more sophisticated Internet messaging (AIM for example) is like the SMS text messaging system used on mobile phones (see Chapter 4) and can be used as a 'push' system as long as users are logged on. It can be used to send status messages telling you your stocks have moved or send you news headlines. Since the client is likely to be on many machines and since the chat systems generally don't leave files lying around you can log on wherever you find yourself. The linking of chat and mobile phones seems inevitable.

Videoconferencing is also possible over the Net. Depending on software and bandwidth you may get stills and text or full motion video with sound. CU-SeeMe was one of the earliest such systems but iVisit is an alternative (written by the same person) and Microsoft's ubiquitous NetMeeting can videoconference as well. Just as with ISDN-based videoconferencing, some of these allow you to share whiteboards and documents as well.

Newsgroups

One area of the Internet that has existed since long before the Web is the Newsgroups (also known as UseNet groups). These are bulletin boards on which users can hold threaded discussions and, in some cases, post files. The threading of a discussion is very useful and can be applied to bulletin boards or forums on websites as well. Basically anyone who posts a message to a newsgroup can either do it in isolation, as a new subject starting its own thread, or the message can be a comment to a previous message, so continuing the thread. This method allows for a discussion in a way e-mails don't do so well.

There are two kinds of newsgroups. The main groups are set up under a formal system where someone from an existing group posts a suggestion that another group be set up to cover such and such a subject or subdivision and other people then vote on it. If it gets enough votes then the group can be set up. The group will need to be placed within the hierarchical structure of newsgroups so that a group on multimedia authoring might go in the comp area (for computing) and thence in the multimedia area. So its name (and path) would be comp.multimedia.authoring. Physically the files for the group need to be put into the news system, of which more in a moment. Alternatively it is possible to set up a newsgroup without a vote in the alt (for alternative) hierarchy. This is one place where the Internet has its 'wrong side of the tracks' with postings of extreme politics, sexual content and pirate software alongside the legitimate stuff.

ISPs will usually offer newsgroups to their customers. This means that they have a server which has access to the newsgroup files in their various hierarchies. A user then needs a newsreader, a client that supports the news

protocol NNTP (similar to the mail protocol SMTP). This may be a stand-alone piece of software or it may be built into the web browser or e-mail client. News postings, like mail, can only include ASCII text so any posting of binary files (to the often-notorious alt.binaries groups) needs a trans-coding program to turn the binary into text. We shouldn't give the impression that the alt newsgroups are all 'dodgy', many are just as good/useful as the 'legitimate' ones, perhaps just a bit more quirky.

ISPs and other Internet hosts share news groups with each other. In this way a message posted to a group in one part of the Internet will soon migrate around the whole Internet as the various news servers swap files. This assumes that the whole news group file-set is made available. Some ISPs, and even some countries, censor the news groups for reasons of taste or, sometimes, for practical business reasons. Some ISPs do not give access to the alt hierarchy. Not all news groups are available free so sometimes a set will not be available simply because the ISP doesn't want to pay for them.

Recent legal precedents (including one in the UK) mean that an ISP has a responsibility for newsgroup content even if the content originates elsewhere. This means that if, as in the UK case, potentially libellous material is posted then the ISP should remove it once they are notified or find it. Sad to say, the use of Nicks (nicknames) by chat, forum and discussion group users seems to occasionally encourage aggressive and even obscene and offensive verbal behaviour (particularly from young males). An article on chat rooms in the UK's *Guardian* newspaper commented that postings on sites dominated by male posters 'often tend to read like Monty Python's famous "Argument" sketch, only less funny'. If you have the equivalent of a chat room, discussion group or forum on a website you are responsible for then you should monitor it. An ISP with thousands of news groups is faced with a formidable task. Some ISPs are now removing potentially troublesome groups rather than risk repercussions.

It is possible to use news servers elsewhere than with your own ISP and some news content is available on the web – through www.dejanews.com for example (now acquired by Google) or at newsone.net.

The Web – HTTP

I've left the web protocol, Hypertext Transfer Protocol (HTTP), until last because it is a very sophisticated and complex protocol. You will be familiar with the letters because a web URL always starts by defining the protocol – you can see file://, ftp:// and even gopher:// but usually it is http:// – followed by the domain name or IP of the machine. The domain usually starts www but it doesn't have to.

To access a web page, the web client (usually a browser but not necessarily, it could be a robot indexing the web pages) sends an HTTP request to the web server and this request is in the form of a string of letters

and numbers. Although we often refer to a machine as being a server, the server is really a program running on that machine. This request to the server has several sections to it including information about the requesting machine, the web page containing the link clicked on to generate the request (if any – known as the referrer), which language is expected (sent as a two-letter code such as 'en' for English) and the version of HTTP being used (sent as a number). A language request can be used by the server to customize the response by sending a web page in the appropriate language.

It is possible for the HTTP request to contain extra information generated by the browser based on a form that the user fills in. This information might be hidden (a form of message called POST) or it might appear in the URL sent back to the server as a string following a question mark in the URL (a form called GET). The server will have to respond appropriately to the POST or GET, usually by passing the information to a CGI program running on the server.

Any file sent by the server to the browser will have a header which will identify the type of file it is. Not all browsers react to this information and may take more notice of the suffix (file-type) of the file to see whether it is a GIF or JPEG, text or HTML. But it is possible for a header to be wrong and in that case the browser may not display the file or may display it incorrectly. This is one cause of a web page which displays correctly in one browser but not in another. Servers need to know how to deal with different types of files. There are standard MIME types – such as image/gif and text/html – and if new types of file are used the server needs to be told. Otherwise it might send the file labelled as text. Similarly the browser needs to know what to do with MIME types it is sent. Unknown ones will usually be downloaded as files.

Each HTTP request in a web page is independent. If you have a page containing five graphic files and a video then the server will receive seven separate HTTP requests, including the page's HTML file. Those parts of a web page don't have to be on the same server and banner ads in web pages are usually served from elsewhere. A browser has a limit to the number of simultaneous HTTP requests it can handle, usually eight. If a page requires more than this number than the browser has to wait for individual files to finish transferring before it can ask for another. If you have a web page with more than eight components the visitor will usually see parts arriving at different times. You basically have no control over the order in which things arrive at the browser although in principle the browser starts at the top of the web page and works its way down. Unfortunately, network delays may frustrate that and things held in a table will often remain invisible until the whole table is displayed.

The HTTP protocol has been extended in order to make transfer of files more efficient. If both browser and server support it there is now the concept of a visitor's session where the link is kept open. It is known as 'persistent'. This will reduce the overhead in accessing the web pages.

Security

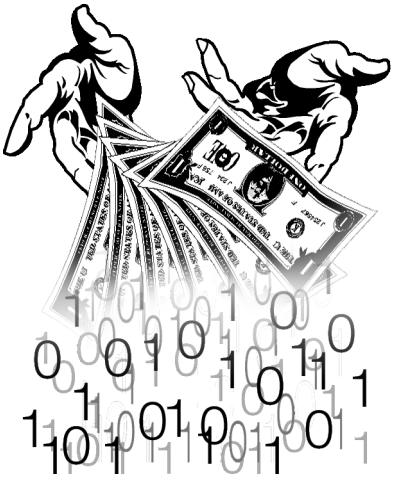
When the protocol is HTTPS this means that a secure server is requested. If one exists at the domain then the server sends a certificate string to the browser which can then decide whether to set up a secure link or to ask the user if they want to take a chance. Other than encrypting the messages passed back and forth, HTTPS does not do anything else and any web pages can be served using HTTPS instead of HTTP. But the encryption and decryption add an overhead to the process and slow down the serving.

To set up a secure server requires two things. First is HTTPS server software, which probably comes built-in to your web server and will then allow you to run HTTPS as well as HTTP on the same website.

Second, you need a certificate from one of the recognized issuing companies like Verisign. They will check your bona fides before issuing the certificate. (This will include verifying the existence of your company, so be sure you have the form of the company name correct.) Without a certificate your secure server will not work: you may not even be able to run the software. As you might expect, the process of issuing the certificate can take some time, although the application can be done online. There is a mechanism for using temporary certificates for testing, but they have limitations, the most significant being that they will only allow a connection to the server from a known machine address. This stops temporary certificates being used for real transactions.

The public face of e-commerce is the ability to take orders online. This usually involves taking people's credit card details. With a secure server you can transfer the credit card information from the customer to your server. But what do you do with it then?

In a sophisticated e-commerce set-up your website will be linked to a banking system that will verify the card information and give you an author-



ization. Based on that authorization you can send the goods and the bank will send you the money.

A simpler set-up can involve the server taking the information and then either saving it locally – somewhere safe from prying Internet eyes – or re-encrypting it and e-mailing the order details to you. A server-side program can easily be written to do this, possibly in Perl. You can then process the transaction as a standard 'Customer not Present' transaction as if they had telephoned you. This will only work if you have a merchant account and can take cards but your merchant will want to know if you intend taking cards over the Internet and may impose conditions even if you already have a standard merchant account.

There is another aspect of security and that is the risk of your website being damaged by an unauthorized person changing files on the server. This is hacking. It has become common for hackers to change the home pages of well-known sites, sometimes to make a political point. Your access to the web server to change pages will probably involve logging on remotely and sending files over by FTP. As in any access to a remote computer, care must be taken with the user names and passwords that you need when you log on. Passwords should not be easy to guess and should not be words from a dictionary – it is possible to reverse-engineer passwords by reference to a dictionary – and should not be written on a sticky note on the computer monitor. The higher the profile of the website the greater the risk of a hacking attack. This risk is not limited to websites, if your own computer is not protected by a firewall then you may find someone will try to hack into it. Although public websites have to be accessible to the outside world they can still be placed behind a firewall and this firewall can be configured to allow only web requests. That will reduce the hacking risk but it will also mean that you have to be on the same side of the firewall as the server to update files on it.

Most web servers are vulnerable to hacking but servers running on Apple Macintosh computers (prior to system X) are reputed to be hacker-proof and that is why the US Army uses a Mac web server. But you should never say 'hacker-proof' – it just makes the challenge greater. Cliff Stoll's book *The Cuckoo's Egg* documents his battle to track down a hacker who was breaking into US military computers back in the early days of the Internet.

XML – separating style from content

When the Web started, HTML (hypertext mark-up language) was a mark-up language which specified the structure of a document by means of 'tags' – codes which define what was the title, what the main headings were, what kind of lists there were and so on. In a simple way, the look of the page was down to the reader and the way the browser was configured. Before long, designers wanted to be able to control the layout of pages so that the page as seen could be defined and everyone would see the same thing. To achieve

this, tags could define fonts and typeface sizes and colours. Tables, or positioning of layers using DHTML (D for Dynamic) could be used to say where things should be. A large website would be difficult to define if you had to put style tags in every paragraph. Cascading Style Sheets were introduced to allow sections of a document, or individual HTML tags, to be given their specifications. All this was defining the appearance of the document often at the expense of its structure since it became increasingly easy to dispense with structural tags like headers completely.

The structure struck back with XML – eXtensible Markup Language. The author of a document can define any XML tags to define the document structure and a style sheet is used to say how the page will then be laid out. Since the latest browsers will render XML as well as HTML it is likely that you have seen XML web pages without realizing it. As you might expect, the file extension is .xml but there is a half-way house. XHTML is a strict version of HTML written in XML.

You can define a completely new mark-up language in XML if you want to. A Document Type Definition (DTD) can be written to specify what you have devised and there are many XML languages around including SMIL (pronounced 'smile'), the Synchronized Multimedia Integration Language.

In the long term, XML is a very important development for the Web because it provides a way of marking a document which can then be served in appropriate formats for the various delivery media such as iTV, mobile and the Web. Since elements within the document can be tagged with meaningful names such as <author> it will be possible to carry out much more sophisticated searches than can be done at present. Currently we are only searching 'blindly' on the text itself without being able to specify the actual meaning of the word.

Network architecture

The Web can function quite simply with a connection between a server and a browser (the client in the client–server architecture). But the Internet can get clogged up, speeds can vary and some websites might attract more hits than a single server can happily handle.

You can insert a proxy server between the client and the server. In its most usual form this proxy server will receive all requests for web pages and will pass on those requests to the real server, keeping a copy of the web page as it goes through. The next time the page is requested then the page is served from the proxy rather than the real server, and this is usually a faster process although it runs the risk of the proxy holding an out-of-date copy of the page. This will also, incidentally, disguise the identity of the client machine, since the distant server will see the proxy as making the request.

To handle large numbers of requests for data, a load balancing arrangement might be used in conjunction with a number of servers, each with

identical contents. The load balancer is another computer which monitors the traffic coming to the site and also monitors how busy the servers are. Since each request for an item on a website is independent of all the others, it doesn't actually matter that the component parts might be served by different machines. To the client, the load balanced system looks like a single server.

You might wish to hide your computers from the Internet, and make it difficult (hopefully impossible) for someone to hack their way in. Here you would use a firewall. This is a computer that monitors the traffic coming in from the Internet and controls the access. It might be configured to allow only certain kinds of requests, for certain ports on the computer: port 80 for web pages or port 53 for DNS requests and so on. It might only allow access from certain other computers. It might make the whole network appear on the Internet as if it were a single 'virtual' computer. This is known as Network Address Translation (NAT) and would be done by having a router provide the bridge between the local network and the Internet at large. The router/firewall could be a computer running a program but is more likely to be a dedicated box which is remotely configured using TelNet or a web browser.

Who am I?

While we're on the subject of security, let's consider what the server knows about you when you access a web page (or, from the point of view of the management, just what can you find out about your visitors?).

As we saw earlier in this chapter, when a browser makes a request for a page it sends a string to the server which tells it such things as: the computer platform and operating system, the language preference, what browser it is, the web page which contains the link (if any) that was clicked to get here, and the requesting computer's IP address. This IP address can be mapped back to its domain name by what is called a reverse DNS look-up.

Actually, you may not really see the IP address of the requesting computer, because if there's a proxy server or firewall or router in the way that is what you will see instead. It is not straightforward to identify exactly who is making a request of the server. Since each request is completely separate, it was necessary to invent another method to keep track of visitors: cookies. A cookie is a short string of text that the server sends to the browser and which it can check on later. A visitor can be identified from a cookie to track such things as user preferences, identity of a shopping basket or validity of a password. (More recently, the extension to HTTP allowing a user session doesn't require cookies to keep track and incidentally also provides for a more secure login where the user's name and password are not actually sent over the Internet to the server at all and so cannot be intercepted.) However, browsers have limited memory in which to store cookies so you

need to be a bit careful in sending them. Some users don't trust them and many check to see what is being sent. If your website is sending cookies, don't forget that the user has certainly visited other sites with cookies recently and you might also get cookies sent by embedded content on your page such as advertisements and counters. The user might also refuse to accept the cookie or might delete it later.

Cookies can be tracked across more than one site. If they are attached to banner ads then it is possible to pick them up as you move from page to page with these ads on the pages. The banner ads are possibly provided by a central advertising agency and the cookies associated with these ads can be used to track which sites you have visited. In turn, this information can be used to target the advertising you are shown by the server. While this customer information is used to target advertising it seems pretty innocuous but if it gets associated with personal data then you could be heading for a data protection problem at worst or, at least, dissatisfied customers who don't want their online habits tracked like this.

However, you may often find that your client wants to track customers as part of their customer relations management (CRM) so that the relationship can be made more personal. With this a customer's experience when visiting the website can be more tailored and targeted. This might involve a log-in process or it might involve tracking using cookies or both. A full CRM system builds a database with a record for each customer or even site visitor but it is possible to do some simple CRM without actually storing anything on the web server. A cookie can be used to store visit parameters on the visitor's machine. These could be preferences such as what style sheet the customer wants for the site. The cookie could even store information about parts of the site visited so that you can offer to show a visitor something new. By doing this in cookies you don't actually need to have a database on the site although you rely on the user keeping the cookies and coming in from the same browser each time. If you do have a database-driven CRM system on the server then you might find there are data protection implications. (Data protection is discussed more fully in the rights chapter in Book 1 Chapter 15.)

When surfers leave one site by following a link and arrive somewhere else, on this surfing journey of theirs, one bit of their baggage is the URL of the page that they just left. This information ends up in the log of the next website, assuming that the referring page information is logged. This information tells you who is linking to your site and, if the visitor has come from a search engine, there will probably be a list of the search terms as well.

There may be a downside to this. If you check your logs to find that a referring page has the URL www.biggestrival.com/takeover/prospects.html then you might be in line for a surprise. You should check your website to make sure that referrers don't carry any unwelcome message from you. (Web consultant Russ Haynal has a website at http://navigators.com/cgi-bin/navigators/persona.pl where he discusses this question of your 'persona'.)

Logs

The information in the browser's request is the basis of logging and the analysis of logs is one way to see how successful a website is. The logs are generated by the server and, as mentioned earlier, one of the things to set up when configuring a server is exactly what is wanted in the logs.

Besides the things such as IP address, browser and operating system – which tells you about the surfer's computer – and referrer information – which tells you where the surfer came from – the logs will record information about the visit to your site. The URL requested will be recorded and this will be every page and every graphic and every other kind of file.

Here's a line from one of my logs:

Let's look at the information in this line bit by bit.

hse-ottawa-ppp158445.sympatico.ca
This is the fully qualified domain name of the machine making the request. In this case the log file is not configured to give an IP address if it can resolve a domain name but sometimes you see an IP address here, not a name. This is not necessarily the user's machine of course since there could be a firewall or proxy in the way. We can guess that this is a dial-up user in Ottawa and the ppp suggests a dial-up connection being allocated an IP and domain name on the fly by the ISP which is Sympatico who describe themselves as Canada's national Internet service. PPP – Point to Point Protocol – is commonly used in dial-up connections.

10/Mar/2001:03:43:59 +0000
The date and time are next, as determined from the server clock at my ISP, not the user's. The +0000 at the end tells me that the clock is set to GMT.

GET /ilight/images/ioffice.gif HTTP/1.1
A GET request is one for the whole file, not just its header information (which would be HEAD) and in this case the file path is shown relative to the root of the website. The version of HTTP being used is 1.1 which means that my server should respond using the same version, which allows persistent connections. If my server only supported HTTP 1.0 then it could use this since the system is backwards compatible.

200 5494
Finally the server records a transaction code of 200 which means the file was sent correctly with a transfer of 5494 bytes. The notorious 404 transaction code means the web page requested was not found and the user would have seen an error. If the code had been 304 and no data had been transferred then this would mean that the file was already held locally in the browser or proxy cache and had not been updated, so it was not sent again. You would not usually expect 304 responses while a visitor moved around the site during a single visit, revisiting menu pages, because the browser would normally not request any particular page from the server more than once in a session unless the cache filled up. If you saw a lot of 304 codes this might indicate frequent visitors returning in different surfing sessions who are not seeing new content. Time to update your website!

The particular log line above is a pared-down one from an ISP. Since adding user agent and referrer information can significantly increase the size of the log file many ISPs are reluctant to include them. Unfortunately this information is very useful to you since it tells you what hardware and software your visitors have, which helps you design your site's features to suit the visitors. It also tells you who is linking to you, which is an indication of peer review since people tend to link to sites which they think are good or useful. Of course a link to your site from 'Web Pages that Suck' needs to be taken with a sense of humour and possibly a review of your design.

In general, people don't look at individual lines of log files, they run the whole file through a log analysis tool. This is a piece of software that reads the lines of the log and produces an analysis based on criteria you set up.

There are the basics such as how many times a page was viewed (called 'page impressions') but with a little more sophisticated analysis you can see how visitors move through the site and how long they stay (called 'stickiness'). Finding where they go when they leave is more difficult. One tactic is to pass every link out of your site through a CGI routine that logs the event since the standard server logging can't do this.

There are many log analysis tools, ranging in price from free upwards. Some of them work by reading log files downloaded from the site onto a local machine and others can be run on the server so that the resulting analyses can be viewed as web pages.

Search engines and spiders

When a website has been set up and is running, it needs to be promoted. Unfortunately, the best way of promoting a web URL seems to be putting it onto the end of a television advertisement (or some other conventional method) but for the rest of us a more realistic option is needed.

The usual mechanisms for promoting websites apply whether the site is a commercial one or a hobby home page: search engines, cross-linking and portals. For those of us with money, listings can be bought.

Search engines work by following links through the Web and indexing what they find along the way. They will often start from links submitted by site-builders and surfers, and work outwards. Since the average website links to many others this can be a fruitful process. Getting a search engine spider (as its roving agent is known) is not the difficult bit. What the spider finds and how it indexes your site is crucial to where in a particular search you will end up. Near, or at, the top of the list is best. But what search terms will people be using?

If you are the only supplier of thermoincabulators in the world (or at least on the Web) then a search on thermoincabulators will find you. But if you are a supplier of CDs you will have more competition.

Spiders will show up in your logs, and there are well-known ones associated with the well-known search engines. A file called robots.txt should be placed in the root of the website, where the home page is. This can be used to configure the spiders' searches and to tell them where they are not allowed to go to index. You can even turn away particular spiders by name, or turn them all away. This assumes the spiders obey robots.txt, but they usually do.

Different search engines use different criteria to rank websites but as a general rule you can improve your chances by following a few sensible guidelines. Firstly make use of the meta tags in the web page header to set up a description of the page and its keywords. If you have a keyword or phrase that you want to hit high in searches then use this in the page title, description, keyword list and in the first paragraph of your text on the page. Ideally include it in an <H1> header tag as the main title. Don't use the phrase a large number of times in a particular way because this can be counterproductive and if done to excess is usually discounted by the indexing.

Not every search engine will index based on all these places. Some will not go further than the <HEAD> section. If the engine only reads the <BODY> then you will have a problem if your page consists solely of a Flash animation and no HTML. Spiders can't read Flash. Framed home pages might similarly fool some spiders – in which case the <NOFRAMES> content can be useful.

Search engine sites usually have a page explaining how they index pages so you can fine-tune pages to suit, and there is a useful website at www.searchenginewatch.com as well. If you want to be extra clever then

one possible tactic is to use the server to detect the spider as it arrives and direct it to a special version of a home page especially structured to milk the sympathy of the indexing in question. You can identify the spider because it should identify itself by name either in its domain name or from what is called the User Agent string and this can be read by your program from the incoming HTTP request. You'd do this differently for each spider based on its indexing criteria.

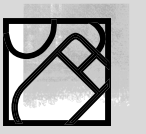
The dot.com bubble

The business perspective on the Internet is bound to be coloured by the rises and falls in the fortunes of so-called dot.com companies. The boom in Internet shares will probably be remembered by historians in the same way as the so-called South Sea Bubble scandal of 1720 (when the South Seas Company also saw its share value rocket without ever making a profit, and took lots of money from British investors until the bubble burst) and the Wall Street Crash. This all risks missing the point about what the Internet means to business as a whole, rather than just to companies who exist only in cyberspace. The Internet may itself change and evolve, but the genie has been let out of the bottle and the way we all do business and communicate is changed forever.

For those of us working at the coal face, producing the web pages, e-databases and e-everything that modern companies need, the changes in cyber-fortunes change our business too. Website makeovers may become less frequent. Design of websites will be governed less by the laws of cool and more by the sensibilities of usability and communication. We will work with structured information across a huge variety of devices. Perhaps more important is the international growth of the Internet. It may be reaching saturation in the USA and other English-speaking western markets but there are still millions of people speaking hundreds of other languages who have yet to be reached.

Technically the Internet changes over time. New IP numbering is under discussion, new top level domains are being introduced, protocols are re-defined and updated. A parallel educational high-speed network, known as Internet2, has been set up. End-users of the Internet may be surfing over telephone lines, broadband connections, televisions, mobile telephones and satellites. The only constant factor is that the Internet has changed the way we make use of information for ever.

THEORY INTO PRACTICE 2



Assuming you are non-technical and you work for a company, tap into the expertise of your technical expert. It should be evident from this chapter that many difficulties can arise from Internet related issues.

1. Ask about any e-mail or FTP problems that has given your company problems.
2. How is your own system set up? Where are the A records kept? What domain names and IP numbers does the company have? What security measures are in place?
3. What are the worst web page problems that they have encountered – what caused them and how were they resolved?
4. What log analysis is kept on your own site? If applicable, what extras have been needed for client sites?
5. Ask them to describe the more major incidents that they have had to deal with on clients' projects, what caused them and how they might be avoided in the future.

Assess whether you feel more confident about discussions of this type now you have a grasp of the background issues.

Summary

- Dial-up and always-on are the two common connection methods to the Internet.
- To set up a website you need to have a domain name and a computer with an IP address.
- Web pages can be static or dynamic. Dynamic pages – or pieces of pages – are built on-the-fly from one or more sources.
- Scalability of websites needs to be planned and tested.
- E-mails are relayed – FTP file transfers and web page requests are sent point-to-point in real time.
- Websites can be managed remotely using FTP.
- Security issues concern credit cards, personal data and hacker protection.
- XML allows document authors flexibility in delivery media by separating style from content.
- Log analysis provides a range of information about the use and users of a website.
- Site recognition by search engine indexing spiders can be enhanced to your advantage.

Recommended reading

Stoll C. (2000). *The Cuckoo's Egg*, New York, NY: Pocket Books.
For a view of life in the 'dark ages' of the Internet, you should read Cliff Stoll's book. It tells of his detective work in tracking down a hacker who was trying to get into US military computers. Along with a ripping good yarn, Cliff explains how the Internet was connected up.

Wallace P. (1999). *The Psychology of the Internet*. Cambridge: Cambridge University Press

Sebesta R.W. (2001). *Programming the World Wide Web*. Reading, MA: Addison-Wesley

Tittel E., Mikula N. and Chandak R. (1998). *XML for Dummies*. Foster City: IDG Books

Albitz P. and Liu C. (1998) *DNS and Bind*, 3rd edn. Sebastopol, CA: O'Reilly & Associates
This book is aimed at system administrators and network techies, but it gives detailed information on how DNS works.

Spainhour S. and Echstein R. (1999). *Webmaster in a Nutshell*, 2nd edn. Sebastopol, CA: O'Reilly & Associates

Niederst J. (1999). *Web Design in a Nutshell*, Sebastopol, CA: O'Reilly
O'Reilly publishes a wide range of useful reference books on all aspects of programming. The author makes extensive use of their HTML, Perl and JavaScript books.

The Internet Society has a page of links to various historic documents at the logically named:
http://www.isoc.org/internet/history

The NetGeo server (http://www.caida.org/tools/utilities/netgeo) maps IP addresses to latitude and longitude or to country. The system queries WHOIS servers and parses the information there in order to obtain its data. A more detailed abstract is at http://www.caida.org/outreach/papers/inet_netgeo.html

The IP Network Index (http://www.ipindex.net) can be used to see to whom a particular IP address is registered. This will usually resolve to an ISP if the end-user is a person or small company.

The Internet Assigned Numbers Authority has a list of country top-level domains and information about who adminsters them at
http://www.iana.org.cctld-whois.htm
Usually this is logically geographic but sometimes, as in the .tv domain, it has become vertically commercial.

Information from IANA on the generic TLDs such as .com and .edu is at
http://www.iana.org/gtld/gtld.htm

The All Whois website (htttp://www.allwhois.com) enables you to look up information on domains to see who registered them and (sometimes) who owns them.

Anonymizer (http://www.anonymizer.com) claims to allow anonymous web browsing and will presumably make non-Americans seem to be in the USA since the domain name or IP will be for anonymizer.com rather than the actual user.

Internet Product Watch (http://ipw.internet.com) provides news, specifications, information and announcements about the latest Internet hardware and software. (Note there is no www in the URL.)
http:///www.internetproductwatch.com also gets you there.

Russ Haynall's 'Check your Persona' web page is at
http://navigators.com/cgi-bin/navigators/persona.pl – again, no www.

Vincent Flanders' site on web page design that doesn't work is at
http://www.webpagesthatsuck.com

Security Space Network Monitor is at
http://www.securityspace.com/s_survey/data

One freely-available stress tester is a Java-based project from Apache called Jakarta at
http://jakarta.apache.org/jmeter

Guardian newspaper (www:/guardianunlimited.co.uk) – online supplement appears each Thursday. The article on forums was written by David Rowley and appeared on 19 April 2001.