Chapter III

Subject Gateway Vs Search Engines
SUBJECT GATEWAYS VS SEARCH ENGINES

3.1. History of the Internet

The Internet has revolutionized the computer and communications world like nothing before. The invention of the telegraph, telephone, radio, and computer set the stage for this unprecedented integration of capabilities. The Internet is at once a world-wide broadcasting capability, a mechanism for information dissemination, and a medium for collaboration and interaction between individuals and their computers without regard for geographic location.

The Internet represents one of the most successful examples of the benefits of sustained investment and commitment to research and development of information infrastructure. Beginning with the early research in packet switching, the government, industry and academia have been partners in evolving and deploying this exciting new technology. Today, terms like "www.amazon.com" and "http://www.acm.org" trip lightly off the tongue of the random person on the street.

The Internet was the result of some visionary thinking by people in the early 1960s who saw great potential value in allowing computers to share information on research and development in scientific and military fields. The first recorded description of the social interactions through networking was a series of memos written by J.C.R. Licklider of MIT in August 1962 discussing his "Galactic Network" concept. He envisioned a globally interconnected set of computers through which everyone could quickly access data and programs from any site. In spirit, the concept was very much like the Internet of today. Licklider was the first head of the computer research program at DARPA, starting in October 1962.
Leonard Kleinrock of MIT and later UCLA developed the theory of packet switching, which was to form the basis of Internet connections. Lawrence Roberts of MIT connected a Massachusetts computer with a California computer in 1965 over dial-up telephone lines. It showed the feasibility of wide area networking, but also showed that the telephone line's circuit switching was inadequate. Kleinrock's packet switching theory was confirmed. Roberts moved over to DARPA in 1966 and developed his plan for ARPANET. These visionaries and many more left unnamed here are the real founders of the Internet.

The Internet, then known as ARPANET, was brought online in 1969 under a contract let by the renamed Advanced Research Projects Agency (ARPA) which initially connected four major computers at universities in the southwestern US. The Internet was designed in part to provide a communications network that would work even if some of the sites were destroyed by nuclear attack. If the most direct route was not available routers would direct traffic around the network via alternate routes.

The early Internet was used by computer experts, engineers, scientists, and librarians. There was nothing friendly about it. There were no home or office personal computers in those days, and anyone who used it, whether a computer professional or an engineer or scientist or librarian, had to learn to use a very complex system.

E-mail was adapted for ARPANET by Ray Tomlinson of BBN in 1972. He picked the @ symbol from the available symbols on his teletype to link the username and address. The telnet protocol, enabling logging on to a remote computer, was published as a Request for Comments (RFC) in 1972. RFC's are a means of sharing developmental work throughout
community. The FTP protocol, enabling file transfers between Internet sites, was published as an RFC in 1973, and from then on RFC's were available electronically to anyone who had use of the ftp protocol.

Libraries began automating and networking their catalogs in the late 1960s independent from ARPA. The visionary Frederick G. Kilgour of the Ohio College Library Center (now OCLC, Inc.) led networking of Ohio libraries during the '60s and '70s. In the mid 1970s more regional consortia from New England, the Southwest states, and the Middle Atlantic states, etc., joined with Ohio to form a national, later international, network. Automated catalogs, not very user-friendly at first, became available to the world, first through telnet and only many years later, through the web.

The Internet matured in the 70's as a result of the TCP/IP architecture, Usenet was started in 1979 based on Unix to Unix Copy Protocol (UUCP). Newsgroups, which are discussion groups focusing on a topic, followed, providing a means of exchanging information throughout the world. While Usenet is not considered as part of the Internet, since it does not share the use of TCP/IP, it linked UNIX systems around the world, and many Internet sites took advantage of the availability of newsgroups. It was a significant part of the community building that took place on the networks.

Similarly, BITNET (Because It's Time Network) connected IBM mainframes around the educational community and the world to provide mail services beginning in 1981. Listserv software was developed for this network and later others. Gateways were developed to connect BITNET with the Internet and allowed exchange of e-mail, particularly for e-mail discussion lists. These listservs and other forms of e-mail discussion lists
formed another major element in the community building that was taking place.

In 1986, the National Science Foundation funded NSFNet as a cross country 56 Kbps backbone for the Internet. They maintained their sponsorship for nearly a decade, setting rules for its non-commercial government and research uses.

As the commands for e-mail, FTP, and telnet were standardized, it became a lot easier for non-technical people to learn to use the nets. It was not easy by today's standards by any means, but it did open up use of the Internet to many more people in universities in particular. Other departments besides the libraries, computer, physics, and engineering departments found ways to make good use of the nets to communicate with colleagues around the world and to share files and resources.

Since the Internet was initially funded by the government, it was originally limited to research, education, and government uses. Commercial uses were prohibited unless they directly served the goals of research and education. This policy continued until the early 90's, when independent commercial networks began to grow. It then became possible to route traffic across the country from one commercial site to another without passing through the government funded NSFNet Internet backbone.

Delphi was the first national commercial online service to offer Internet access to its subscribers. It opened up an email connection in July 1992 and full Internet service in November 1992. All pretenses of limitations on commercial use disappeared in May 1995 when the National
Science Foundation ended its sponsorship of the Internet backbone, and all traffic relied on commercial networks. Since commercial usage was so widespread by this time and educational institutions had been paying their own way for some time, the loss of NSF funding had no appreciable effect on costs.

The early days of the web was a confused period as many developers tried to put their personal stamp on ways the web should develop. The web was threatened with becoming a mass of unrelated protocols that would require different software for different applications. The visionary Michael Dertouzos of MIT's Laboratory for Computer Sciences persuaded Tim Berners-Lee and others to form the World Wide Web Consortium in 1994 to promote and develop standards for the Web.

The following is a list of types of information that is available on the Internet in digital form.

**Table 1 : Types of information on the Internet**

<table>
<thead>
<tr>
<th>Type</th>
<th>Full text on Internet</th>
<th>How to find</th>
<th>Cost?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference books</td>
<td>Rare</td>
<td>Scattered throughout sites and through universities and libraries</td>
<td>Free</td>
</tr>
<tr>
<td>Encyclopedias</td>
<td>Some</td>
<td>Encyclopedia Britannica is online</td>
<td>Basic is Free</td>
</tr>
<tr>
<td>Electronic popular magazines</td>
<td>Full text through the Electric Library, EBSCO, and many selected articles are on magazine websites</td>
<td>Use search tools inside websites</td>
<td>Free</td>
</tr>
<tr>
<td>Electronic technical journals</td>
<td>Full text through some libraries</td>
<td>Use search tools inside websites</td>
<td>Must be student, employee of large company</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>Newspapers</td>
<td>Many newspaper websites have only same day articles</td>
<td>Search inside newspaper website</td>
<td>Free</td>
</tr>
<tr>
<td>Technical papers</td>
<td>Only a few repositories available through libraries</td>
<td>Search within site</td>
<td>Must be student, employee of large company</td>
</tr>
<tr>
<td>Web articles</td>
<td>The most common type information on the Internet</td>
<td>Use popular search engines</td>
<td>Free</td>
</tr>
<tr>
<td>Electronic books</td>
<td>Rare</td>
<td>Use popular search engines</td>
<td>Typical $10 per book</td>
</tr>
<tr>
<td>Speeches</td>
<td>Quite common</td>
<td>Search databases such as Electric Library</td>
<td>Free</td>
</tr>
<tr>
<td>Discussion groups</td>
<td>Common</td>
<td>Search individual discussion groups or popular search engines</td>
<td>Free</td>
</tr>
<tr>
<td>Databases</td>
<td>Lots</td>
<td>Search individual websites</td>
<td>Paid for by those listing in database</td>
</tr>
<tr>
<td>Company information</td>
<td>Common</td>
<td>Use popular search engines</td>
<td>Free</td>
</tr>
<tr>
<td>Computer files</td>
<td>Common</td>
<td>Search Individual websites</td>
<td>Thousands are free</td>
</tr>
<tr>
<td>Stock analysis</td>
<td>Common</td>
<td>Search Individual websites</td>
<td>General service is free</td>
</tr>
<tr>
<td>Personal ads</td>
<td>&quot;One &amp; Only&quot; predominant</td>
<td>Search Individual websites</td>
<td>Rarely free,</td>
</tr>
<tr>
<td>Classifieds</td>
<td>Lots similar to Buy and Sell</td>
<td>Search Individual websites</td>
<td>delayed service is often free</td>
</tr>
<tr>
<td>Patents</td>
<td>Many countries are going online</td>
<td>Search Individual websites</td>
<td>Free</td>
</tr>
</tbody>
</table>

Source: http://www.benwiens.com/websearch.html
While the number of sites on the Internet was small, it was fairly easy to keep track of the resources of interest that were available. But as more and more universities and organizations—and their libraries—connected, the Internet became harder and harder to track. There was more and more need for tools to index the resources that were available.

The first effort, other than library catalogs, to index the Internet was created in 1989, as Peter Deutsch and his crew at McGill University in Montreal, created an Archiver for ftp sites, which they named Archie. This software would periodically reach out to all known openly available FTP sites, list their files, and build a searchable index of the software. This is the first Search Engine.

2.2. Subject gateways Vs Search Engines

*What is the Search Engine?*

Search engines are the card catalogs of the web. They search for keywords by using "spiders" or "Robots" (electronic software) that crawl through web space from link to link, collecting and indexing the words on millions of web pages and feeding them into searchable databases.

The honour of being the first Internet search engine goes to Archie, a pre-web search application created in 1990 by McGill university student named Alan Emtage. By 1990, academics and technologists were regularly using the internet to store papers, technical specs, and other kinds of documents on machines that were publicly accessible. Unless we had the exact machine address and filename, however it was nearly impossible to find those archives. Archie scoured Internet-based archives and hence the name called "Archie".
Based on the Internet’s FTP standard, Archie’s architecture was similar to most modern web search engines—it crawled sources, built an index, and had a search interface. But the pre-Web era was not a very user-friendly time. Only true techies and academics used Archie, though among that crowd it was quite popular. Typical users would query the engine by connecting directly to an Archie server via a command-line interface. They would query Archie via keywords thought to be in a matching file’s title, and then receive a list of places where a particular matching file could be found. They then connected to that machine, and rummage around till they found what they were looking for. Not particularly robust, but much better than nothing.

The name: “Archie” had a quirky appeal that seemed to fit the young Internet. In 1993, students at the University of Nevada created Veronica, a play on the comic book couple. Veronica worked much as Archie did but substituted gopher, another popular and more fully featured Internet file sharing standard, for FTP.

Veronica moved search a bit closer to what we now expect—the Gopher standard allowed searchers to connect directly to the document queried, as opposed to just the machine on which the document released. Not a huge step, but progress.

Both Archie and Veronica locked semantic abilities but they didn’t index the full text of the document, just the documents title. That meant a searcher had to know or infer the title of the document he or she was looking for. With the rise of the web, Archie and Veronica soon fell out of favour.
As the web took off, so did the basic problem of search. When the internet was the domain of academics and technologists, finding things was a limited problem. But from 1993 to 1996, the web grew from 130 sites to more than 600,000. Watching all this growth was Matthew Gray, a researcher at the Massachusetts Institute of Technology and a pioneer of the earliest web based Search Engine, the WWW Wanderer.

The Wanderer solved a basic problem Gray had noted with the web, namely that it was growing faster than any human could track. The wanderer was a robot that automatically created an index of sites and had another unexpected effect. In the early days of the web, bandwidth was at a premium and many web masters felt the wanderer ate up too many processing and bandwidth cycles as it indexed a site's contents.

The wanderer was soon eclipsed by more powerful engines. One of the first was web crawler, developed by university of Washington researcher Brian Pinkerton while he was working for Steve Job's company. The technologies the company developed—built-in Ethernet, high quality colour. WebCrawler was important to the evolution of search because it was the first to index the full text of the web documents it found. WebCrawler opened up a new universe for web surfers, particularly its full text search and simple browser-based interface was an important step toward making the web fit for mainstream consumption, beyond academics and tech Greeks.

The First truly Good Search Engine

When the internet was young, when the web comprised less than 10 million pages, when the yaho was a funky set of links and "google" was just a common misspelling for a very large number, Louis Monier put the
entire web on a simple computer. Louis Monier, a researcher at DEC's western Lab in Palo Alto, California, suggested building a search engine: it could load the entire internet onto an Alpha computer, then build a program showcasing Alpha’s speed—Alta Vista was born a proof point to DEC’s hardware dominance.

As is true for much of the IT industry, nearly every well-known company in search can trace its roots to a University, the kind of institution that allows big ideas to flourish without the strait jacket of commercial demands. Google, Excite and Yahoo emerged from Stanford; Inktomi, Berkeley; and Lycos came from Carnegie Mellon.

Even so often a great innovation will spring not from University, but from within a corporation. A few technology companies understand nurture the ethos of academic research—open inquiry, freedom to fail, research without resource constraints and open collaboration.

But not many companies can afford the luxury of pure research labs and even fewer have the foresight and long term vision to create them. In the late 1980’s, DEC was among the few IT giants making a long-term investment in pure research.

*Alta Vista Salient Features*

Yet Altavista is remarkable for a number of reasons. To borrow from the present, Altavista was the Google of its era. In 1996 it was arguably the best and most-loved brand on the web. It presaged many of the current innovations and opportunities in search from automatic language translation to audio and video search to clustering of results.
Altavista was an extraordinary success item in its original business but ultimately failed because of hide bound management.

**Rise of the Big Guys: Lycos, Yahoo, Excite**

By 1995 several other major web destinations had framed including Lycos, was created in May 1994 by Carnegie Mellon University (CMU) project. CMU’s Dr. Michael Mauldin working under a grant from the defence Advanced Research Projects Agency (DARPA). The name was derived from Lycosidae the Latin word for the wolf spider family, whose members actively seek their prey rather than catching it in a web. Like its predecessors, Lycos deployed a spider like crawler to index the web, but it used more sophisticated mathematical algorithms to determine the meaning of a page and answer user queries. And it became the first major engine to use links to a web site as the basis of relevance the underlying basis for Google’s current success. For a short period in 1999 Lycos became the most popular online destination in the world. Today Lycos remains a top-twenty destination, but it has struggled to regain its past glories in light of the extraordinary success of Google.

**Excite**

Founded in 1994 by six Stanford University alumni all tight friends. Excite was the first search engine to transcend classic keyword-based searching with technology that grouped web pages by their underlying concepts. It used statistic analysis of word relationships on the page to deliver fine-tuned results to web surfers.

One of its most persistent innovations was personalization – MyExcite was among the first services to allow users to create custom web pages with news, business information and regional weather reports. And
in the summer of 1997, Excite became the first of the major portals to offer free e-mail—a move that rivals yahoo and Lycos would make that October.

Yahoo

Yahoo got its start when two bored PhD candidates at Stanford hacked together a system that helped them win a fantasy basketball league. Jerry Yang and David Filo were both pursuing doctorates in electronic design automation, a once hot field that had cooled by the fourth year of their doctoral work. Srinija Srinivasan who joined Yahoo in 1995 as editor in Chief says,” The shift from exploration and discovery to the intent based search of today was inconceivable. Now, we go online expecting everything we want to find will be there. That is a major shift”.

Another reason Yahoo succeeded was its sense of fun—a characteristic that would come to define not only Yahoo, but nearly every Internet company seeking the fickle approval of the Web public. Yahoo pioneered some of the Web’s earliest social mores—including, for example, links to competitor’s sites in case a searcher could not find what he or she was looking for, and listing “what’s hot” prominently on its home page, thereby driving extraordinary amounts of traffic to otherwise obscure sites.

Google

Larry Page and Sergey Brin two more Stanford PhD candidates created ranking system rewarding links that came from sources that were important and penalizing those that did not.
Table 2: Growth of Search Engines

<table>
<thead>
<tr>
<th>S.No</th>
<th>Name of the Search Engines</th>
<th>Year of Establishment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Archie</td>
<td>1990</td>
</tr>
<tr>
<td>2.</td>
<td>Veronica</td>
<td>1993</td>
</tr>
<tr>
<td>3.</td>
<td>Lycos</td>
<td>1994</td>
</tr>
<tr>
<td>4.</td>
<td>Excite</td>
<td>1994</td>
</tr>
<tr>
<td>5.</td>
<td>Yahoo</td>
<td>1995</td>
</tr>
<tr>
<td>6.</td>
<td>Altavista</td>
<td>1996</td>
</tr>
<tr>
<td>7.</td>
<td>Google</td>
<td>1998</td>
</tr>
</tbody>
</table>

How do Search Engines Work?

Search Engines for the general web do not really search the World Wide Web directly. Each one searches a database of the full text of web pages selected from the billions of web pages out there residing on servers. When we search the web using a search engine, we are always searching a somewhat stale copy of the real web page. When we click on links provided in a search engine's search results, we retrieve from the server the current version of the page.

Search engine databases are selected and built by computer robot programs called spiders. Although it is said they "crawl" the web in their hunt for pages to include, in truth they stay in one place. They find the pages for potential inclusion by following the links in the pages they already have in their database (i.e., already "know about"). They cannot think or type a URL or use judgment to "decide" to go look something up and see what's on the web about it. (Computers are getting more sophisticated all the time, but they are still brainless.)
If a web page is never linked to in any other page, search engine spiders cannot find it. The only way a brand new page - one that no other page has ever linked to - can get into a search engine is for its URL to be sent by some human to the search engine companies as a request that the new page be included. All search engine companies offer ways to do this.

After spiders find pages, they pass them on to another computer program for "indexing." This program identifies the text, links, and other content in the page and stores it in the search engine database's files so that the database can be searched by keyword and whatever more advanced approaches are offered, and the page will be found if our search matches its content.

Some types of pages and links are excluded from most search engines by policy. Others are excluded because search engine spiders cannot access them. Pages that are excluded are referred to as the "Invisible Web" -- what we don't see in search engine results. The Invisible Web is estimated to be two to three or more times bigger than the visible web.

**What is subject Gateways?**

The Web is quickly becoming the world's fastest growing repository of data. People are increasingly going to the Internet before they go to the library. Subject gateways "doing for Internet information resources what librarians do for books". An Internet search tool to help people find information on the Internet e.g.: electronic journals software datasets electronic books mailing lists / discussion groups (and their archives) articles / papers / reports bibliographic databases bibliographies organisational home pages educational materials news resource guides.
Cataloguing is all very well for books and collecting information on their storage but they mostly do not give as direct useful access to the information we need. This is the reason why Subject Gateways were set up. A subject gateway, in the context of network-based resource access, can be defined as a facility that allows easier access to network-based resources in a defined subject area. The simplest types of subject gateways are sets of Web pages containing lists of links to resources. Subject gateways are often a good starting point for subject searching on the Internet. They are provided by a range of academic and commercial institutions and function as collections of quality web sites. In the majority of cases sites listed in gateways have been evaluated and assessed for inclusion.

**Evaluation of Search Engines**

The measure of a search engine's ability is in the results it provides and how the search process and results set compare with rival search engines. A true and fair evaluation of a search engine's performance and database is very much dependent on the evaluator being familiar with the subject matter and the search facilities offered by the search engine involved.

Search engines do not index all the documents on the Web, nor do they all index the same sites or in the same fashion.

Observation generally noted – more resources are put into adding new documents than confirming the validity of those already in the database, one of the consequences being the existence of dead and out-of-date links in search results. This is the classic tradeoff between completeness and accuracy.
A good search robot combined with full-text indexing but poor retrieval software will give rise to high recall and low precision. "The main criticism of automated tools is that they tend to overload users with irrelevant, misleading results". Documents which "share the same relevant words but not necessarily shared relevant context ... (can be) displayed next to each other in the results". (Jenkins C, 1998)

**Why gateways better than search engines?**

Search engines, like the Web, are ever changing. What holds true today for a search engine may not do so in the future. A judgment at a point in time may not hold true for very long in the changing environment that is the World Wide Web.

Using a subject gateway instead of a general search engine can result in the return of more genuine and relevant web pages from our search. If we do wish to use a search engine, then we will have to think more carefully about appraising the information we retrieve from the web before we can rely on it as a basis for our study or research.

Subject gateways are portals to web pages in a particular subject area. They usually provide a search interface which allows us to find relevant web pages. An added advantage is that they usually link to web pages which are of reliable quality.
<table>
<thead>
<tr>
<th>Google</th>
<th>OMNI</th>
</tr>
</thead>
</table>
| 1. Epilepsy is a common chronic neurological disorder that is characterized by recurrent unprovoked seizures. [1] [2] These seizures are transient signs and/or ... | Surgery for epilepsy  
This leaflet, aimed at parents, provides information about epilepsy, and how it can be helped by surgery. It explains who would be considered for surgery, tests and scans that would be required, types of operation, risks, the success rate, and follow up treatment. Published on the Web by Great Ormond Street Children's Hospital (GOSH) and the Institute of Child Health (ICH). |
| 2. In-depth epilepsy and seizure information for patients, families, and the professionals who care for them. | Improving services for people with epilepsy: Department of Health action plan in response to the national clinical audit of epilepsy-related death  
This document Improving services for people with epilepsy: Department of Health action plan in response to the National Clinical Audit of epilepsy-related death was published by the Department of Health (DH) in February 2003. This action plans aims to improve standards of care and support for people with epilepsy and their families. It is hoped that this will be the start of a process of improving awareness and understanding of Sudden Unexpected Death in Epilepsy (SUDEP). The three main sections of this document cover pathology and post mortem investigations; improving care, management and treatment of epilepsy; and improvements in the provision of information.
### 3. Epilepsy

Epilepsy is a brain disorder that causes people to have recurring seizures. The seizures happen when clusters of nerve cells, or neurons, in the brain send ...

<table>
<thead>
<tr>
<th>Diagnosis and management of epilepsy in adults</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis and Management of Epilepsy in Adults is the Scottish Intercollegiate Guidelines Network (SIGN's) 70th guideline, published in April 2003. The Scottish Intercollegiate Guidelines Network (SIGN) develops and publishes evidence-based clinical guidelines for use by the health service in Scotland. This guideline provides recommendations on the diagnosis and treatment of epilepsy and covers classification of epileptic seizures; clinical signs and symptoms; brain imaging; eeg monitoring; antiepileptic drugs; drug-resistant epilepsy; contraception; pregnancy; HRT; models of care; implementation for audit; recommendations for patients and carers; and outcome measures.</td>
</tr>
</tbody>
</table>

Here, we can see that the resources catalogued in OMNI have been described in a "human readable" fashion, whilst the entries in Google are presented more as "raw data". OMNI points directly to the home page or start point of a resource while Google often points to pages without context, leaving the user to find their own way.

Google records are created by an automatic process and typically consist of a mixture of metadata offered by the author of the page (if this is available) and text picked up from the page itself. In contrast, OMNI records offer information created by a cataloguer, which is designed to highlight the main features of a resource in a easily-readable, concise fashion.
Google indexes individual *pages*, not *resources*. As an illustration of the difference between a page and a resource, consider that an online textbook could consist of many web pages, hyperlinked together via a table of contents. The Google software does not know which set of pages on a server constitute a resource and when it encounters a large collection of pages, is likely to index a random sample. Subject gateways such as OMNI, on the other hand, catalogue at the resource level, and will therefore describe resource composed of many pages in a much more coherent fashion.

Lastly, the resources described in a subject gateway are likely to have been hand-picked and catalogued with a particular audience in mind. The OMNI gateway, for example, includes only biomedical resources of interest to the higher education and research community, and catalogues resources with this in mind. Thus, resources included in OMNI are indexed and classified in a similar way to books in an academic medical library, and are selected so that they are at an appropriate level for students, researchers, lecturers, etc. This tailored and selective approach is not possible for a service such as Google, which successfully serves a much broader community.

3.3. Subject gateways Vs Semantic Web

**Semantic Web**

The **Semantic Web** is a project that intends to create a universal medium for information exchange by putting documents with computer-processable meaning (semantics) on the World Wide Web. The Semantic Web is not about the meaning of English documents. It's not about the artificial intelligence areas of machine learning or natural language understanding -- they use the word semantics with a different meaning.
Currently under the direction of the Web's creator, Tim Berners-Lee of the World Wide Web Consortium, the Semantic Web extends the Web through the use of standards, markup languages and related processing tools. The Semantic Web is designed to smoothly interconnect personal information management, enterprise application integration, and the global sharing of commercial, scientific and cultural data. We are talking about data here, not human documents.

The SW is not at all a threat to existing browsers, that this is adding something to the WWW, not replacing it. The existence of data on the web will not threaten the documents, music pictures, and so on... on the Web. With the standardization and deployment of Semantic Web standards in various commercial products and services, a shift occurred from the perspective of many that this work was research to recognition that this is a practical technology deployed in mass-market tools that enables more flexible access to structured data on the Web.

The Semantic Web is a mesh of information linked up in such a way as to be easily processable by machines, on a global scale. We can think of it as being an efficient way of representing data on the World Wide Web, or as a globally linked database. Data that is generally hidden away in HTML files is often useful in some contexts, but not in others. The problem with the majority of data on the Web that is in this form at the moment, is that it is difficult to use this on a large scale, because there is no global system for publishing data in such a way as it can be easily processed by anyone. For example, just think of information about local sports events, weather information, plane times, Major League Baseball statistics, and television guides... all this information is presented by numerous sites, but all in HTML. The problem with that is that, is some...
contexts, it is difficult to use this data in the ways that one might want to do so.

Most people are capable of using the web to, say, find the Swedish word for "car", renew a library book, or find the cheapest DVD and buy it. But if we ask a computer to do the same thing, it wouldn't know where to start. That is because web pages are designed to be read by humans, not machines. The Semantic Web is a project aimed to make web pages understandable by computers, so that they can search websites and perform actions in a standardized way. Using a markup language called RDF (Resource Description Framework), we can put logical statements like these on the Internet, "spiders" could collect them, and the statements could be searched, analyzed, and processed. Yet another development related to semantic Web is the recent explosion of blogs and syndicated feeds. There are 8 million to 12 million active blogs on the internet and millions more RSS feeds. (Real Simple Syndication)

Perfect search will require more than ubiquity, click streams, and personalization. The vast corpus of information now available to us is often meaningless unless it is somehow tagged-identified in such a way that search engines can best make sense of it and serves it up to us. Many in the search industry believe search will be revolutionized by what is called metadata. Click streams are a form of metadata-information about where we go and what we choose as we browse the Web. But to get more perfect search, we need to create a more intelligent Web. That means tagging the relatively dumb Web pages that make up most of the Web as we know it today with some kind of code that declares, in a machine-readable universal lingo, what they are, what they are capable of doing, and how they might change over time.
In 1998, Berners-Lee’s “Semantic Web Road Map” outlined a universal and relatively simple approach to structuring metadata so that the Web becomes more intelligent. While it’s always dangerous to lean too heavily on metaphor, the basic idea is that with semantic tags, the Web becomes more like a structured database such as Lexis and Nexis or the Sabre reservation system, making it far easier to find things. This is turn allows rules of logic, or reason, into the equation. This structure also makes it much easier to do things, to execute complex tasks built upon finding things—scheduling a meeting, planning a trip, organizing a wedding, etc.

The real power of semantic Web will be realized when people create many programs that collect Web content from diverse sources, process the information exchange the results with other programs. The effectiveness of such software agents will increase exponentially as more machine readable Web content and automated services become available. The Semantic Web promotes this synergy: even agents that were not expressly designed to work together can transfer data among themselves when the data come with semantics.

**What is Semantic Web?**

First, blogs are personal statements by individuals, digital declarations of who they are and who they wish to be in the searchable world. The blog becomes a much nuanced statement of individual’s social standing, relationships, interests, and history.

Second, once blogs reach critical mass, intelligent engines will be able to discern patterns among them that will provide second-and third-
order relevance inputs that will help refine and return far better search results.

The Semantic Web is a web of data. There is lots of data we all use every day, and it’s not part of the web. We can see our bank statements on the web, and our photographs, and we can see our appointments in a calendar. But can we see our photos in a calendar to see what we are doing when we took them? Can we see bank statement lines in a calendar? Why not? Because we don't have a web of data. Because data is controlled by applications, and each application keeps it to itself.

The Semantic Web is about two things. It is about common formats for interchange of data, where on the original Web we only had interchange of documents. Also it is about language for recording how the data relates to real world objects. That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing.

The Principle of Least Power

The Semantic Web works on a principle of least power: the less rules, the better. This means that the Semantic Web is essentially very unconstraining in what it lets one say, and hence it follows that anyone can say anything about anything. When we look at what the Semantic Web is trying to do, it becomes very obvious why this level of power is necessary. If we started constraining people, they wouldn't be able to build a full range of applications, and the Semantic Web would therefore become useless to some people.
The Semantic Web proposes to explicate the meaning of Web resources by annotating them with metadata that have been described in an ontology. This will enable machines to "understand" the meaning of resources on the Web, thereby unleashing the potential for software agents to perform tasks on behalf of humans. Consequently, significant effort in the Semantic Web research community is devoted to the development of machine processible ontology representation formalisms. There are also related efforts in both the academic and commercial communities, which are making available tools for semi-automatic and automatic semantic (ontology-driven and/or domain-specific) metadata extraction and annotation.

These languages (RDQL) offer most of the essential features for semantic querying such as the ability to query using ontological concepts, inferring as part of query answering, and some allow the ability to specify incomplete queries through the use of path expressions. With the progress towards realizing the Semantic Web, the development of semantic query capabilities has become a pertinent research problem. Semantic querying techniques will exploit the semantics of web content to provide superior results than present-day techniques which rely mostly on lexical (e.g. search engines) and structural properties (e.g. XQuery) of a document. There are now a number of proposals for querying RDF data including RQL, SquishQL, TRIPLE.

One key advantage of this last feature is that users do not need to have in-depth knowledge of schema and are not required to specify the exact paths that qualify the desired resource entities. However, even with such expressive capabilities, many of these languages do not adequately support a query paradigm that enables the discovery of complex
relationships between resources. The pervasive querying paradigm offered by these languages is one in which queries are of the form: "Get all entities that are related to resource A via a relationship R" where R is typically specified as possibly a join condition or path expression, etc. In this approach, a query is a specification of which entities (i.e. resources) should be returned in the result. Sometimes the specification describes a relationship that the qualifying entities should have with other entities, e.g. a join expression or a path expression indicating a structural relationship. However, the requirement that such a relationship be specified as part of the query is prohibitive in domains with analytical or investigative tasks such as national/homeland security and business intelligence, where the focus is on trying to uncover obscured relationships or associations between entities and very limited information about the existence and nature of any such relationship is known to the user. In fact, in this scenario the relationship between entities is the subject of the user's query and should be returned as the result of the query as opposed to be specified as part of the query. That is, queries would be of the form "How is Resource A related to Resource B?". For example, a security agent may want to find any relationship between a terrorist act and a terrorist organization or a country known to support such activities.

One major challenge in dealing with queries of this nature is that it is often not clear exactly what notion of a relationship is required in the query. For example, in the context of assessing flight security, the fact that two passengers on the same flight are nationals of a country with known terrorist groups and that they have both recently acquired some flight training, may indicate an association due to a similarity. On the other hand, the fact that a passenger placed a phone call to someone in another country that is known to have links with terrorist organizations and activities may
indicate another type of association characterized by connectivity. Therefore, various notions of "relatedness" should be supported.

**Merits**

The Semantic Web is generally built on syntaxes which use URIs to represent data, usually in triples based structures: i.e. many triples of URI data that can be held in databases, or interchanged on the World Wide Web using a set of particular syntaxes developed especially for the task. These syntaxes are called "Resource Description Framework" syntaxes. So the Semantic Web can be seen as a huge engineering solution... but it is more than that. We will find that as it becomes easier to publish data in a repurpose form, so more people will want to publish data, and there will be a knock-on or domino effect. We may find that a large number of Semantic Web applications can be used for a variety of different tasks, increasing the modularity of applications on the Web, but enough subjective reasoning... onto how this will be accomplished. Critical to the success of the Semantic Web will be the development of software *agents* that can extract content from the web in response to complex queries by *reasoning* upon its explicit, canonical content.

What impact might this have on scientific publishing? In the next few years, we expect that tools for publishing papers on the web will automatically help users to include more of this machine-readable markup in the papers they produce. Whereas current tools using XML (Extensible Markup Language) can allow a user to assert that some part of a document is about an ‘experiment’, the new languages will let the scientist express that how the experiment uses.
Papers that include this new markup language will be found by new and better search engines, and users will thus be able to issue significantly more precise queries. More importantly, experimental results will themselves be published on the web, outside of the context of a research paper. So a scientist could design and run an experiment, and create an emerging web page containing the information that he or she wants to share with trusted colleagues (see Figure:1). Finding out about experiments and studies in progress will be easy, and work will be able to be modified as a result of interaction with peers, with less need to wait for formal publication. Just as preprints challenge established journals' online versions, these new ‘papers in progresses will be a significant challenge to online scientific publishers.

The semantic web will facilitate the development of automated methods for helping users to understand the content produced by those in other scientific disciplines. On the semantic web, one will be able to produce machine-readable content that will provide, say, automated translation between the output of a scientific device and the input of a data mining package used in some other discipline, or a self-evolving translator that allows one group of scientists to directly interact with the technical data produced by another.

These new products will allow users to create relationships that allow communication when the commonality of concept has not (yet) led to a commonality of terms. The semantic web will provide unifying underlying technologies to allow these concepts to be progressively linked into a universal web of knowledge, and will therefore help to break down the walls erected by lack of communication, and allow researchers to find and understand products from other scientific disciplines. The very notion
of a journal of medicine separate from a journal of bioinformatics, separate from the writings of physicists, chemists, psychologists and even kindergarten teachers, will some day become as out of date as the print journal is becoming to our graduate students.

**Figure: 1 Different Layers of Semantic Web**

The layers of the semantic web are built as new languages and tools that are anchored in XML and build towards a world of trusted information shared among collaborating groups of users.

Semantic Web technologies can considerably improve the information sharing process by overcoming the problems of current web portals. In this sense, Gateways based on Semantic Web technologies represent the next generation of web portals.

Semantic Web technologies will enhance the utility of Web Services when they are widely deployed. On the web, the many service providers will need to be able to advertise their services to an extremely wide and varied audience of service users. Providing brokering capabilities, the ability for service users to be automatically matched with service users, is difficult, and revolves around the same sort of vocabulary mapping issues.
that databases exhibit. In the current implementations, the services describe inputs, outputs, ports and other aspects of calling each other; is however the description of what the services do is left for a "content" field, to be filled in by an arbitrary description. Thus, the problem is similar to that of databases without agreed upon content many different mappings between diverse user communities need to be expressed.

Another enhancement to Web Services provided by the Semantic Web occurs when the specific service one needs is not immediately available. For example, suppose a company which supplies little gift boxes full of candies wants to buy a hundred gross of chocolate hearts and a hundred gross of candy canes, and ship them both to a plant in Nagoya for packaging, then we can find vendors of hearts, vendors of candy canes, and lots of shippers - but no one service that can do all three. Essentially, we are composing one new service (the candycane-chocolate-shipping service) out of three others.

The composition of Semantic Web Service descriptions allows us to pull this information together even if we don't agree upon the same terms for our descriptions a priori. Further, Semantic Web applications can then be used to analyze ways to reach the goal we need by putting services together in an efficient and effective means (e.g. shipping chocolate requires refrigeration, so merging additional information about chocolate helps ensure effective delivery). Although complex service composition is still very much a research issue, basic composition, done by matching inputs and outputs of the various services, is already doable using existing Semantic Web tools. We are not inventing relational models for data or composition methodologies for services. Rather, we are bringing to the web a number of things we already largely know how to do -- that is, we
are allowing them to work together in a decentralized system - without a human having to custom handcraft every connection.

Potential threats and criticism of the Semantic Web

Censorship, and privacy

Enthusiasm about the Semantic Web could be tempered by concerns regarding censorship and privacy. For instance, text-analyzing techniques can now be easily bypassed by using other words, metaphors for instance, or by using images in place of words. An advanced implementation of the Semantic Web would make it a lot easier for governments to control the viewing and creation of online information as this information would be much easier for an automated content-blocking machine to understand. In addition, the issue has also been raised that with the use of FOAF (software) files and Geolocation meta-data, there would be very little anonymity associated with the authorship of articles on things such as a personal blog.

Doubling output formats

Another criticism of the Semantic Web is that it would be much more time-consuming to create and publish content as there would need to be two formats for one piece of data. One format would need to be specialized for human viewing and the other would have to be specialized for machines. With this being the case, it would be much less likely for companies to adopt these practices as it would only slow down their progress. However, many web applications in development are addressing this issue by creating a machine-readable format upon the publishing of data or the request of a machine for such data.
It is very important to realize that the Semantic Web does not require content owners to individually encode information! The vast bulk of data to be on the Semantic Web is already sitting in databases and files in proprietary data formats. Downloaded bank statements, weather and stock quote information, human resource information, geospatial data, and maps, and so on...all that is needed to write an adapter to convert a particular format into RDF and all the content in that format is available.

Note that search engines for the traditional web of documents have the task of finding relevant items in a sea of documents in (some form of more or less broken) natural language, with links. The Semantic Web is very different. Search techniques for the Semantic Web are going to be very different: it may be that the value added will be made in different ways by systems roaming around and looking for patterns, or by performing some specific types of inference, or by indexing Semantic Web data in new interesting ways. It probably won't be eigenvector-based link analysis which drives the good hypertext search engines. In a way, the search engines are making up, by special techniques, for the lack of machine-understandable semantics in the documents on the web.

Client-server

Yet another criticism stems from the fact that the Semantic Web is based on a traditional client-server architecture, which ultimately is not scalable. URIs within RDF are still machine-specific references, which calls the persistence of these documents into question. Many have pointed to MAYA's Information Commons as a more practical implementation of the basic RDF schema that is both distributed and ultimately scalable.
Another problem encountered by Semantic Web proponents is the fact that ontology's used to represent information in open environments such as the Web do not generally coincide (e.g., the field “date” in one ontology may be represented by “year,” “month,” “day” in another). This problem is addressed though a line of work called ontology merging.