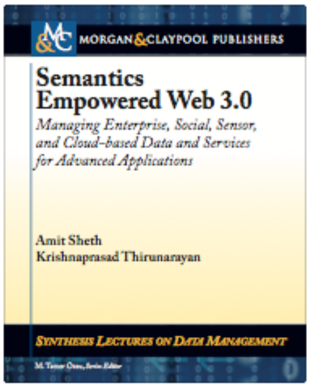
Lecture notes for INFO116

Lecture 1: The Role of Semantics and Metadata

Dr. Csaba Veres,
Institute for Information and Media Science,
The University of Bergen

(Ch 1 & Ch 2 to page 18)


MORGAN & CLAYPOOL PUBLISHERS
Semantics Empowered Web 3.0
Managing Enterprise, Social, Sensor, and Cloud-based Data and Services for Advanced Applications
Amit Sheth
Krishnaprasad Thirunarayan
SYNTHESIS LECTURES ON DATA MANAGEMENT

Semantics = Meaning
Meaning = Metadata+++

Web 3.0 has become a repository of an ever growing variety of Web resources that include data and services associated with enterprises, social networks, sensors, cloud, as well as mobile and other devices that constitute the Internet of Things. Data is now created by everyone, not just governments and organisations.

These pose unprecedented challenges in terms of

- heterogeneity (variety)
- scale (volume)
- continuous changes (velocity)

Evolution of different kinds of data on the Internet

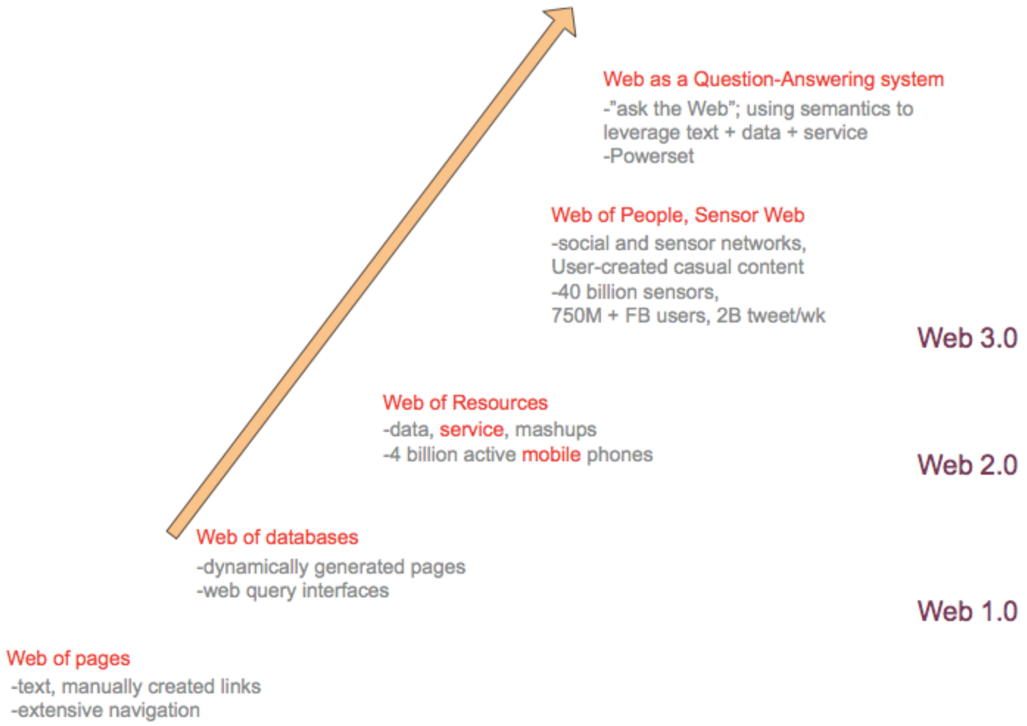
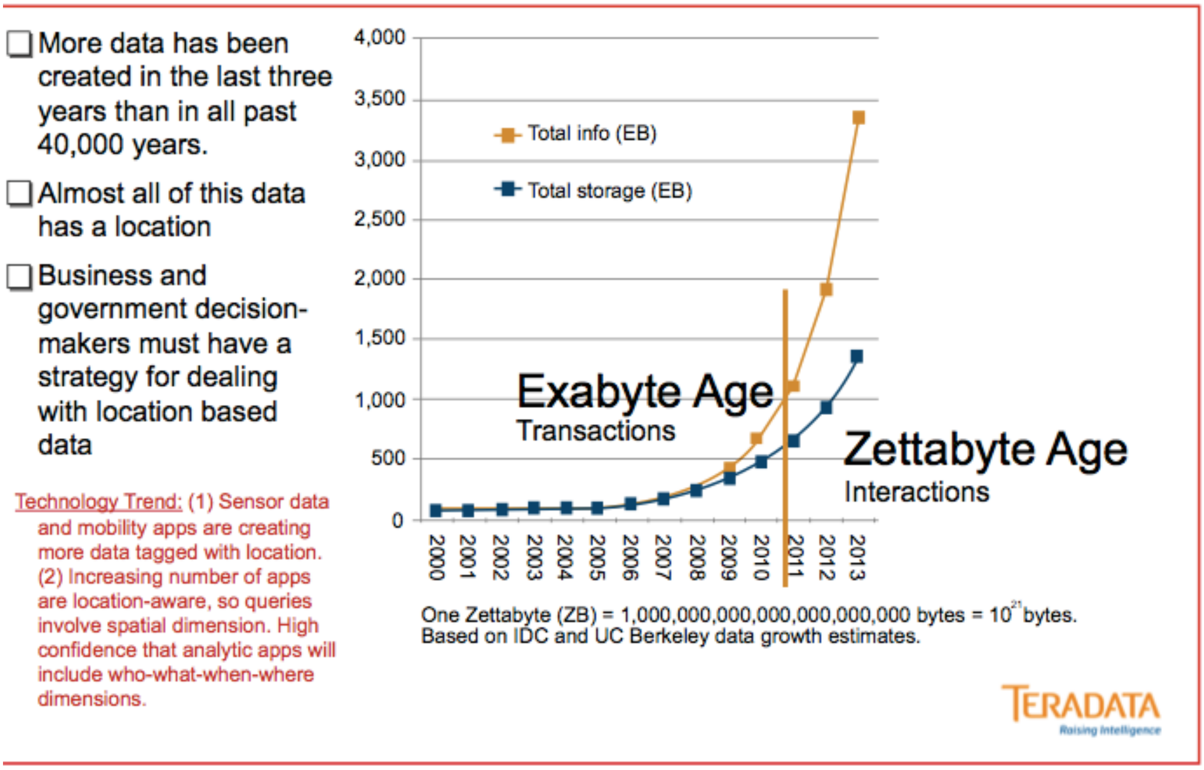

Web as a Question-Answering system
-"ask the Web"; using semantics to leverage text + data + service
-Powerset
Web of People, Sensor Web
-social and sensor networks, User-created casual content
-40 billion sensors, 750M + FB users, 2B tweet/wk
Web 3.0
Web of Resources
-data, service, mashups
-4 billion active mobile phones
Web 2.0
Web of databases
-dynamically generated pages
-web query interfaces
Web 1.0
Web of pages
-text, manually created links
-extensive navigation

Figure 1.1: Evolution of Web.

Data growth is expanding at a phenomenal rate, especially since the introduction of sensor networks.

A Boeing jet generates 10 terabytes of information per engine every 30 minutes of flight, according to Stephen Brobst, the CTO of Teradata. So for a single six-hour, cross-country flight from New York to Los Angeles on a twin-engine Boeing 737 — the plane used by many carriers on this route — the total amount of data generated would be a massive 240 terabytes of data. There are about 28,537 commercial flights in the sky in the United States on any given day. Using only commercial flights, a day's worth of sensor data quickly climbs into the petabyte scale — for a single day.

http://gigaom.com/2010/09/13/sensor-networks-top-social-networks-for-big-data-2/



- □ More data has been created in the last three years than in all past 40,000 years.
- □ Almost all of this data has a location
- □ Business and government decision-makers must have a strategy for dealing with location based data

Technology Trend: (1) Sensor data and mobility apps are creating more data tagged with location. (2) Increasing number of apps are location-aware, so queries involve spatial dimension. High confidence that analytic apps will include who-what-when-where dimensions.

One Zettabyte (ZB) = 1,000,000,000,000,000,000,000 bytes = 10^{21} bytes. Based on IDC and UC Berkeley data growth estimates.

We can no longer store all the data that are generated even if we wanted to. Nevertheless, these data are very valuable. We want to search them, browse them, integrate them, mine them, and ultimately use them to gain insight, develop situational awareness, discover new knowledge, answer difficult questions, and make decisions.

The fundamental hurdle in this quest is our inability to automatically relate, disambiguate, understand, and abstract data, and distill them into knowledge that we can reliably reason with.

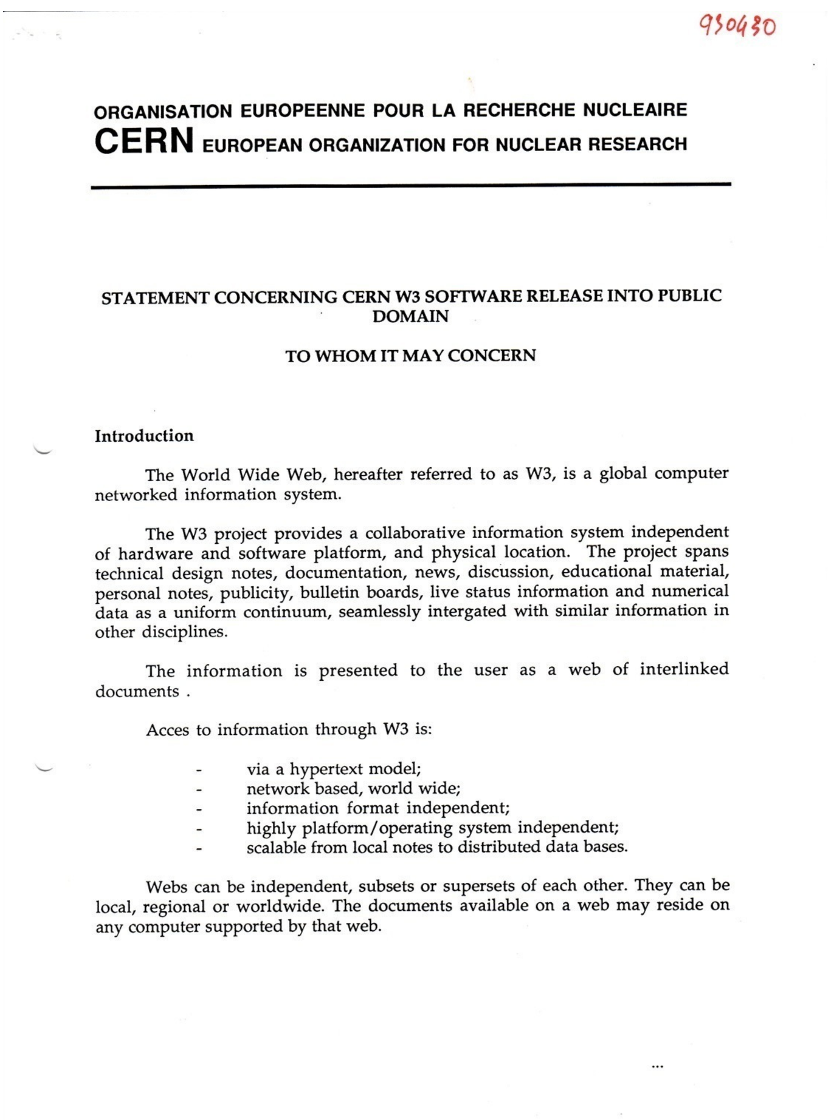
The Evolution of Data on the Web

1980s, the client-server paradigm was at the foundation of data sharing and social communication on the Internet. For each function, there was a separate server and client program.

- FTP was used to share files
- TELNET was used to access different machines
- EMAIL was used for private communication
- USENET was used for public exchanges.

30 April 1993 WWW was born.
http://info.cern.ch

The official release documents, in which CERN gave the world a BIG PRESENT!



930430

ORGANISATION EUROPEENNE POUR LA RECHERCHE NUCLEAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

STATEMENT CONCERNING CERN W3 SOFTWARE RELEASE INTO PUBLIC DOMAIN

TO WHOM IT MAY CONCERN

Introduction

The World Wide Web, hereafter referred to as W3, is a global computer networked information system.

The W3 project provides a collaborative information system independent of hardware and software platform, and physical location. The project spans technical design notes, documentation, news, discussion, educational material, personal notes, publicity, bulletin boards, live status information and numerical data as a uniform continuum, seamlessly intergated with similar information in other disciplines.

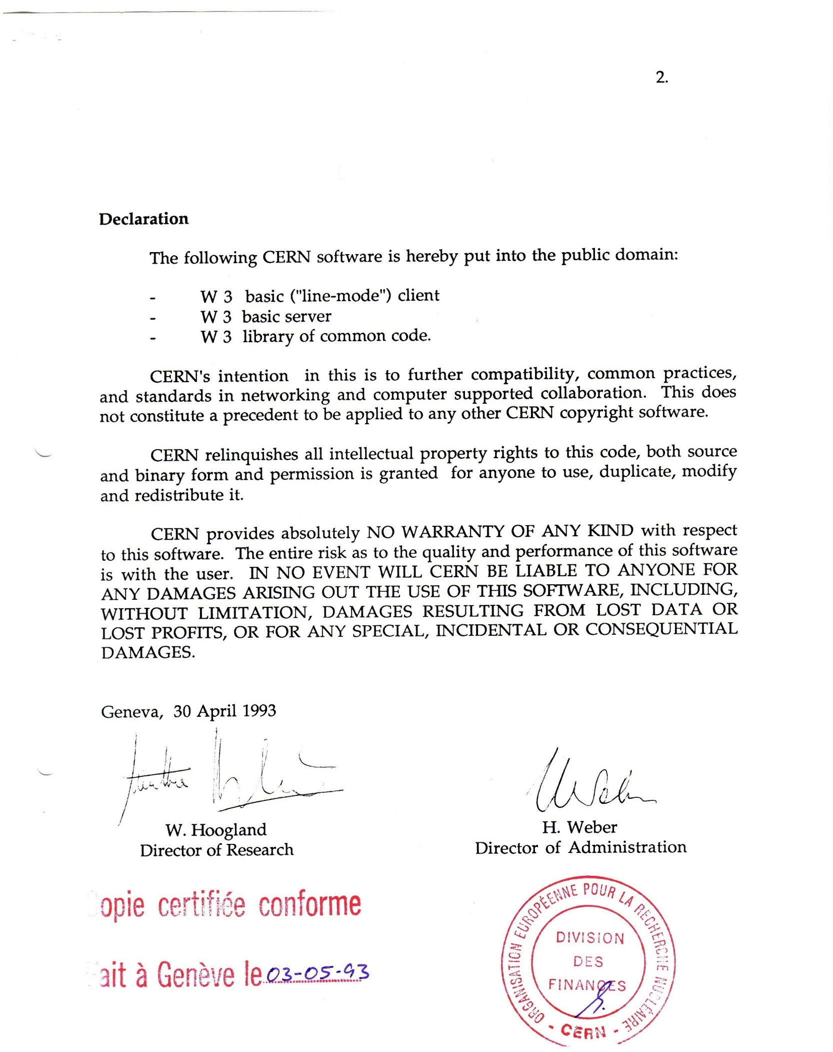
The information is presented to the user as a web of interlinked documents .

Acces to information through W3 is:

- via a hypertext model;
- network based, world wide;
- information format independent;
- highly platform/operating system independent;
- scalable from local notes to distributed data bases.

Webs can be independent, subsets or supersets of each other. They can be local, regional or worldwide. The documents available on a web may reside on any computer supported by that web.

...



2.

Declaration

The following CERN software is hereby put into the public domain:

- W 3 basic ("line-mode") client
- W 3 basic server
- W 3 library of common code.

CERN's intention in this is to further compatibility, common practices, and standards in networking and computer supported collaboration. This does not constitute a precedent to be applied to any other CERN copyright software.

CERN relinquishes all intellectual property rights to this code, both source and binary form and permission is granted for anyone to use, duplicate, modify and redistribute it.

CERN provides absolutely NO WARRANTY OF ANY KIND with respect to this software. The entire risk as to the quality and performance of this software is with the user. IN NO EVENT WILL CERN BE LIABLE TO ANYONE FOR ANY DAMAGES ARISING OUT THE USE OF THIS SOFTWARE, INCLUDING, WITHOUT LIMITATION, DAMAGES RESULTING FROM LOST DATA OR LOST PROFITS, OR FOR ANY SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES.

Geneva, 30 April 1993

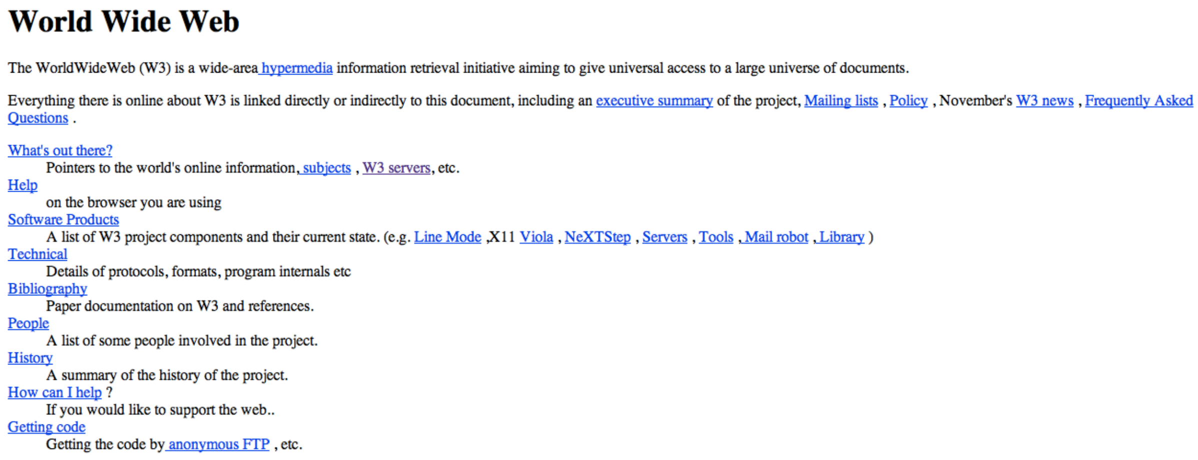
W. Hoogland
Director of Research

H. Weber
Director of Administration

opie certifiée conforme
ait à Genève le 03-05-93

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
DIVISION
DES
FINANCES
- CERN -

And the first web site, on Tim Berners-Lee's NeXT machine, built by Steve Jobs' company NeXT Inc.



World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy , November's W3 news , Frequently Asked Questions .

What's out there?
: Pointers to the world's online information, subjects , W3 servers, etc.

Help
: on the browser you are using

Software Products
: A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)

Technical
: Details of protocols, formats, program internals etc

Bibliography
: Paper documentation on W3 and references.

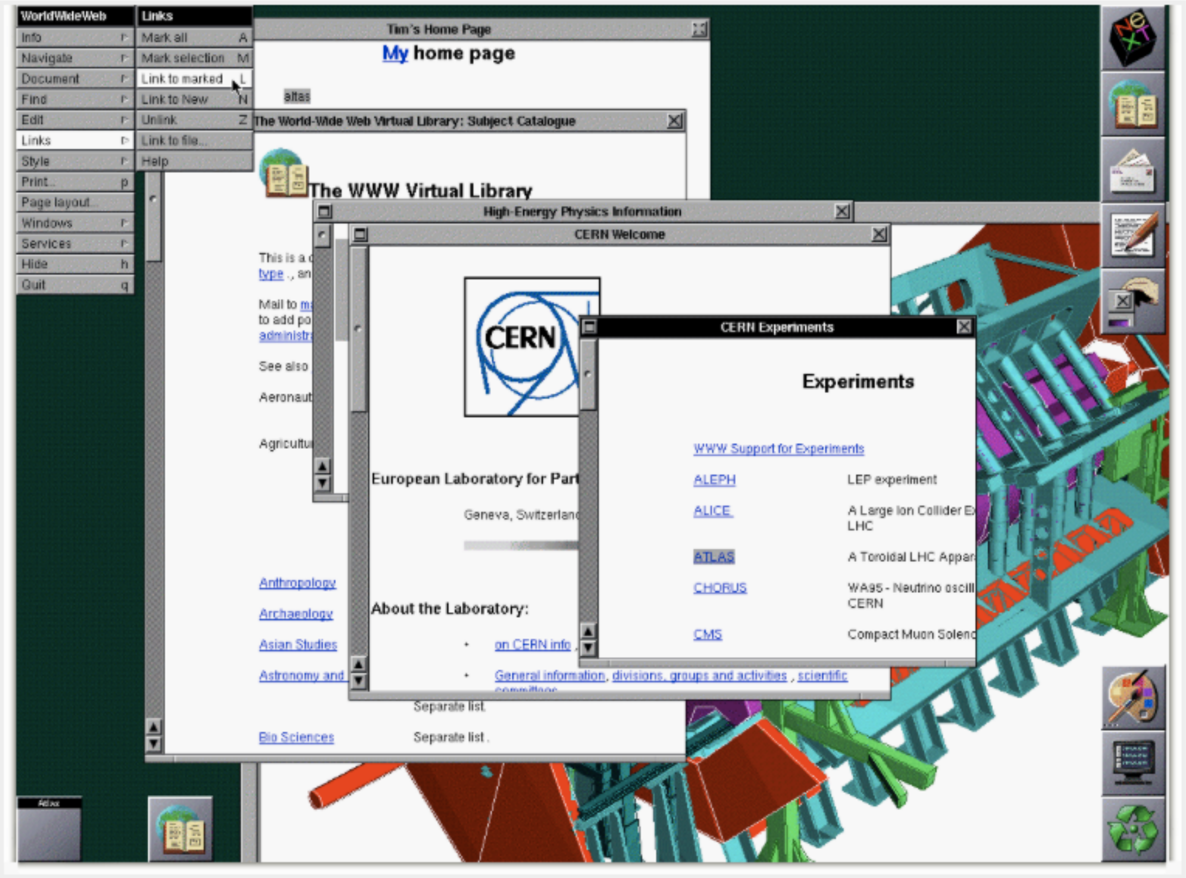
People
: A list of some people involved in the project.

History
: A summary of the history of the project.

How can I help ?
: If you would like to support the web..

Getting code
: Getting the code by anonymous FTP , etc.

The original NeXT web browser:



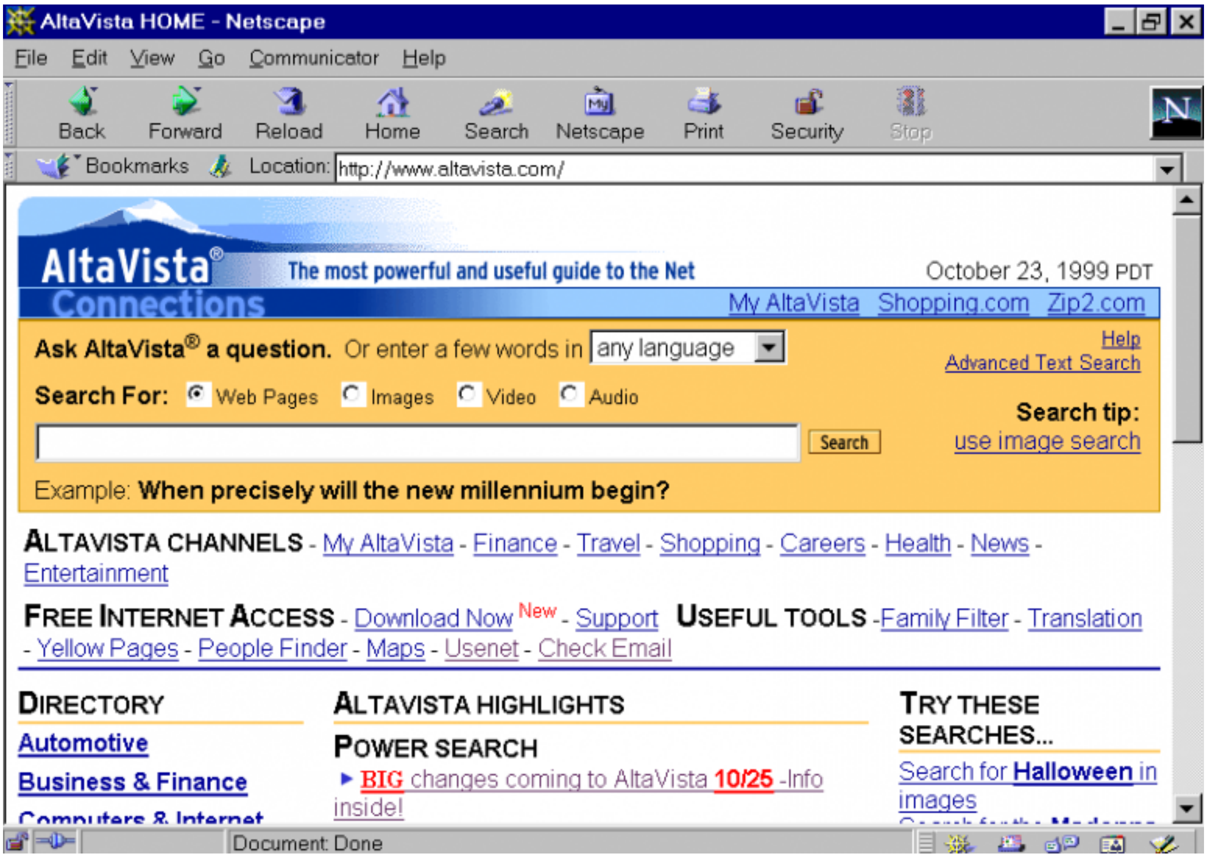
Screenshot of the original NeXT web browser in 1993

The most important concept is the Uniform Resource Locator (URL) (that encodes name or location of a Web resource or service reflecting information such as communication protocol [http, telnet, ftp, mail, etc.], machine [IP address], and full-path address [in a directory]), that enabled an "integrated client" Web browser to manage interactions with different servers. HyperText Transmission Protocol (HTTP) was developed for text-based request-response communication between a client and a server. HyperText Markup Language (HTML), an application/instance of SGML, was developed for defining logical structure and presentation of documents.

The web was originally the Web of documents in which web pages were linked using hyperlinks, with Web browsers providing a convenient means to navigate through them.

The HTML documents (or pages) were either created manually and linked or were dynamically generated from databases. The existence of a link between two documents suggests a weak association between the linked documents, and the count of the incoming links can be used as a measure of attention the document gets or deserves. However, it is not always clear if the directional link is supportive or dismissive. In response to keyword-based queries, a search engine gleans relevance of a document using its content and authoritativeness of a document using the topology of the web of documents.

People found information through keyword search, or various category systems (Yahoo, dmoz)



Location: http://www.altavista.com/

The most powerful and useful guide to the Net

October 23, 1999 PDT

My AltaVista Shopping.com Zip2.com

Ask AltaVista® a question. Or enter a few words in any language

Help
Advanced Text Search

Search For: ● Web Pages ○ Images ○ Video ○ Audio

Search tip:
use image search

Example: When precisely will the new millennium begin?

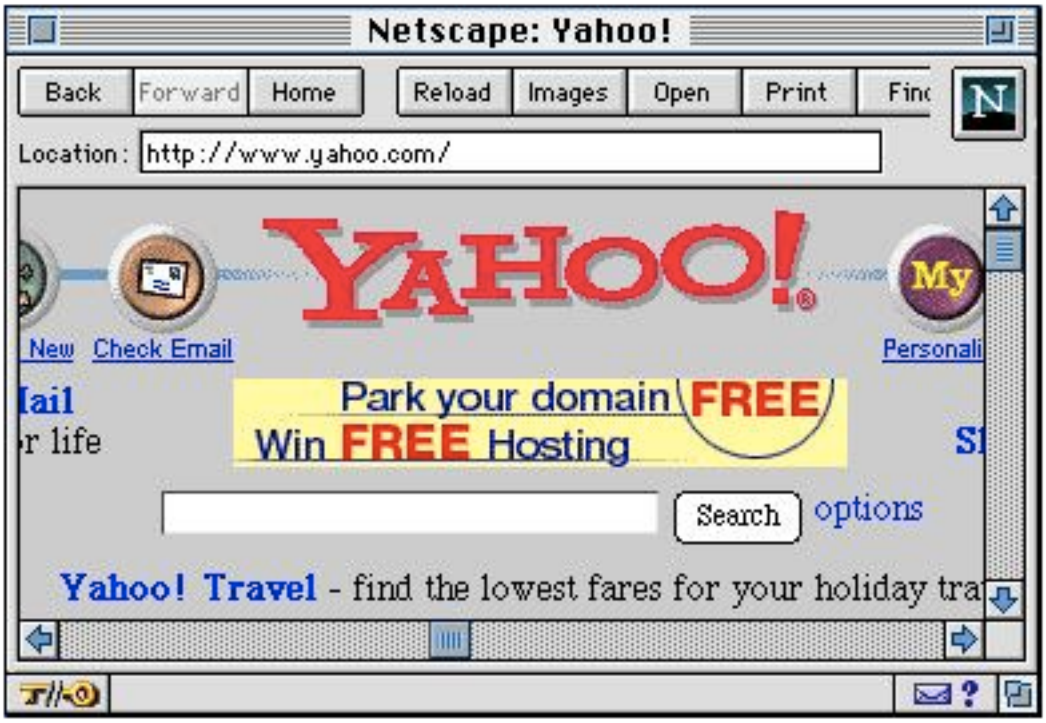
ALTAVISTA CHANNELS - My AltaVista - Finance - Travel - Shopping - Careers - Health - News - Entertainment

FREE INTERNET ACCESS - Download Now New - Support USEFUL TOOLS - Family Filter - Translation - Yellow Pages - People Finder - Maps - Usenet - Check Email

DIRECTORY
Automotive
Business & Finance
Computers & Internet

ALTAVISTA HIGHLIGHTS
POWER SEARCH
▸ BIG changes coming to AltaVista 10/25 -Info inside!

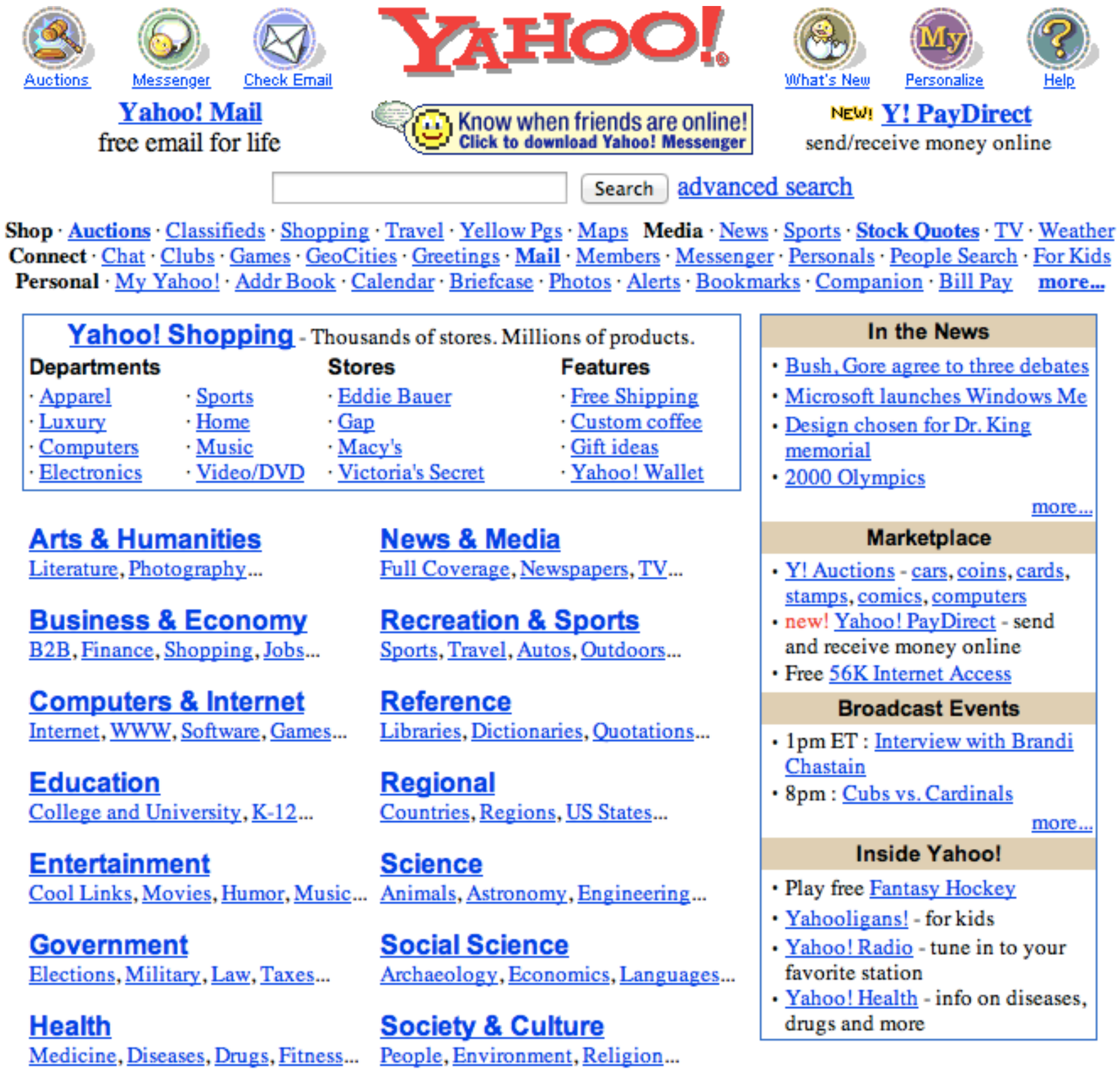
TRY THESE SEARCHES...
Search for Halloween in images



Location: http://www.yahoo.com/

Yahoo! Travel - find the lowest fares for your holiday tra

http://web.archive.org/web/20000915051602/http://www.yahoo.com/



Yahoo! Mail
free email for life

Know when friends are online!
Click to download Yahoo! Messenger

NEW! Y! PayDirect
send/receive money online

Search advanced search

Shop · Auctions · Classifieds · Shopping · Travel · Yellow Pgs · Maps **Media** · News · Sports · Stock Quotes · TV · Weather
Connect · Chat · Clubs · Games · GeoCities · Greetings · Mail · Members · Messenger · Personals · People Search · For Kids
Personal · My Yahoo! · Addr Book · Calendar · Briefcase · Photos · Alerts · Bookmarks · Companion · Bill Pay **more...**

Yahoo! Shopping - Thousands of stores. Millions of products.

Departments		Stores	Features
· Apparel	· Sports	· Eddie Bauer	· Free Shipping
· Luxury	· Home	· Gap	· Custom coffee
· Computers	· Music	· Macy's	· Gift ideas
· Electronics	· Video/DVD	· Victoria's Secret	· Yahoo! Wallet

Arts & Humanities
Literature, Photography...

Business & Economy
B2B, Finance, Shopping, Jobs...

Computers & Internet
Internet, WWW, Software, Games...

Education
College and University, K-12...

Entertainment
Cool Links, Movies, Humor, Music...

Government
Elections, Military, Law, Taxes...

Health
Medicine, Diseases, Drugs, Fitness...

News & Media
Full Coverage, Newspapers, TV...

Recreation & Sports
Sports, Travel, Autos, Outdoors...

Reference
Libraries, Dictionaries, Quotations...

Regional
Countries, Regions, US States...

Science
Animals, Astronomy, Engineering...

Social Science
Archaeology, Economics, Languages...

Society & Culture
People, Environment, Religion...

In the News

- Bush, Gore agree to three debates
- Microsoft launches Windows Me
- Design chosen for Dr. King memorial
- 2000 Olympics

more...

Marketplace

- Y! Auctions - cars, coins, cards, stamps, comics, computers
- new! Yahoo! PayDirect - send and receive money online
- Free 56K Internet Access

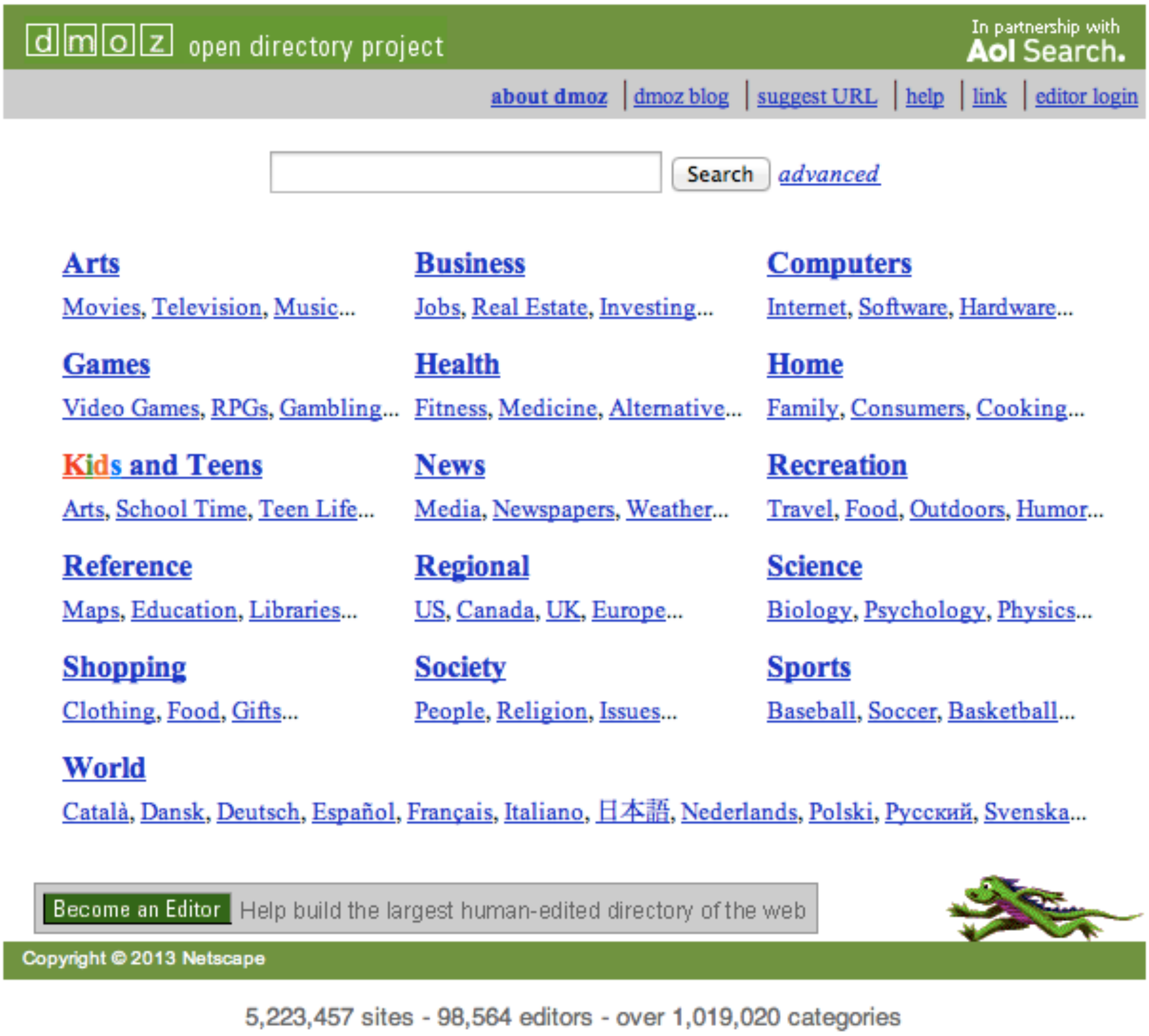
Broadcast Events

- 1pm ET : Interview with Brandi Chastain
- 8pm : Cubs vs. Cardinals

more...

Inside Yahoo!

- Play free Fantasy Hockey
- Yahooligans! - for kids
- Yahoo! Radio - tune in to your favorite station
- Yahoo! Health - info on diseases, drugs and more



dmoz open directory project

In partnership with Aol Search.

about dmoz | dmoz blog | suggest URL | help | link | editor login

Search advanced

Arts
Movies, Television, Music...

Business
Jobs, Real Estate, Investing...

Computers
Internet, Software, Hardware...

Games
Video Games, RPGs, Gambling...

Health
Fitness, Medicine, Alternative...

Home
Family, Consumers, Cooking...

Kids and Teens
Arts, School Time, Teen Life...

News
Media, Newspapers, Weather...

Recreation
Travel, Food, Outdoors, Humor...

Reference
Maps, Education, Libraries...

Regional
US, Canada, UK, Europe...

Science
Biology, Psychology, Physics...

Shopping
Clothing, Food, Gifts...

Society
People, Religion, Issues...

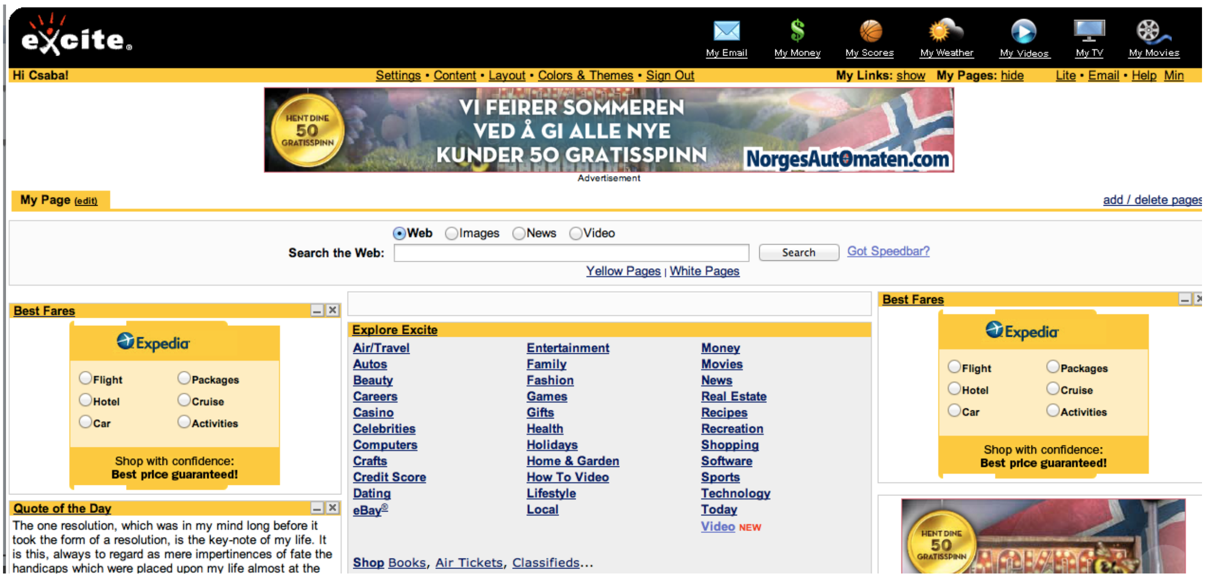
Sports
Baseball, Soccer, Basketball...

World
Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...

Become an Editor Help build the largest human-edited directory of the web

Copyright © 2013 Netscape

5,223,457 sites - 98,564 editors - over 1,019,020 categories



excite

My Email My Money My Scores My Weather My Videos My TV My Movies

Hi Csaba! Settings • Content • Layout • Colors & Themes • Sign Out My Links: show My Pages: hide Lite • Email • Help Min

VI FEIRER SOMMEREN VED Å GI ALLE NYE KUNDER 50 GRATISSPINN NorgesAutOmaten.com HENT DINE 50 GRATISSPINN

Advertisement

My Page (edit) add / delete pages

● Web ○ Images ○ News ○ Video

Search the Web: Search Got Speedbar?

Yellow Pages | White Pages

Best Fares

Expedia

○ Flight ○ Packages
○ Hotel ○ Cruise
○ Car ○ Activities

Shop with confidence:
Best price guaranteed!

Quote of the Day

The one resolution, which was in my mind long before it took the form of a resolution, is the key-note of my life. It is this, always to regard as mere impertinences of fate the handicaps which were placed upon my life almost at the

Explore Excite

Air/Travel	Entertainment	Money
Autos	Family	Movies
Beauty	Fashion	News
Careers	Games	Real Estate
Casino	Gifts	Recipes
Celebrities	Health	Recreation
Computers	Holidays	Shopping
Crafts	Home & Garden	Software
Credit Score	How To Video	Sports
Dating	Lifestyle	Technology
eBay®	Local	Today
		Video NEW

Shop Books, Air Tickets, Classifieds...

Best Fares

Expedia

○ Flight ○ Packages
○ Hotel ○ Cruise
○ Car ○ Activities

Shop with confidence:
Best price guaranteed!

HENT DINE 50 GRATISSPINN

By 1997, the Web grew to 1 million sites, and by 2004, to 50 million sites.

2003. Web 2.0.

Web 2.0 refers to the second stage in the evolution of the Web that enabled users to interact and collaborate with each other in a social media dialogue as consumers of user-generated content.

Examples of Web 2.0 include social networking sites, blogs, wikis, video sharing sites, hosted services, Web applications, mashups, folksonomies, etc.

If Web 1.0 is "read web," which involved significant effort in creating a Web content, then Web 2.0 is "read/write web," which en- abled any (even a casual) user to create content and easily share it. Examples of Web 2.0 technologies include XML (Extensible Markup Language that can be used to annotate documents or serialize data), AJAX (Asynchronous JavaScript and XML that enables Web applications to retrieve data from the server asynchronously without interfering with the existing page), JSON (JavaScript Ob- ject Notation, a lightweight data interchange format), RSS (Really Simple Syndication that allows subscription to feeds such as News, Events, Sports, etc.), and P2P (Peer to Peer) content sharing systems.

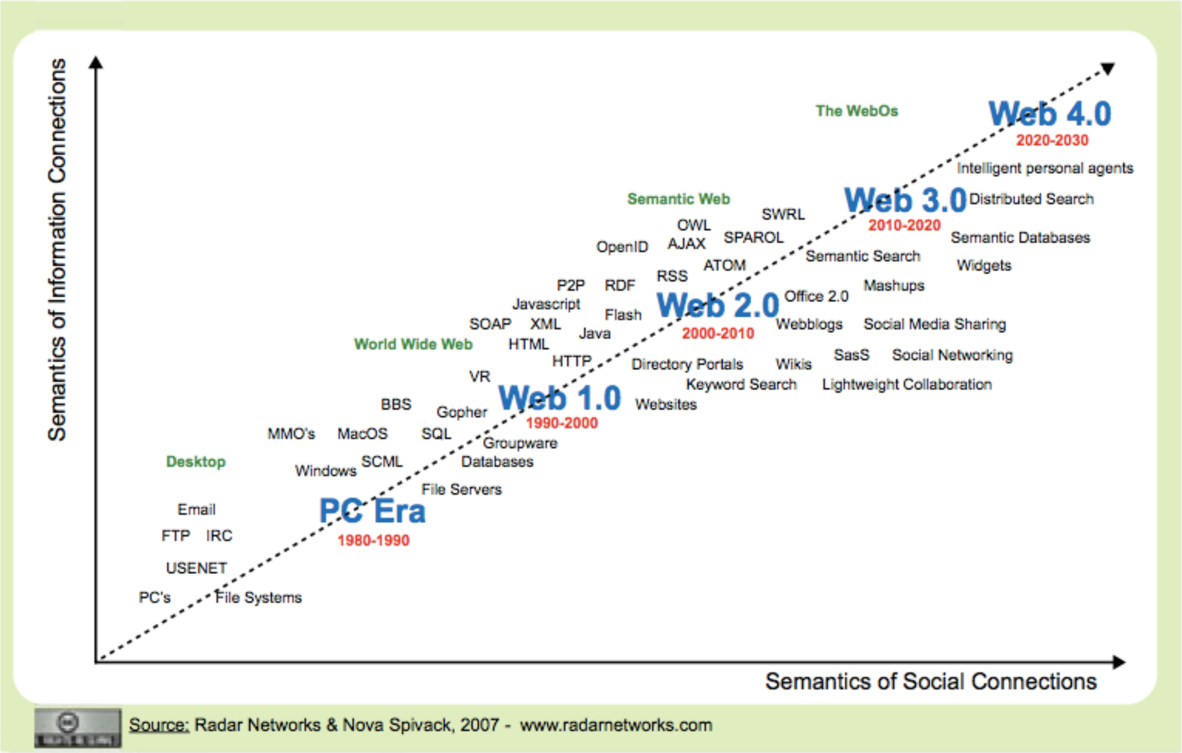
In terms of semantics Web 2.0 supported the use of tags to describe various features. However, besides providing a rich interactive experience over the Web, and some ability for tag-based search of resources, Web 2.0 did not provide any significant additional power to meaningfully describe (or associate meaning with) data, and to reason with Web data and links. Applications employed pattern-based information retrieval techniques to summarize user-generated content, while mashups exploited spatio-temporal context to organize user-generated content. The social media succeeded in connecting and engaging people.

By 2006, there were 100 million websites, 9 billion web pages, and 1 billion global users.

2006. Web 3.0

Currently there are approximately 250 million websites and 32 billion web pages. Web 3.0 has significant extension in the types of data, people, and interactions, along with heterogeneity and scale. Perhaps the most significant extension of the Web is the ability to connect a significant portion of humanity for the first time with over 5 billion mobile phones, of which over a billion and a half of the new connections have access to the Internet. This is complemented by more than 40 billion sensors covering increasingly large parts of the earth, constantly reporting on human activities and environmental information.

As time progresses, the connections between information is becoming explicit and more fine-grained, promoting personalization and more expressive social connections, eventually leading to robust *collective intelligence*:



Source: Radar Networks & Nova Spivack, 2007 - www.radarnetworks.com

If we are to have any chance of organizing, integrating and understanding this mass of data, then the data will have to be self describing in some way: we have to know what the data is about, and machines will have to be able to reason about the data.

Berners-Lee and Fischetti [2001] noted that “if HTML and the Web made all the online documents look like one huge book, RDF, schema, and inference languages will make all the data in the world look like one huge database.” Berners-Lee also said “I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web—the content, links, and transactions between people and computers.” A “Semantic Web,” which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy, and our daily lives will be handled by machines talking to machines.

Different ways in which semantics plays a part in the future Web:

- Integration: understanding the connection between data has always been a big part of semantics. In the new Web it is even more essential in integrating information about the same concept or object in a different modality and media—for example, to relate a person's image with his or her descriptive information, or to correlate information about an event on social media with corresponding sensor observations. In coming years, semantics will play a crucial role in integrating objects that straddle the cyber-physical or physical-virtual divide.
- Intelligent processing: objects, relationships, semantic search instead of text and keyword search.
- Knowledge enabled computing: understand background knowledge relevant to a particular set of facts. NLP and machine learning. We are now able to apply domain-independent (e.g., related to time, space, and geographic concepts) to domain-specific models of various complexities and comprehensiveness such as nomenclatures, taxonomies, and ontologies, to improve information processing. Community knowledge bases, DBPedia, Urban dictionary, etc.
- Abstraction and experience: Semantic approaches support abstractions to convert low-level data and observations into high-level symbolic representation that constitute our perception and cognition. e.g. Sensor data from commercial airlines.

Different kinds of Semantics

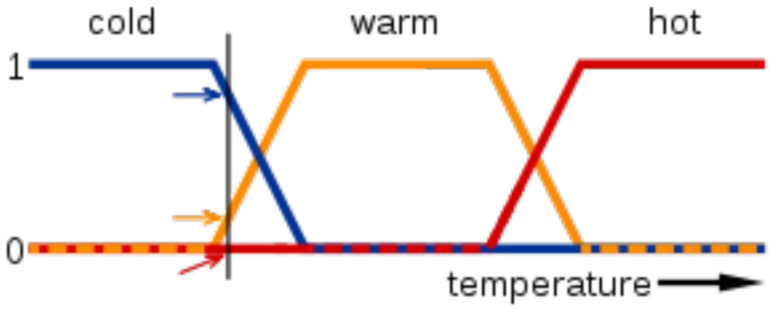
In order to interpret data of any kind, we need a model of semantics and a means to associate semantics with the data. Ultimately, providing semantics to data involves understanding entities, actions, and relationships the data describe, and making explicit relationships that are implicit using reasoning.

In fact, querying data involves verifying or seeking entities that satisfy certain relationships.

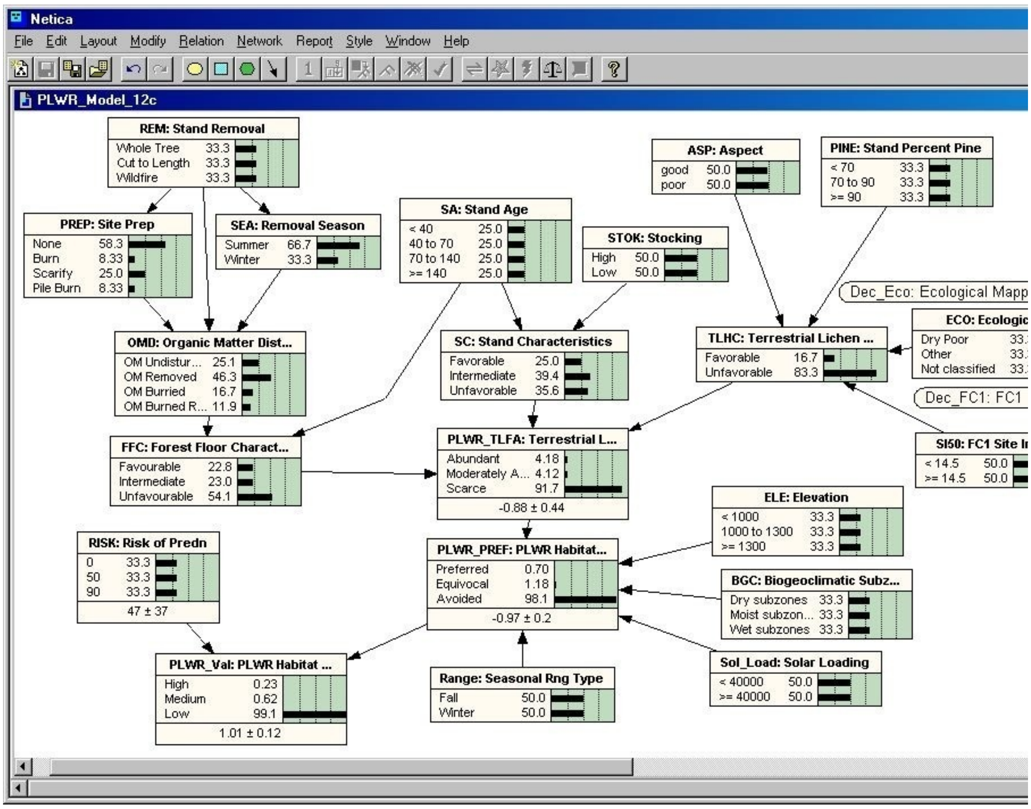
There are different ways to think about, and to formalise semantics:

- Implicit semantics: the meaning of the data is in the patterns which are not explicitly represented. Nevertheless, the meaning can be used by machines and automatically derived. For example:
  - A document linked to another document via a hyperlink, potentially associating semantic metadata describing the concepts that relate the two documents.
  - Word co-occurance. Basis for extracting topics in natural language, clustering documents by meaning.
  - Disambiguation based on use, context
- Formal semantics. Use of an artificial logical language designed for knowledge representation and reasoning. Formal languages share some features:
  - Model Theoretic Semantics: Primitive symbols (atoms, n-ary function symbols, and n-ary predicate symbols) in a formal language are interpreted using semantic structures (domain of discourse D, n-ary functions over D, and n-ary relations over D) that reflect certain basic presuppositions about the "nature of the world" that are implicitly described by the language. Expressions in the form of terms are mapped to the domain of discourse, and sentences are used as constraints to define models. The logical consequence of a set of sentences S is the set of sentences T that are true in all models of S. e.g The semantics of subsumption in DLs, reflecting the human tendency of categorizing by means of broader or narrower descriptions.
  - The Principle of Compositionality: The meaning of an expression is defined in terms of the meanings of its immediate parts and of the way they are syntactically combined. The emphasis is on locality and referential transparency.

- Powerful soft semantics
    - In addition to statistical and formal approaches, there are a host of non truth preserving logics, which should be useful in particular situations.
    - Fuzzy logic



    -
    - Probabilistic rasoning
    -



Currently the major proposals involve the first two, though the latter group of techniques could also be used in conjunction with the first two. For example ontology alignment, or learning prior probabilities.

RETRIEVAL APPLICATIONS: INFORMATION VS. DATA

Data retrieval (DR) aims at determining all objects that satisfy a semantically well-defined query, while information retrieval (IR) aims to decipher user information need and interpret document content in order to satisfy a query. Matching between a document and a query in the abstracted space of the set of index words (which exemplifies IR) is very imprecise, while relational database and its query language (which exemplifies DR) has semantic clarity and precision.

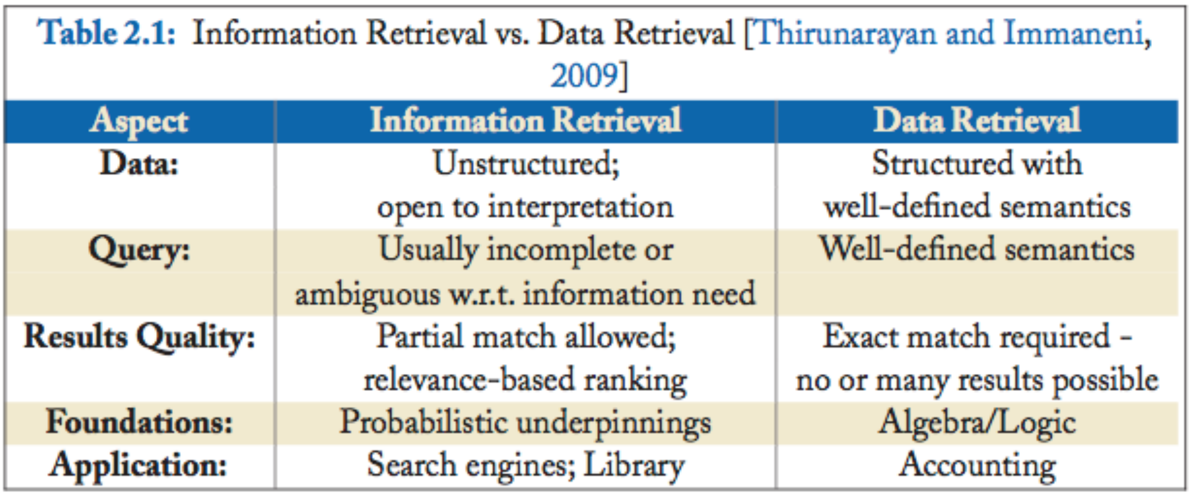


Table 2.1: Information Retrieval vs. Data Retrieval [Thirunarayan and Immaneni, 2009]

Aspect	Information Retrieval	Data Retrieval
Data:	Unstructured; open to interpretation	Structured with well-defined semantics
Query:	Usually incomplete or ambiguous w.r.t. information need	Well-defined semantics
Results Quality:	Partial match allowed; relevance-based ranking	Exact match required - no or many results possible
Foundations:	Probabilistic underpinnings	Algebra/Logic
Application:	Search engines; Library	Accounting

First generation search engines such as Aliweb retrieved documents that matched keyword- based queries. Second generation search engines such as Excite, Lycos, and Altavista incorporated content-specific relevance ranking based on a vector space model (TF-IDF) to hone in on a relevant subset of documents in spite of high recall [Manning et al., 2008]. To overcome spamming, and to exploit collective Web wisdom, third generation search engines such as Google, Yahoo!, and Bing incorporated content-independent, source authority information using a PageRank algo- rithm [Brin and Page, 1998] and notions of hubs and authorities [Kleinberg, 1999], and attempted to glean relative semantic emphasis of various words based on syntactic features, such as fonts, and distance between query term hits. Contemporary search engines have also incorporated context, annotations, user profiles, and past query history associated with a user to personalize the search and apply additional reasoning to improve satisfaction of the information need [Guha et al., 2003]. distinguish navigational searches, where a user provides a phrase to be found in a document, and re- search searches, where a user provides a phrase to designate an object.

More recently, major search engines have started to exploit large background knowledge bases, parts of which are based on a structured knowledge extracted from a community-created corpus (e.g., Wikipedia) and domain-specific structured datasets repre- sented as linked data using Semantic Web technologies. Although details are not publicly discussed, the Yahoo! initiative of developing ConceptBase, Bing's incorporation of Powerset technology which extracts entities and relationships from Wikipedia, and Google's acquisition of Freebase point to the type of background knowledge being exploited. Some, like Bing, are more aggressively creating a domain-specific entity or knowledge base and support limited forms of faceted search. Emerging semantic search engines, and their "intelligent" counterparts—question-answering systems—benefit from all three forms of semantics [Ferrucci et al., 2010].