The World Wide Web: Strategies and Tactics

Harry Halpin

March 12th, 2004
The World Wide Web

Historically the most significant computational phenoumena...yet.

Pervasive in everyday life, much more so than talking robots or voice recognition systems!

- **Amazon.com**: They sell more books than any single bookstore entirely through the Web - and they have a US patent (5,960,411) on one-click purchasing, the basis of e-commerce.

- **eBay**: Anything can be bought and sold on the Web - including a free Scottish tabloid, sold within a day of publication for eighty quid.

- **Google**: Their search engine serves as Everyman’s Oracle, processing millions of queries a day - yet now they’re not telling us how they work! Or claiming pigeons do their work...

Where did it come from and where is it going? Who controls it? What are its problems and its successes?
History of the Web

1770s: David Hume presents philosophical idea that all complex ideas are the results of associations between simpler ideas. Thinking is association.

1945: Vannevar Bush, Roosevelt’s scientific adviser, proposes in his paper “As We May Think” that machines should store information so it may be arbitrarily linked to any other piece.

1965: Hypertext coined by Ted Nelson: some text connects to other text - non-sequential text. Nelson refuses to use the World Wide Web today, claiming that it’s still based on a “paper” model without annotation or automatic royalty. 1989: Tim Berners-Lee, a computer engineer at CERN, proposes idea for World Wide Web, where any information can be linked to any other information on any machine.

History of the Web Part 2

1991-93: Tim BL with Roy Fielding refines specifications of URI, HTTP, and HTML, releasing them through the IETF.

1994: Netscape releases first modern Web Browser, can display images.


1995: Java makes programs web portable over the Web.


1999: XML: Web now used to transfer machine-readable data

2003: Semantic Web (RDF/OWL) proposals and Web Services concretize.
The Components of the Web

**Web Servers:** Computers that host information on the Web, like Apache or httpd.

**Web Clients:** Programs that access Web information and displays it, like browsers such as Mozilla or Internet Explorer.

**HTTP:** (HyperText Transfer Protocol) that allows machines to transmit information over the Web.

**HTML:** (HyperText Mark-Up Language) that allows someone to mark up and link documents on the Web.

**URL:** (Uniform Resource Locator) the address (www.myaddress.com/mypage.html) that information on the Web has.
How Big is the Web?

Very hard to measure, due to web’s anarchic nature.

- More than 3 billion web-pages according to Google - and Google does not even index all of them.

- More than 8 million unique servers - as of 2002.

- English has just been superseded, with less than half of Web currently in English.

- Growth seems to have slowed down - but then the Third World such as many parts of Africa and Asia are just starting to get online.

UN has begun WSIS (World Summit on Information Society) to begin to think through its policy on the Web and information architecture in general.
Why is the Web Successful

- **Universal**: Any machine with a browser and HTTP support can access the Web - no matter what the hardware. Any information that can be given a URL can be put on the Web. Anyone with access to a browser can access information on the Web.

- **Decentralized**: Anyone can put up a web site or web server on the Web, without asking permission from any central authority. Of course, there’s domain name problems!

- **Low-Learning Curve**: Surprisingly easy to learn and use!

- **Multimedia**: At first just used for text, but now images, music, and other forms of media are dominating the Web.
Who Controls the Web

**Nobody.** It’s anarchy. However, there are groups with influence.

**IETF:** (Internet Engineering Task Force) No membership, just open participation by computer enthusiasts. Develop and ratify Internet standards through RFCs (requests for Comments). At www.ietf.org.

**ISO:** (International Organization for Standardization) Membership consists of 148 nations whose standards bind their government, NGOs, and businesses and covers everything from welding to “information”. Generally too slow for the Web.

**Companies** The Browser wars between Microsoft and Netscape involved each company adding proprietary extensions to HTML, threatening to fragment the Web.

**The World Wide Web Consortium** (W3C) Founded by Tim Berners-Lee to be the de facto organization to keep bring the Web to “its fullest potential” through open standards. Found at www.w3.org.
How the W3C Works

Tim Berners-Lee: is the Director. He is the final assessor of consensus, and authorizes all technical publications and standards, and can act as a “tie-breaker” in disputes.

Headquarters: In European Research Consortium for Informatics and Mathematics (ERCIM) (France), Keio University, and MIT.

Team Members: Are employees of the W3C. There’s 68 of them across 10 countries, and there’s one in Edinburgh, Henry S. Thompson.

Members: Currently 371 organizations that pay a fee to take part in the W3C. Everyone from IBM to the University of Hawaii is part of the W3C, an odd collection of hackers, academics, and industry.

Team Members: When interest in moving the Web in a certain direction takes off, the W3C creates a Working Group to discuss and propose standards. The standards are available for public comments, and eventually become a Recommendation.
The Web has Problems...

It’s impossible to find anything on the Web - even with Google it seems like a hit and miss affair - and as the Web grows larger it’s going to get worse.

HTML is just a layout language - it can’t compute like an actual programming language - so adding “1 + 1” on the Web is impossible.

Things keep disappearing on the Web. URLs keep changing - thus the problem of Broken Links at “404 Object Not Found”.

The Web can’t answer questions. Type in “Where can I buy the cheapest computer in the UK”.


There’s no identity. Spam runs rampant, domain names get stolen or squatted in exchange for exorbitant bribes.
The XML Success Story

One problem is the division of content and presentation. For example, the header

```html
<h1>
```

tag both marks up a conceptual “header” idea, and a layout type, usually **bold**.

**XML** takes HTML and changes it is two ways - everything now can be an extensible tag name and all tags must be properly enclosed - leading to “tree” data structures.

To the W3C’s surprise, soon everyone was using XML to transfer data across the Web!

Simple, elegant, and clear, the W3C’s greatest success to date.

XML is now a multi-million dollar industry, and increasingly universal common data format for the next millennium.
Strategy and Tactics

Tim BL was right: the open standard process can influence the growth of the Web. The W3C has a strategy for Web: After understanding why the Web works so well, build a series of standards to make the Web more easily usable by machines.

Through the W3C, industry wants to use Web to communicate, broker, and use its applications over the Web. This is called Web Services.

The identity issue has still to resolved.

Strategies are visions of how the web could work better.

A series of standards has been released as tactics to achieving these strategies.

Strategies are in tension if not outright conflict.
The Architecture of the WWW

Still not sure exactly why the Web works as well as it does.

W3C appointed the TAG (Technical Architecture Group) to draft a document explaining it.

So far, have released their first draft. Defines the web as a “universal network of information resources”.

A resource is anything that has identity.

Identity on the Web, I assume - i.e. a URL.

URLs are supposed to be absolute (www.mypage.com/directory/file), and are supposed to infrequently change.

Content-negotiation and multiple representation heavily emphasized.

Yet used in less than .001 percent of web-sites!
The Semantic Web

Tim BL’s dream: A web that allows machines to seamlessly help us navigate the Web and its relation to the world. Not a web of web-pages, but a web of meaning.

Imagine an AI agent that if you asked for the best price for some item, could go out on the Web and actually find it, and get it delivered to your front door.

An ambitious Web strategy - and it also sounds like bad sci-fi. Despite the hype, getting this to work is a hard if not impossible problem - as AI found out in the 80s.

One strategy is to give every object on the Web a unique name (URN - Universal Resource Names). Or the world. Still, how to keep track of all these names? And what’s an object?

How can we determine the relationship between a resource and what it is about in the Wide World outside the Web. Humans are good at this, and machines are manifestly bad at this.
The Semantic Web Part Two

**Metadata** is one way out of the problem - that’s data “about” other data.

Machine-readable data can then be put on the Web **Ontologies** - knowledge representation classifications, like the taxonomies of animals or catalogs of car parts.

**RDF** (Resource Description Framework): A way of storing meta-data in a machine readable format - as triples.

(author, Harry, slide_show)

**OWL** (Web Ontology Language): A way of implementing first order logic on the web.

Aristotle is a man. All men are mortal. Thus, Aristotle is a man.

My web page is about the Eiffel Tower in France. The Eiffel Tower is a tourist attraction in France. Thus, my web-page is about tourist attractions in France.

Still far away from actual “semantics”!
Web Services

Data needs to be able to be sent anywhere. Yet, how? Firewalls block strange data from entering your computer.

Unless, as realized by the Web Services crowd, you can disguise it as Web Page by wrapping it in XML and serving it over HTTP - i.e. SOAP (Simple Object Access Protocol)

Web Services allows you to run programs over networks, which has been done for years by things such as RPC (Remote Procedure Calls) long before the Web. The difference is now it uses Web protocols and can interact with rest of Web.

E-business is very excited by this, wants to be able to display their services and haggle on the Web better.

Without a centralized repository of Web Service listing (like a phonebook).
Web Services Part 2

How to find anything without a phone book?

UDDI (Universal Description, Discovery, and Integration) allows lists of Web Services to be listed. Accessing and integrating them is going to be tricky. Most working programs are also not Web Services, and while it’s easy to create Web Services, updating legacy code is difficult.

Unclear how this interacts with Semantic Web. Instead of unchanging URIs, applications use URI’s to deliver constantly changing application-specific data.

Cheap hack or wave of the future?
Multimodal Web

The Web is still mostly text - and this is changing, as the Web is now being used to host video, music sharing, mobile phones.

**SMIL**: (Synchronized Multimedia Integration Language) allows web-enabled devices to use on the multimedia that they support.

**VoiceXML**, the Voice Extensible Markup Language. For creating audio dialogs that feature synthesized speech and digital audio - surfing the Web via voice.

**EMMA**: (Extensible MultiModal Annotation) allows data to be time-stamped, have differing input modes.

**SVG**: (Scalable Vector Graphics) XML-based two dimensional graphics that can be rendered optimally on all sorts of devices.
More Multimedia

The rendering of the same content can be improved automatically as devices get more powerful unlike Flash!

Back to basics: Can we ever get the Web standards to be used as a layout language for print documents?

No real video or audio formats...yet.
The Grid

The Web is never totally off, but it can slow down. Can we get more power?

Needed for giant data-sets, such as the Human Genome Project. Currently scientists are using ftp to transfer their files - but the files are getting too big for even one computer to handle.

Can we use the Web to solve these problems? Distributed data-sets, distributed applications, harnessing Web Standard protocols.

Scientists then can do analysis over data-sets previously unheard of, called e-science, running on a more powerful Web, the Grid.

The fastest way to get to your data is not use the network at all. Caching in Web browsers being an example.
The Next Generation Web

Most people are fine with the way the Web is right now, yet as it grows its problems are going to get worse.

The identity problem is arguably the most needed problem to solve - after all, spam is threatening everyone’s e-mail box.

The metaphysical quandaries raised by the Semantic Web about the nature of Web resources are confusing and unclear.

The Web Services initiative may want everyone to forsake their personal computers, but people like having their own copies of data.

Too many standards? Are they too long and confusing? Should they seek more power, but allow themselves to be complex, or stick to being minimal?

The Web and the W3C needs you! It’s energy has always come from the immense amount of users who are willing to contribute their time, energy, and their opinion.