

Adaptive Portals: Next Generation Web Deployment



Michael Hogarth, MD

Assistant Professor, Internal Medicine and Clinical Pathology

The background is a dark blue field with several large, semi-transparent gears of various shades of blue. On the left side, there is a vertical strip with a colorful, abstract, and somewhat pixelated pattern in shades of orange, yellow, and brown. A thin white horizontal line is positioned below the title, and a thin white vertical line is positioned to the left of the list items.

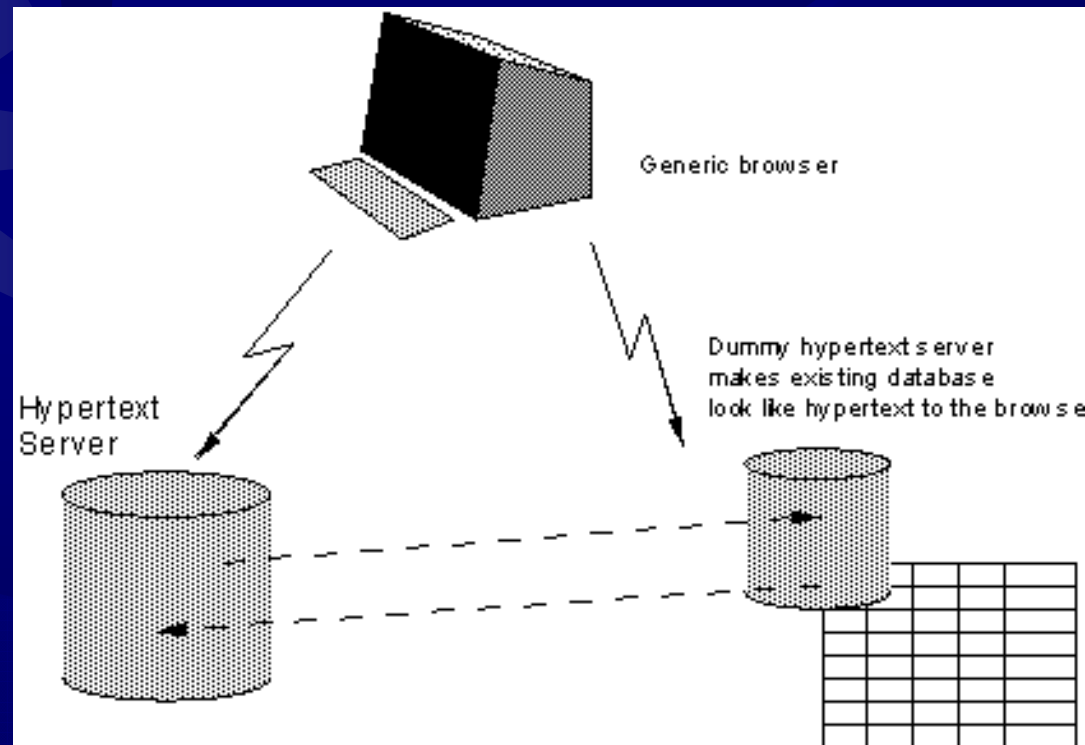
Presentation Summary

- ✦ Evolution of the Web
- ✦ Web server basics
- ✦ What are portals exactly?
- ✦ Adaptive portal models
- ✦ Implementing adaptive portals

Web beginnings...

Information Management: A Proposal

Tim Berners-Lee, CERN March 1989, May 1990



Reference: <http://www.w3.org/History/1989/proposal-msw.html>



Original vision

“Some examples of systems which could be connected in this way are:

- **uucp News** This is a Unix electronic conferencing system. A server for uucp news could make links between notes on the same subject, as well as showing the structure of the conferences.
- **VAX/Notes** This is Digital's electronic conferencing system. It has a fairly wide following in FermiLab, but much less in CERN. The topology of a conference is quite restricting.
- **CERNDOC** This is a document registration and distribution system running on CERN's VM machine. As well as documents, categories and projects, keywords and authors lend themselves to representation as hypertext nodes.
- **File systems** This would allow any file to be linked to from other hypertext documents.
- **The Telephone Book** Even this could even be viewed as hypertext, with links between people and sections, sections and groups, people and floors of buildings, etc.
- **The unix manual** This is a large body of computer-readable text, currently organised in a flat way, but which also contains link information in a standard format ("See also..").
- **Databases** A generic tool could perhaps be made to allow any database which uses a commercial DBMS to be displayed as a hypertext view.”

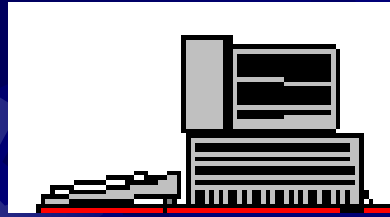
The protocols

- ★ HTTP - hypertext transfer protocol
 - ★ “language” for data exchange between client (browser) and server (web server)
 - ★ determines how “pages” are “asked for” and “returned”
- ★ HTML - hypertext markup language
 - ★ deals strictly with the rendering of information in a browser
- ★ Neither HTML or HTTP *require* the other

Web server basics -- HTTP

- ★ HTTP = Hypertext Transfer Protocol
 - client/server data exchange protocol
 - has only been versioned a few times
 - includes commands that:
 - retrieve a file
 - return the date/time stamp of a file
 - return information about the server (in a header)

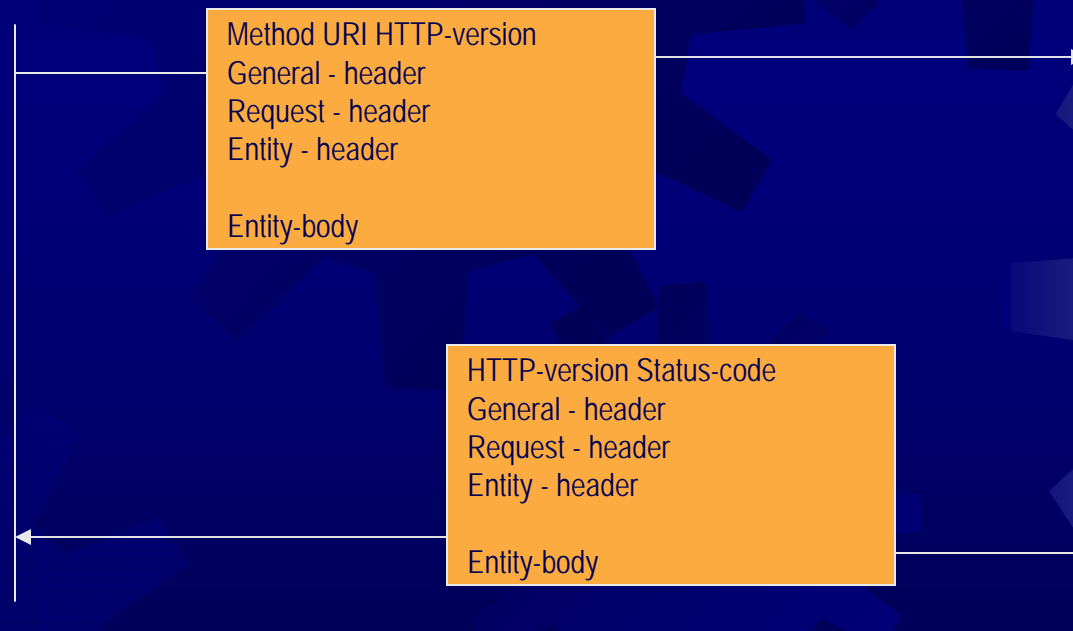
Structure of an HTTP transaction



Client



Server



An Example

```
szhogart@mycin:~  
[szhogart@mycin szhogart]$ telnet www.ucdavis.edu 80  
Trying 169.237.105.90...  
Connected to www.ucdavis.edu.  
Escape character is '^'.  
GET /  
  
<HTML>  
  <BODY bgcolor="#ffffff">  
We're moving.<br>  
The K-8 Aeronautics Internet Textbook site (http://wings.ucdavis.edu and  
http://muttley.ucdavis.edu) site has been moved to  
http://wings.avkids.com. Click on  
the following link to access the new site:  
<a href="http://wings.avkids.com/">(http://wings.avkids.com/)  
</a><BR>  
<p>  
Nos estamos mudando.<br>  
La pagina del Libro de Texto de Aeronautica K-8 en la Internet  
(http://wings.ucdavis.edu y http://muttley.ucdavis.edu) se ha mudado a  
http://wings.avkids.com. Presione el siguiente enlace para conectarse al  
nuevo sitio:  
<a href="http://wings.avkids.com/">(http://wings.avkids.com/)  
</a><BR>  
  </BODY>  
</HTML>  
Connection closed by foreign host.  
[szhogart@mycin szhogart]$ █
```

A correct example --- giving the server the version number

```
szhogart@mycin:~  
[szhogart@mycin szhogart]$ telnet www.ucdavis.edu 80  
Trying 169.237.105.90...  
Connected to www.ucdavis.edu.  
Escape character is '^]'.  
GET / HTTP/1.1  
  
HTTP/1.1 400 Bad Request  
Date: Tue, 12 Mar 2002 08:56:52 GMT  
Server: Apache/1.3.22 (Unix)  
Connection: close  
Transfer-Encoding: chunked  
Content-Type: text/html; charset=iso-8859-1  
  
127  
<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML 2.0//EN">  
<HTML><HEAD>  
<TITLE>400 Bad Request</TITLE>  
</HEAD><BODY>  
<H1>Bad Request</H1>  
Your browser sent a request that this server could not understand.<P>  
client sent HTTP/1.1 request without hostname (see RFC2616 section 14.23): <P>  
</BODY></HTML>  
  
0  
  
Connection closed by foreign host.  
[szhogart@mycin szhogart]$  
  
[szhogart@mycin szhogart]$ █
```

HTTP 0.9 - 1991 - not an RFC

- ✦ Only defined very elemental aspects of the connection (port, the query command, server disconnects after serving up the file, etc..)
- ✦ Commands:
 - GET - The GET method means retrieve whatever information (in the form of an entity) is identified by the Request-URI.

HTTP 1.0 - 1992 (RFC)

- ★ GET
- ★ HEAD - The HEAD method is identical to GET except that the server must not return any Entity-Body in the response
- ★ POST - The POST method is used to request that the destination server accept the entity enclosed in the request as a new subordinate of the resource identified by the Request-URI. POST is designed to allow a uniform method to submit data to the server for processing (data entry).
- ★ Basic authentication protocol (AUTH)

HTTP 1.1 - refinement (1997)

- ★ Added robustness to the connection
 - persistent connections (Keep-alive)
 - multi-homing
 - allows a single web server to handle multiple domains
 - byte ranges
 - allowed clients to only retrieve a portion of an entity
- ★ More commands
 - UPLOAD, DELETE, etc..

“CGI” - 1995

- ★ CGI - common gateway interface (CGI) programming
 - Actually “invented” by a Netscape Mosaic engineer (not Marc Andreessen)
 - establishes a common mechanism for Web servers to interact with programs that are not part of the Web server --- Perl, C++, etc..
 - Instantly makes Web servers “gateways” to established “legacy” systems such as:
 - mainframes
 - database servers
 - any server-based program that can be invoked

Next generation web sites c.1997

★ CGI programs for dynamic HTML generation

- web pages don't exist -- created 'on the fly' by a program
- many done with "scripting" languages
- scripting languages popular at the time:
 - Perl
 - iBasic -- became Active Server Pages (ASP)
 - Database Markup Language (DBML) -- CFML



Evolution of dynamic HTML c.1999

- ✦ Notion of “personalized” web pages is born -- pages created dynamically in response to user’s preferences/profile
- ✦ Term “portal” is coined
- ✦ Portals = sites that serve different pages for different users dynamically
 - ✦ “personalization of the Web”
 - ✦ MyYahoo, MyNetscape, etc..



Enterprise Information Portals

- ✦ EIP's are just now becoming mainstream -- commercialization
- ✦ We built one of the first (1998)
 - ✦ 1st version built with Cold Fusion
 - ✦ Goal was to improve delivery and management of medical Web content at UCDCMC
 - ✦ Original vision was a “one-stop shopping” for information that was configured depending on the individual's group membership



Adaptive Portals - the next generation

- ✦ Semi-automatically improve their organization/presentation by learning from visitor access patterns
- ✦ Require the following:
 - ✦ document models
 - ✦ semantic “tagging” of content
 - ✦ clustering of “documents” or data based on similarity of access or content
 - access clusters
 - content clusters

An example: Amazon.com

- ✦ Amazon.com -- latest version
 - ✦ constructs Web pages with content that is adapted real-time from
 - recent browsing by the user (access clustering)
 - similar books/content viewed by others with similar profiles (access clustering)
 - books/content that is similar in content (content clustering)




Important issues for Adaptive Portals

- ✦ Document model
- ✦ Semantic tagging
- ✦ Clustering

Document model

★ Dublin Core

- ★ a metadata model for use with Web content
- ★ resulted from a meeting in Dublin, Ohio!
- ★ Provides a generic model
- ★ Core components
 - ★ Title, Author, Keywords, Description, Publisher, Resource Type, Format, Resource Identifier, Language, etc..



Dublin Core for Medicine: Medical Core Metadata (MCM)

- ★ Added some resource types: meeting, pathology images, radiology images, patient educational material, review, practice guidelines, etc...
- ★ Implemented of MeSH information in the Dublin Core Metadata

NOTE: Dr. Malet published the first "list of medical sites" on the Internet --- survives today as Medical Matrix (www.medmatrix.org)

Malet G, Munoz F, Appleyard R, Hersh W. J Am Med Inform Assoc 1999 Mar-Apr;6(2):163-72

Semantic Tagging - the “semantic web”

- ★ Semantic Web - coined by Tim Berners-Lee
 - entails adding “concept” or “meaning” information to Web content metadata
 - Would greatly enhance the ability to link Web content semantically --- or by meaning, rather than just by keyword or Web master arrangement
- ★ Highly recommended reading:
 - Scientific American, May 2001 Issue: “The Semantic Web”
<http://www.scientificamerican.com/2001/0501issue/0501bern-ers-lee.html>



Automated semantic tagging

- ★ Semantic information has traditionally added manually --- VERY costly and not practical for the way content is created today
 - Example -- MeSH “coding” by the NLM
 - Unclear whether it can be done consistently among various human indexers
- ★ Can it be done automatically and be beneficial -- even if not perfect?

A possible methodology

- ✦ Three necessary steps
 - ✦ Identify “concept containing” phrases in free text
 - ✦ Link concept phrases with concepts in an ontology
 - ✦ “Tag” the documents with the concept identifier(s) from the ontology

Conceptualized annotations...

- ✦ A current research endeavors in the Human Brain Project (Gorin, Gertz, Srinivas, Stone, Hogarth)
- ✦ Two relevant research endeavors
 - ✦ Identification of “semantically important” phrases in free text --- using part of speech (POS) tagger “trained” with the UMLS SPECIALIST Lexicon
 - to semi-automatically “prime” the concept domain in a particular knowledge domain
 - generic technique that can be applied to many things
 - ✦ Annotation Graph Model --- a model that links annotations, concepts, and resources

Next generation CRC

★ Use these techniques to

- ★ (1) semantically index medical knowledge sources
- ★ (2) use the semantic indexing and user-based clustering to offer semantically linked information for users

★ Current status

- ★ Implement an indexer/search engine in the portal engine (using Apache's Lucene - just underway)
- ★ Implement a terminology server (jTerm -- done)
- ★ Implement the conceptualized annotator in auto-tagging mode (pending research in HBP)