INFO/CS 4302
Web Information Systems

FT 2012

Week 3: The History and Architecture of the Web

(Lecture 4)

Theresa Velden

Acknowledgement

This lecture is a “slide mash-up” indebted to content of the lectures given by:

- Carl Lagoze (U Michigan, long-term teaching this course at Cornell)
- Bernhard Haslhofer, Fall 2010 at U Vienna, Austria

INTRO: (SHORT) HISTORY OF THE WORLD WIDE WEB

History of the Web

- Vision: Vannevar Bush's **MEMEX** envisioned in 'As We May Think', The Atlantic 1945
  - Visionary components and their realization by the World Wide Web

Associative Trails & Hyperlinking

The human mind does not work that way. It operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain. It has other characteristics, of course; trails that are not frequently followed are prone to fade, items are not fully permanent, memory is transitory. Yet the speed of action, the intricacy of trails, the detail of mental pictures, is awe-inspiring beyond all else in nature.

Man cannot hope fully to duplicate this mental process artificially, but he certainly ought to be able to learn from it. In minor ways he may even improve, for his records have relative permanency. The first idea, however, to be drawn from the analogy concerns selection. Selection by association, rather than by indexing, may yet be mechanized. One cannot hope thus to equal the speed and flexibility with which the mind follows an associative trail, but it should be possible to beat the mind decisively in regard to the permanence and clarity of the items resurrected from storage.

All this is conventional, except for the projection forward of present-day mechanisms and gadgetry. It affords an immediate step, however, to associative indexing, the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another. This is the essential feature of the memex. The process of tying two items together is the important thing.

Personal Library

thinner film will certainly be usable. Even under these conditions there would be a total factor of 10,000 between the bulk of the ordinary record on books, and its microfilm replica. The *Encyclopoedia Britannica* could be reduced to the volume of a matchbox. A library of a million volumes could be compressed into one end of a desk. If the human race has produced since the invention of movable type a total record, in the form of magazines, newspapers, books, tracts, advertising blurbs, correspondence, having a volume corresponding to a billion books, the whole affair, assembled and compressed, could be lugged off in a moving van. Mere compression, of course, is not enough; one needs not only to make

Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and to coin one at random, ``memex'' will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.

Most of the memex contents are purchased on microfilm ready for insertion. Books of all sorts, pictures, current periodicals, newspapers, are thus obtained and dropped into place. Business correspondence takes the same path. And there is provision for direct entry. On the top of the memex is a transparent platen. On this are placed longhand notes, photographs, memoranda, all sort of things. When one is in

New Forms of Encyclopedias

Wholly new forms of encyclopedias will appear, ready-made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified. The lawyer has at his touch the associated opinions and decisions of his whole experience, and of the experience of friends and authorities. The patent attorney has on call the millions of issued patents, with familiar trails to every point of his client's interest. The physician, puzzled by its patient's reactions, strikes the trail established in studying an earlier similar case, and runs rapidly through analogous case histories, with side references to the classics for the pertinent anatomy and histology. The chemist, struggling with the synthesis of an organic compound, has all the chemical literature before him in his laboratory, with trails following the analogies of compounds, and side trails to their physical and chemical behavior.

Ubiquitous Availability

Presumably man's spirit should be elevated if he can better review his shady past and analyze more completely and objectively his present problems. He has built a civilization so complex that he needs to mechanize his record more fully if he is to push his experiment to its logical conclusion and not merely become bogged down part way there by overtaxing his limited memory. His excursion may be more enjoyable if he can reacquire the privilege of forgetting the manifold things he does not need to have immediately at hand, with some assurance that he can find them again if they prove important.

History of Hypertext

- 1960's: start of research on hypertext and hypermedia
  - Ted Nelson (coined the term 'hypertext'), Xanadu
  - Douglas Engelbart (augment human intellect)
- Goals:
  - Free author from sequential writing
  - Provide autonomy to reader to chose
  - Augment thinking and reflection
- 1980': experimentation with various hypertext systems
- Features:
  - Machine-readable links from one chunk of text to another
  - Nodes and links that are simple navigable connections between nodes

Sources: http://www.mprove.de/diplom/text/2_hypertext.html, http://en.wikipedia.org/wiki/Proto-hypertext

Invention of The Web

‘CERN is a wonderful organisation. It involves several thousand people, many of them very creative, all working toward common goals. Although they are nominally organised into a hierarchical management structure, this does not constrain the way people will communicate, and share information, equipment and software across groups.’

Source: Tim Berners-Lee, CERN March 1989, May 1990 “Information Management: A Proposal, http://www.nic.funet.fi/index/FUNET/history/internet/w3c/proposal.html”

Photo: by dirtybronson

Invention of The Web

‘The actual observed working structure of the organisation is a multiply connected "web" whose interconnections evolve with time. In this environment, a new person arriving, or someone taking on a new task, is normally given a few hints as to who would be useful people to talk to. Information about what facilities exist and how to find out about them travels in the corridor gossip and occasional newsletters, and the details about what is required to be done spread in a similar way.’

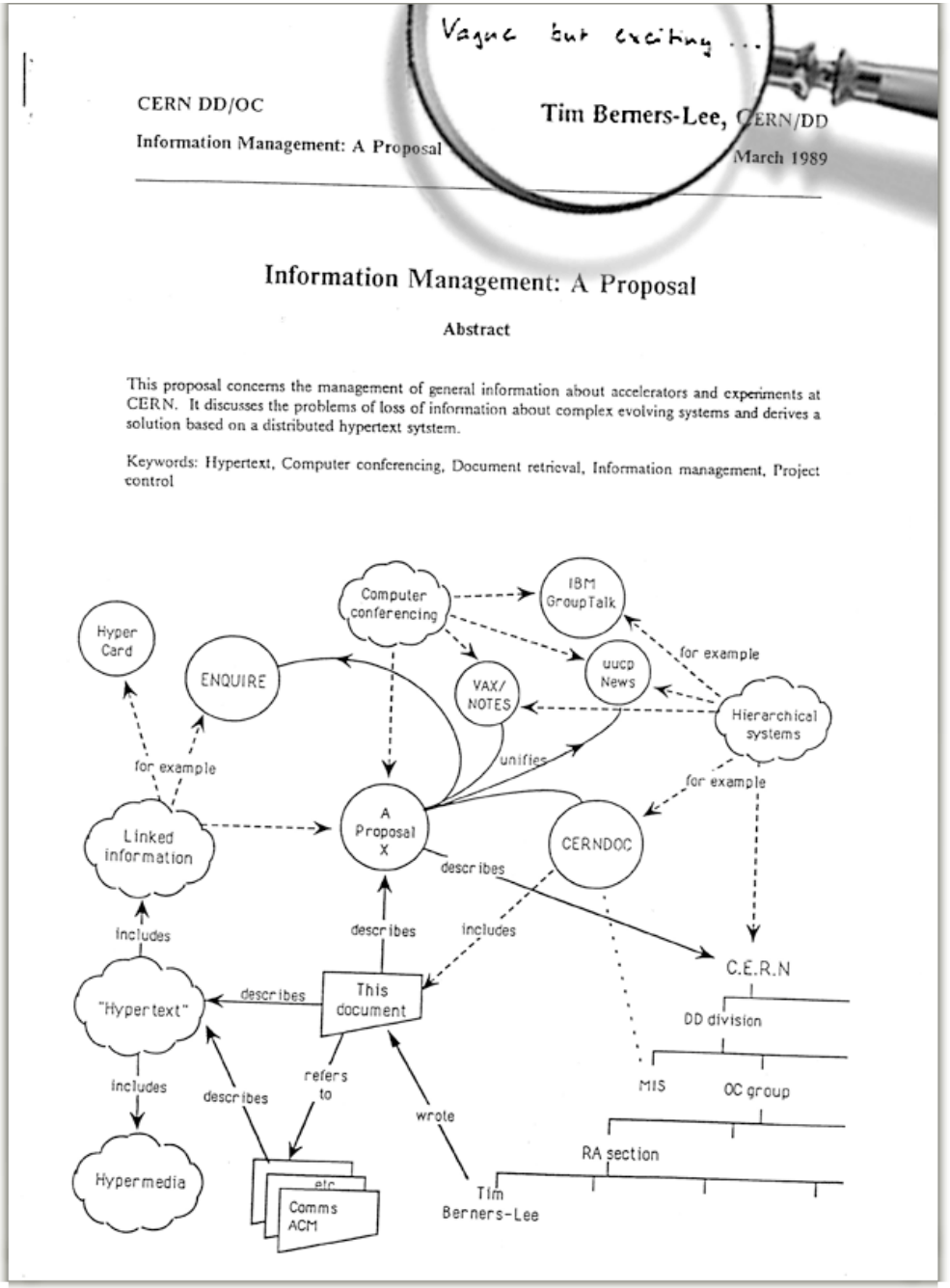
Source: Tim Berners-Lee, CERN March 1989, May 1990 “Information Management: A Proposal, http://www.nic.funet.fi/index/FUNET/history/internet/w3c/proposal.html”

Photo: by dirtybronson

Historical Proposal

- Turn-over of people makes finding existing information difficult
- CERN experiments constantly evolving (cannot be documented in a single book)
- allow a pool of information to develop which could grow and evolve with the organisation and the projects it describes
- Important requirements:
  - Remote access
  - Access is required to the same data from different types of system (VM/CMS, Macintosh, VAX/VMS, Unix)

http://www.w3.org/History/1989/proposal.html



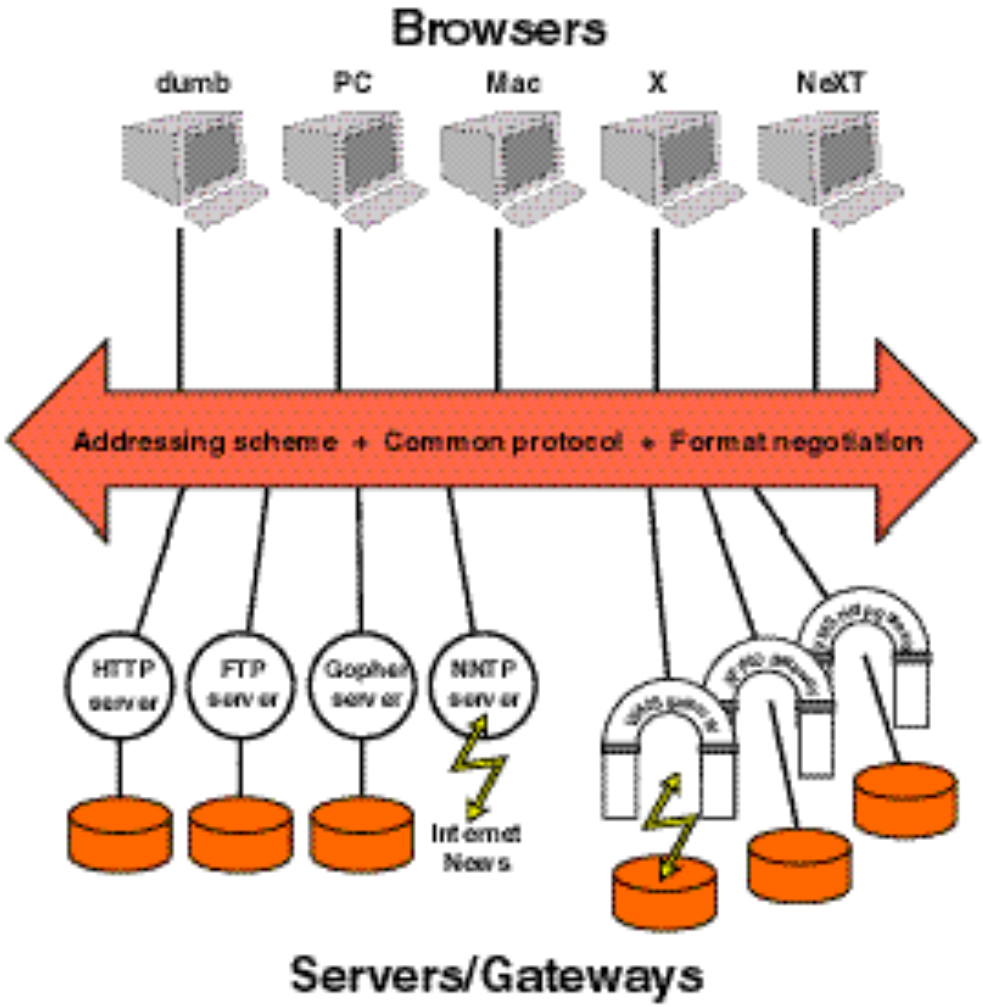
Source: CERN 2008 - Web Communications, DSU-CO

Revised Proposal

"HyperText is a way to link and access information of various kinds as **a web of nodes in which the user can browse at will**. It provides **a single user-interface** to large classes of information (reports, notes, data-bases, computer documentation and on-line help). We propose a simple scheme incorporating servers already available at CERN... A program which provides access to the hypertext world we call **a browser**... "

Tim Berners-Lee , R. Cailliau. 12 November 1990, CERN [6]

Early Sketch of Web Architecture



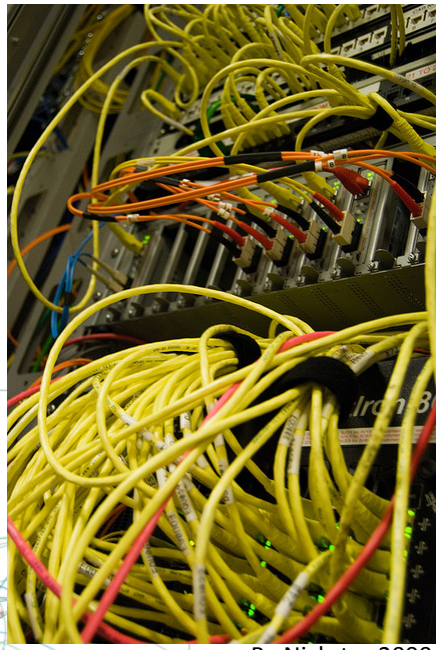
Tim Berner's Lee on the invention of the 'document web'

- http://www.youtube.com/watch?v=OM6XIICm_qo

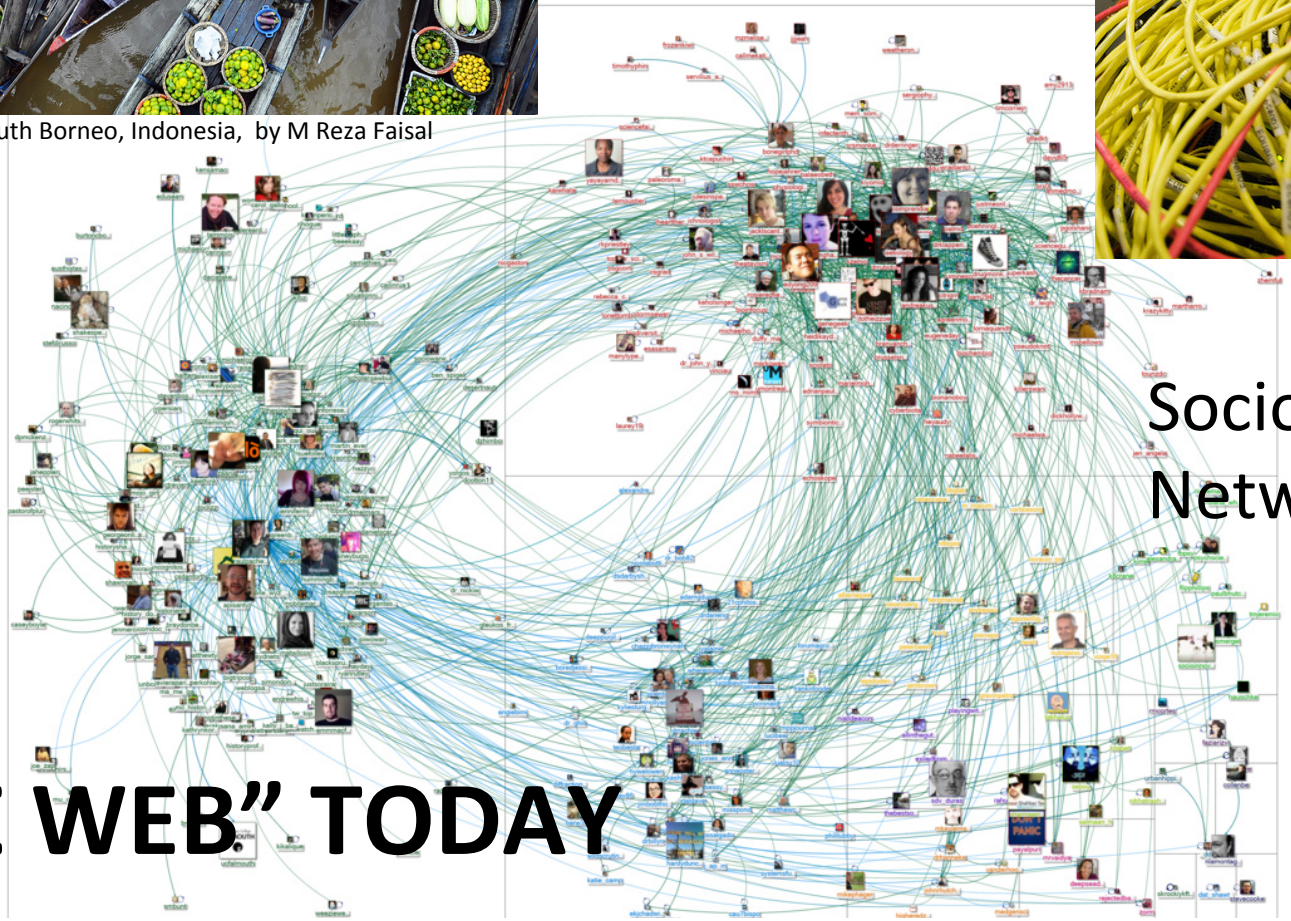
Social Network

Lok Baintan, South Borneo, Indonesia, by M Reza Faisal

Technical Network



By Nickster 2000

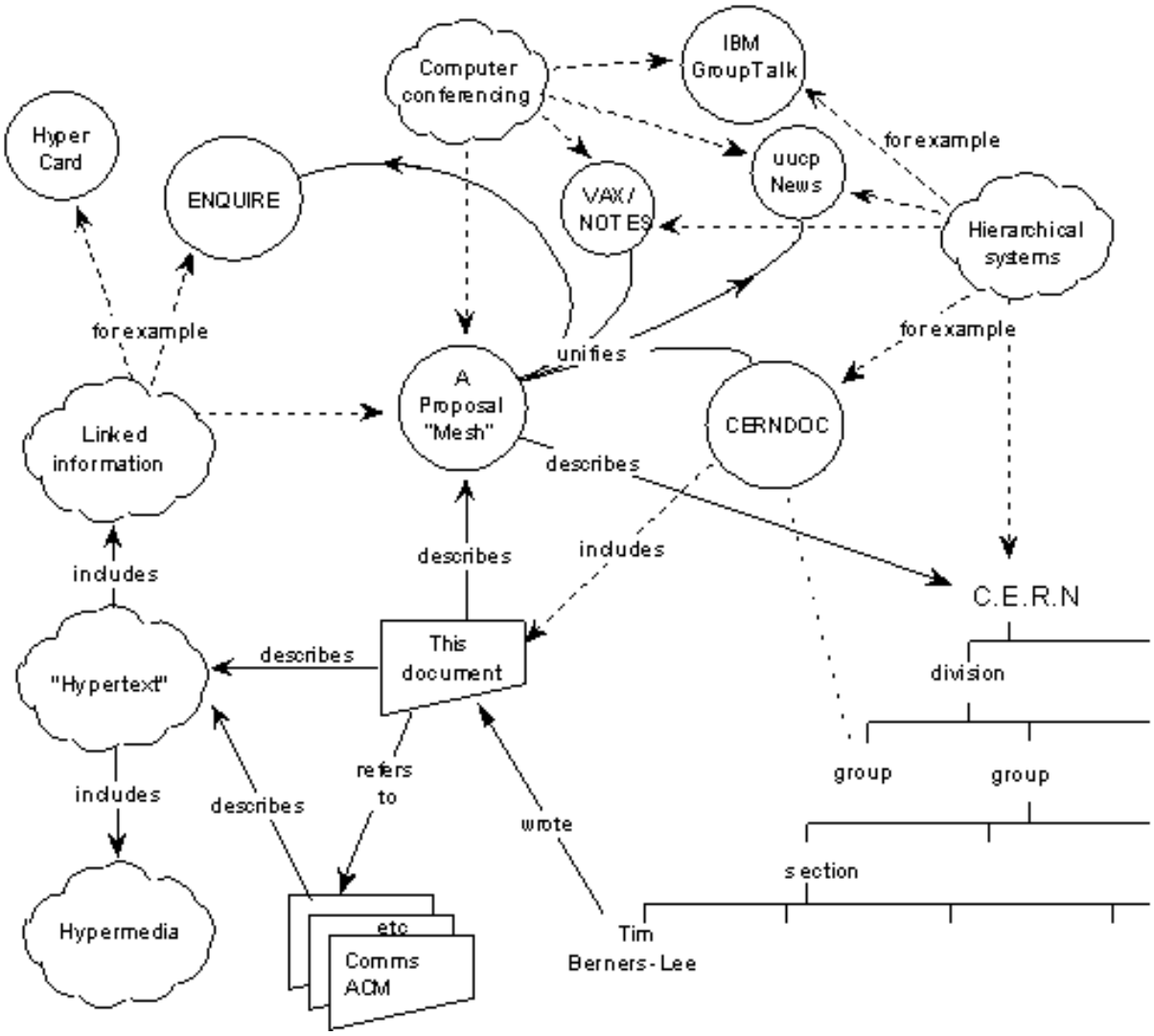


Socio-Technical Network

“THE WEB” TODAY

ARCHITECTURAL COMPONENTS OF THE WEB

The Web As A Graph



Source: Tim Berners-Lee, CERN March 1989, May 1990 “Information Management: A Proposal, http://www.nic.funet.fi/index/FUNET/history/internet/w3c/proposal.html”

The Web As A Graph

"We can call the circles nodes, and the arrows links. Suppose each node is like a small note, summary article, or comment. I'm not over concerned here with whether it has text or graphics or both. Ideally, it represents or describes one particular person or object. Examples of nodes can be

- People
- Software modules
- Groups of people
- Projects
- Concepts
- Documents
- Types of hardware
- Specific hardware objects"

Source: Tim Berners-Lee, CERN March 1989, May 1990 "Information Management: A Proposal, http://www.nic.funet.fi/index/FUNET/history/internet/w3c/proposal.html"

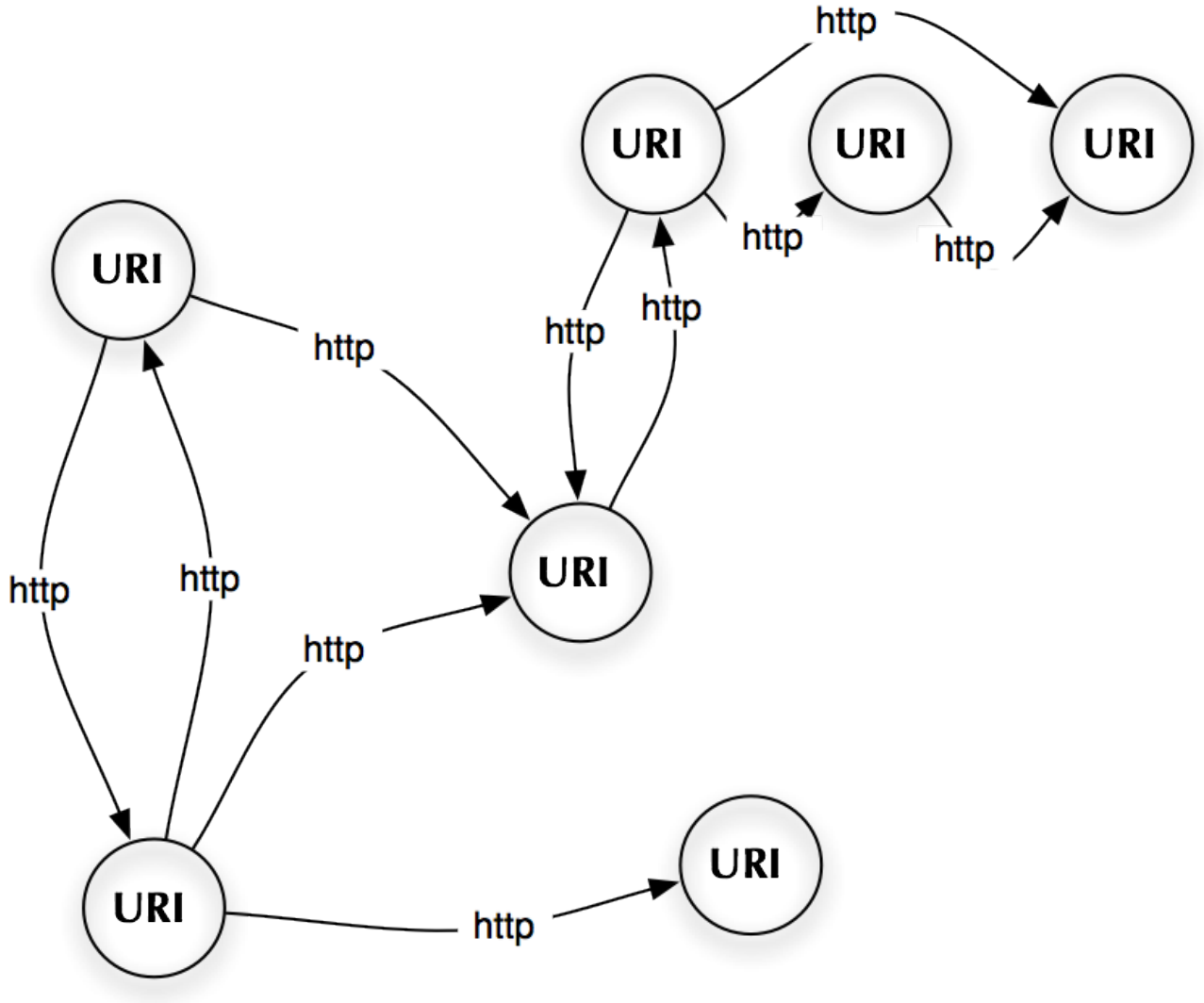
The Web As A Graph

"The arrows which links circle A to circle B can mean, for example, that A...

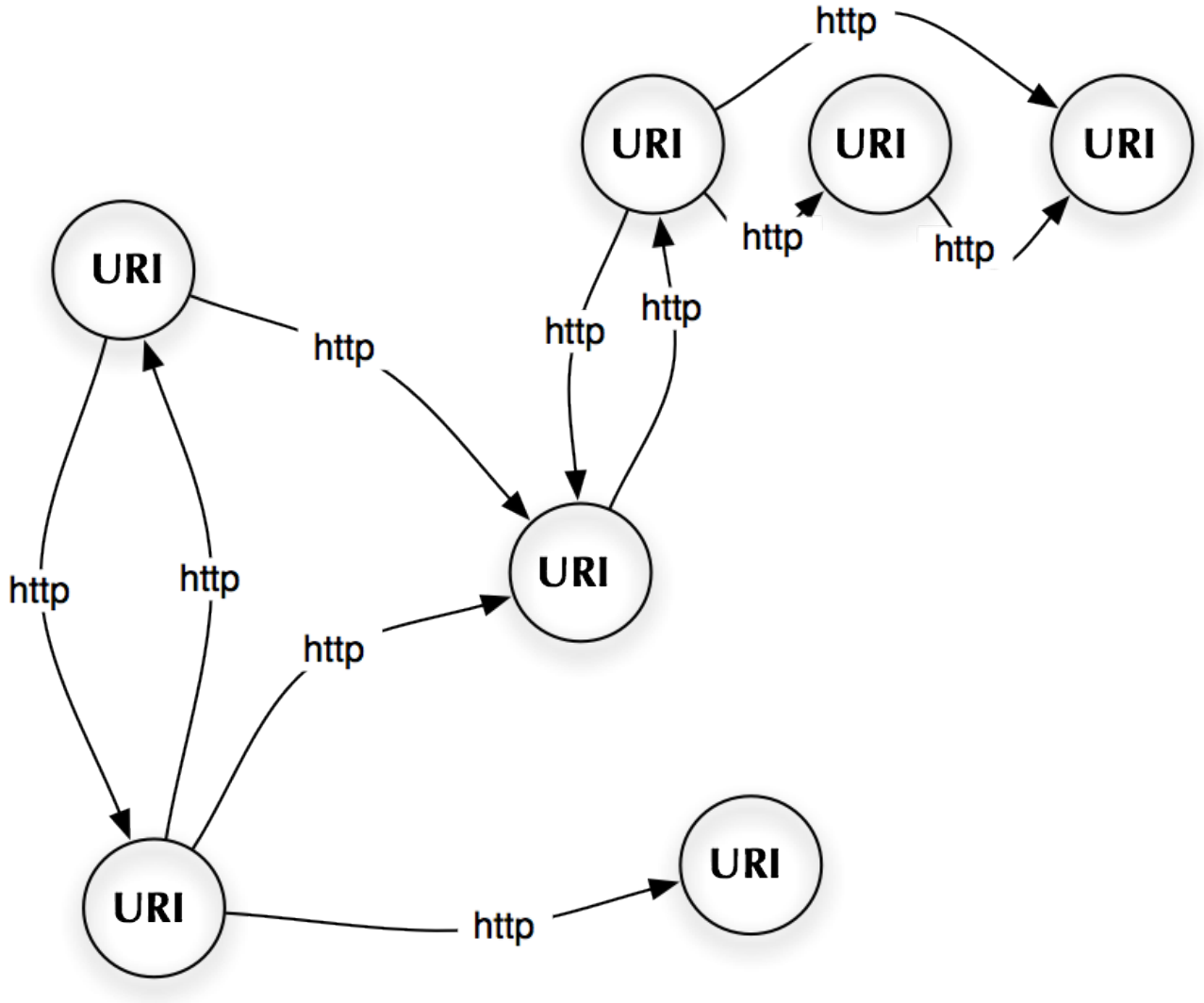
- depends on B
- is part of B
- made B
- refers to B
- uses B
- is an example of B"

Source: Tim Berners-Lee, CERN March 1989, May 1990 "Information Management: A Proposal, http://www.nic.funet.fi/index/FUNET/history/internet/w3c/proposal.html"

Universal Web Architecture



(Naiive View of) The Web Architecture



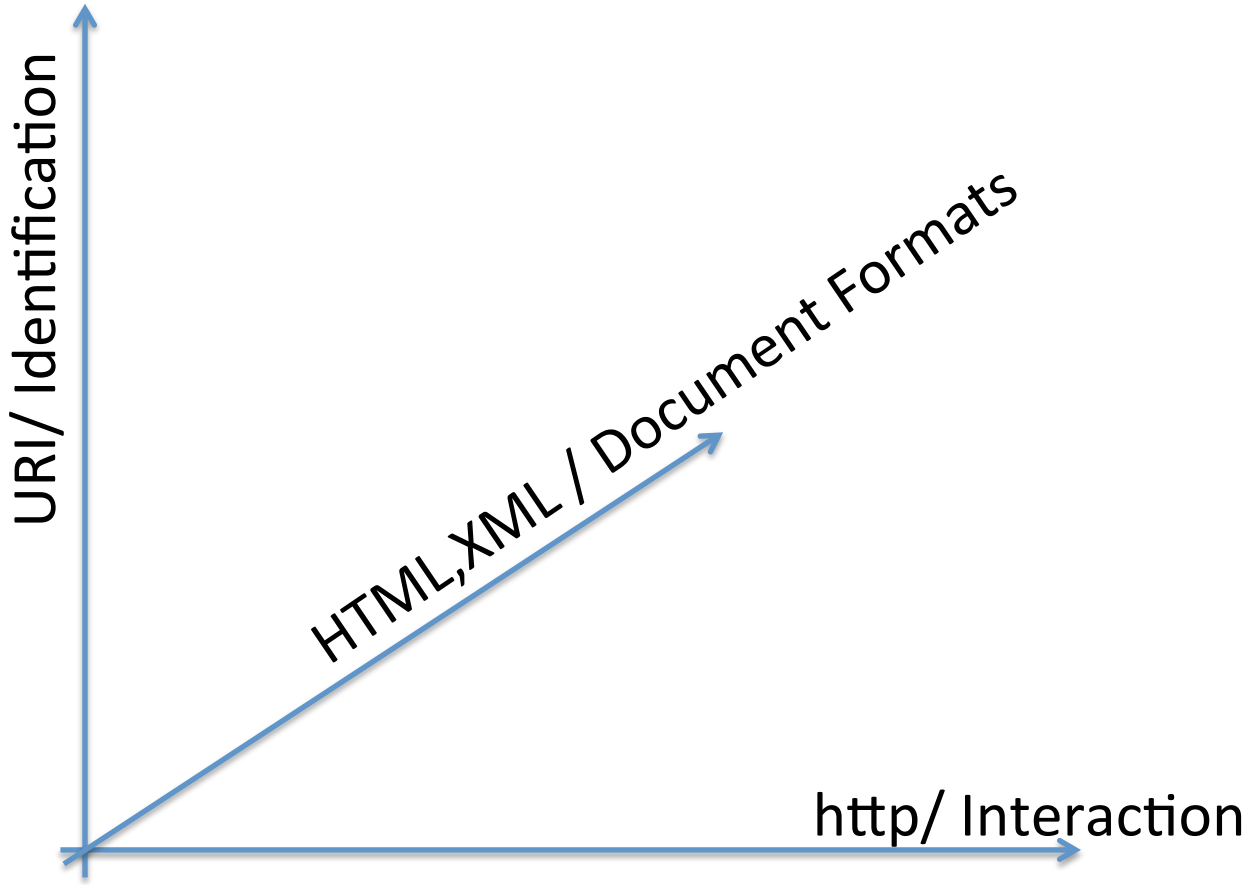
Key Architectural Components

1. **Identification** (URI) of web resources
2. **Interaction** (http) with web resources
3. **Standardized Document Formats** (HTML, XML)

W3C's Technical Architecture Group (TAG) released in 2004 a document describing architectural issues of the World Wide Web: http://www.w3.org/TR/webarch/

Note: notions of site, server, repository are **not** in the architecture

Principle ‘Orthogonal Specifications’



Key Architectural Components

1. **Identification (U**niform **R**esource **I**dentifier**) of web resources**
2. **Interaction (h**yper **t**ext **t**ransport **p**rotocol**) with web resources**
3. Standardized **Document Formats** (HTML, XML) [FOCUS OF NEXT WEEK]

IDENTIFICATION

Identification of what?

A **resource**

- is an entity that can be identified by a URI
- is an abstract concept: we cannot see, smell, touch, examine a resource
- is not necessarily retrievable through the internet
- can be non-digital (a person, a University, a book)
- or digital (personal homepage, blog post, online newspaper article) URI < - > URL

Resource-Centric Architecture

- Abstract resources: their essence is not information
  - E.g. a dog, concepts,…
  - http://www.example.com/ontology/meter
- Informational resources: their essential characteristics can be conveyed in a message
  - Web pages, images, technical specifications
  - E.g. http://www.flickr.com/user1/photos/image123.jpg

Uniform Resource Identifier (URI)

- "URIs are a cornerstone of Web architecture, providing identification that is common across the Web" [W3C TAG, 2004]
- A URI is a compact sequence of characters that identifies an abstract or physical resource. **[RFC 3986]**
- A resource's URI is independent of access mechanism or document format of its representation
- Provide for 'network effect' to happen
- Resource identification
  - A resource is identified by one (or several) URI
  - A URI only identifies one resource

URI Syntax

- Federated and extensible (IANA scheme registry)
- organized hierarchically, components listed in order of decreasing significance from left to right

  <scheme name> : <hierarchical part> [? <query>] [# <fragment>]

- Examples of URI from different schemes:

```
ftp://ftp.is.co.za/rfc/rfc1808.txt
      -- ftp scheme for File Transfer Protocol services
http://www.math.uio.no/faq/compression-faq/part1.html
      -- http scheme for Hypertext Transfer Protocol services, RFC 2616
mailto:mduerst@ifi.unizh.ch
      -- mailto scheme for electronic mail addresses
file:///path
      -- addressing files on local or networked system, RFC 1738
telnet://melvyl.ucop.edu/
      -- telnet scheme for interactive services via the TELNET Protocol
```

- Each scheme may further restrict syntax and semantics of URI identifiers
  - <scheme name> : <hierarchical part> [? <query>] [# <fragment>]

scheme-specific

Uniform Resource Identifier (URI)

"A common misunderstanding of URIs is that they are only used to refer to accessible resources. **The URI itself only provides identification; access to the resource is neither guaranteed nor implied** by the presence of a URI. Instead, any operation associated with a URI reference is defined by the protocol element, data format attribute, or natural language text in which it appears."

[RFC 3986]

Uniform Resource Identifier (URI)

"However, we do use a few general terms for describing common operations on URIs. **URI "resolution"** is the process of determining an access mechanism and the appropriate parameters necessary to dereference a URI; this resolution may require several iterations.

To use that access mechanism to perform an action on the URI's resource is to **"dereference" the URI.**

[RFC 3986]

Are URIs = URLs?

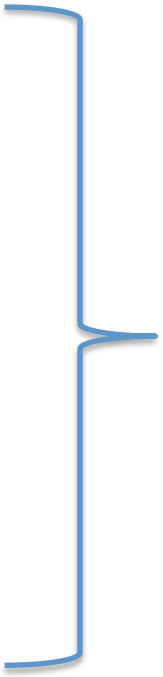
- Some are
- A URL (Uniform Resource Locator) is a type of URI that identifies a resource via a representation of its primary access mechanism (e.g., its network "location")
  - Other URIs do not: e.g. urn:isbn:9789521061547
- So a subset of URIs are URLs
- http URIs are URLs (as are ftp URIs)
- That said, the term 'URL' is considered informal and outdated

Source: http://www.w3.org/TR/uri-clarification/#confusion

URI Equivalence

- In principle, URIs equivalent if they identify the same resource
- Practically: based on string comparison and rules defined by scheme definitions (for http scheme: http://www.ietf.org/rfc/rfc2616.txt)

http://foo.com
http://foo.com/
http://foo.com:80
http://www.foo.com
http://132.24.3.1
http://foo.com/index.html
/index.html
http://foo.com?a=b&c=d



Equivalent http URIs

URI Types

- Descriptive: conveys some notion of the resource, e.g.
  - http://www.cs.cornell.edu/People/faculty/
  - http://www.cs.cornell.edu/People/faculty/~bailey/
- Opaque: convey no semantics
  - http://tinyurl.com/822zwtf

URI Allocation

- **URI ownership** is a relation between a URI and a social entity, such as a person, organization, or specification.
- URI ownership is delegated from the IANA (Internet Assigned Numbers Authority) URI scheme registry (http://www.iana.org/assignments/uri-schemes.html) to IANA-registered URI scheme specifications
- For the HTTP URI (URL) scheme, the Internet community delegates authority via the **Domain Name Service (DNS)** to a particular owner
  - E.g. The Wikimedia Foundation Inc. owns domain name 'wikipedia.org':
    - http://en.wikipedia.org/wiki/The_Shining
    - http://en.wikipedia.org/wiki/The_Shining_(film)
    - http://en.wikipedia.org/wiki/Shining_(Norwegian_band)
- URI owners are responsible for avoiding the assignment of equivalent URIs to multiple resources

[based on: Bernhard Haslhofer, U Vienna, Austria]

INTERACTION

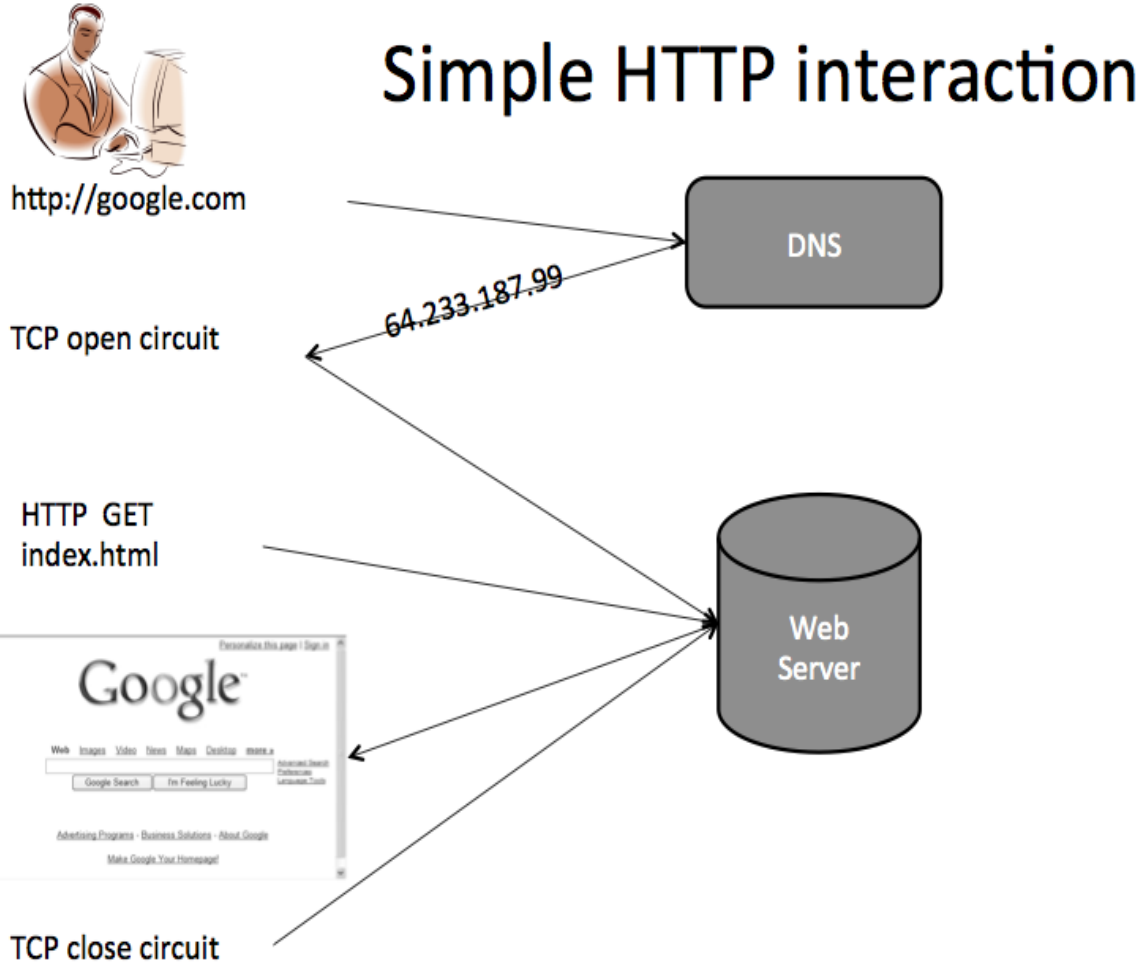
Hypertext Transfer Protocol (http)

- HTTP is the foundation for data communication for the World Wide Web
- E.g. dereferencing a URI and retrieving a presentation of a web resource
- HTTP is a **request-response** protocol and basis for interaction between web agents and servers
- HTTP/1.1 is defined in RFC 2616 http://www.iex.org/rfc/rfc2616.txt
- Layered on top of TCP and uses DNS
- Text based
  - All messages and requests are human readable
- Stateless
  - No persistent client server connection
  - All state carried in protocol request/reply (cookie)

[based on: Bernhard Haslhofer, U Vienna, Austria]

http session:
sequence of request-response

- an HTTP client initiates a request
- it uses DNS to resolve domain name
- it establishes a TCP connection to a particular port (typically 80) on a host (e.g. google.com)
- an HTTP Server listening on that port waits for a clients request message
- upon receiving the request, the server sends back a status line (e.g., "HTTP/1.1 200 OK") and a message of its own (body, error message, some other information)



[indebted to Bernhard Haslhofer & Carl Lagoze]

http Verbs

Desired action to be performed on the identified resource

1. **HEAD**: Asks for the response identical to the one that would correspond to a GET request, but without the response body.
2. **GET**: Requests a representation of the specified resource
3. **POST**: Submits data to be processed (e.g., from an HTML form) to the identified resource.
4. **PUT**: Uploads a representation of the specified resource.
5. **DELETE**: Deletes the specified resource.
6. **TRACE**: Echoes back the received request, so that a client can see what (if any) changes or additions have been made by intermediate servers.
7. **OPTIONS**: Returns the HTTP methods that the server supports for the specified URL.
8. **CONNECT**: Converts the request connection to a transparent TCP/IP tunnel, usually to facilitate SSL-encrypted communication (HTTPS) through an unencrypted HTTP proxy.
9. **PATCH**: Is used to apply partial modifications to a resource.

[based on: Bernhard Haslhofer, U Vienna, Austria]

http Status Codes

- the first line of the HTTP response is called the **status line** and includes a numeric **status code** and a textual **reason phrase**
  - e.g., ”HTTP/1.1 404 Not Found” or , ”HTTP/1.1 200 OK”
  - the status code is intended for the user agent (e.g., browser)
  - the reason phrase for human interpretation

http://www.flickr.com/photos/girliemac/sets/72157628409467125

http Status Codes

- **1xx: informational**
- 2xx: successful
- 3xx: redirection
- 4xx: client error
- 5xx: server error

This means that the server has received the request headers, and that the client should proceed to send the request body (in the case of a request for which a body needs to be sent; for example, a POST request).

http://www.flickr.com/photos/girliemac/sets/72157628409467125

http Status Codes

- 1xx: informational
- **2xx: successful**
- 3xx: redirection
- 4xx: client error
- 5xx: server error

Standard response for successful HTTP requests.

http://www.flickr.com/photos/girliemac/sets/72157628409467125

http Status Codes

- 1xx: informational
- 2xx: successful
- **3xx: redirection**
- 4xx: client error
- 5xx: server error

This and all future requests should be directed to the given URI.

http://www.flickr.com/photos/girliemac/sets/72157628409467125

http Status Codes

- 1xx: informational
- 2xx: successful
- 3xx: redirection
- **4xx: client error**
- 5xx: server error

The requested resource could not be found but may be available again in the future. Subsequent requests by the client are permissible.

http://www.flickr.com/photos/girliemac/sets/72157628409467125

http Status Codes

- 1xx: informational
- 2xx: successful
- 3xx: redirection
- 4xx: client error
- **5xx: server error**

The server is currently unavailable (because it is overloaded or down for maintenance). Generally, this is a temporary state.

http://www.flickr.com/photos/girliemac/sets/72157628409467125

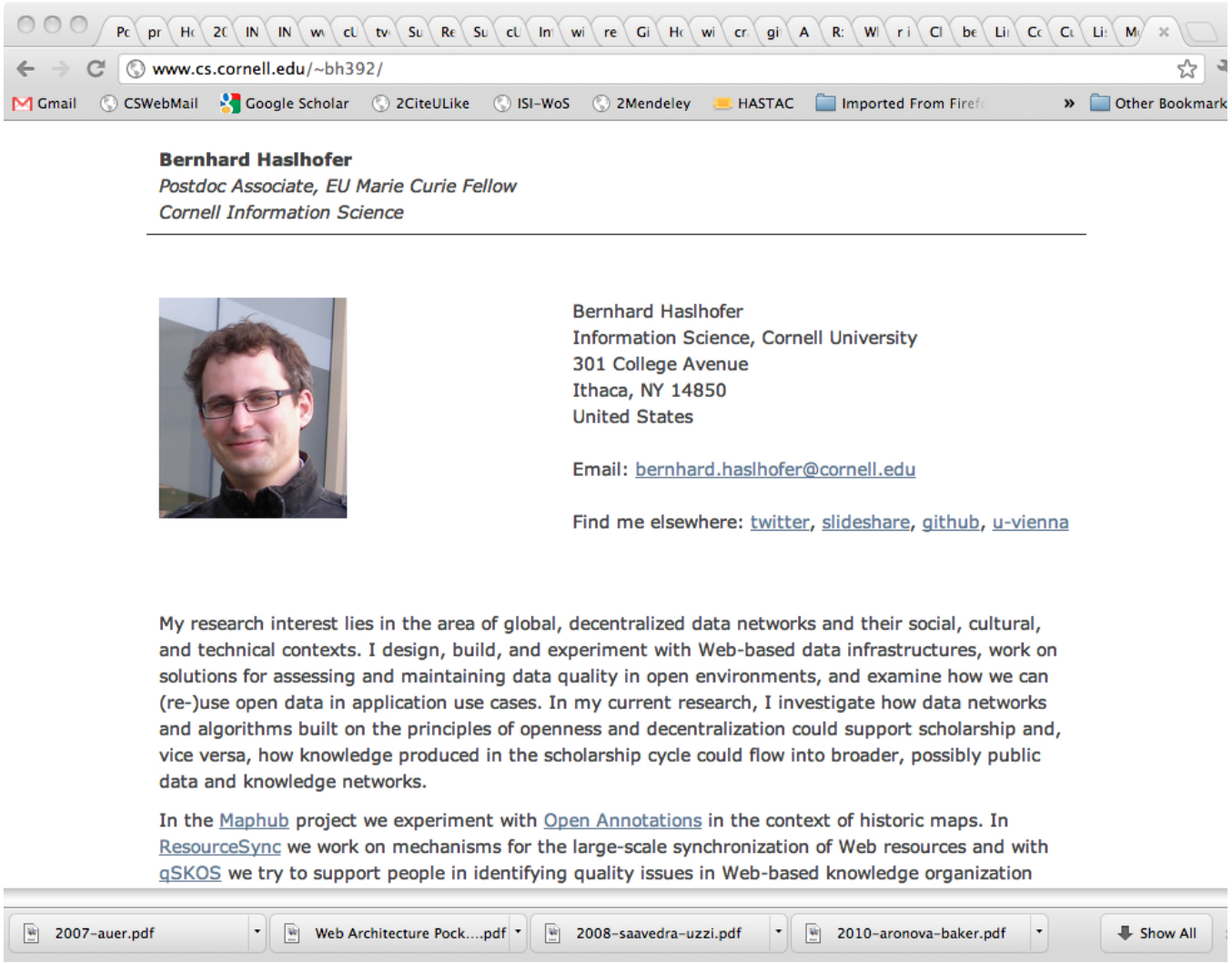
Status Codes Defined (- the cats)

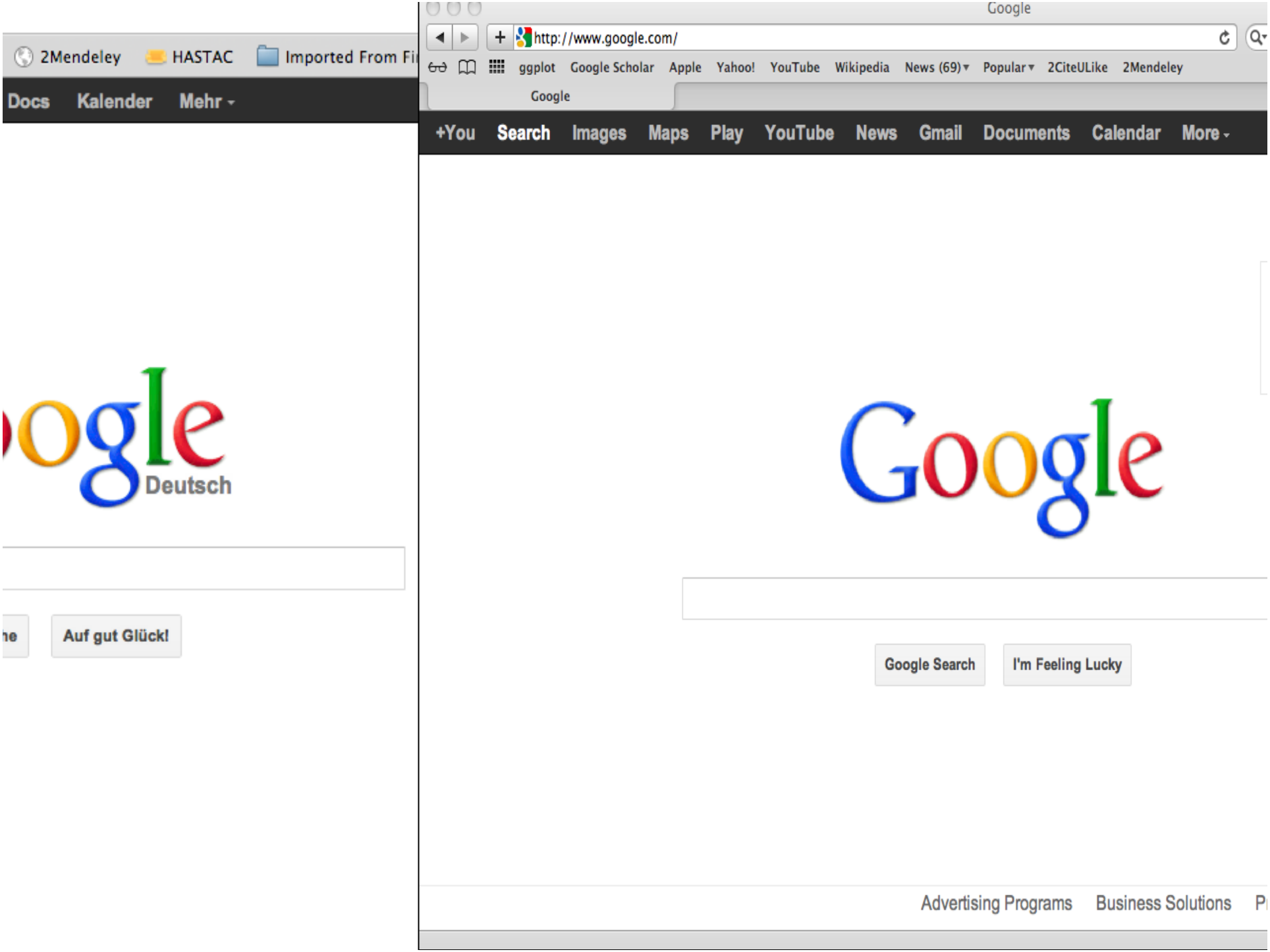
- http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html
- educational material: http://www.flickr.com/photos/girliemac/sets/72157628409467125

URI + http GET Request

= what gets delivered?

One web resource, several representations



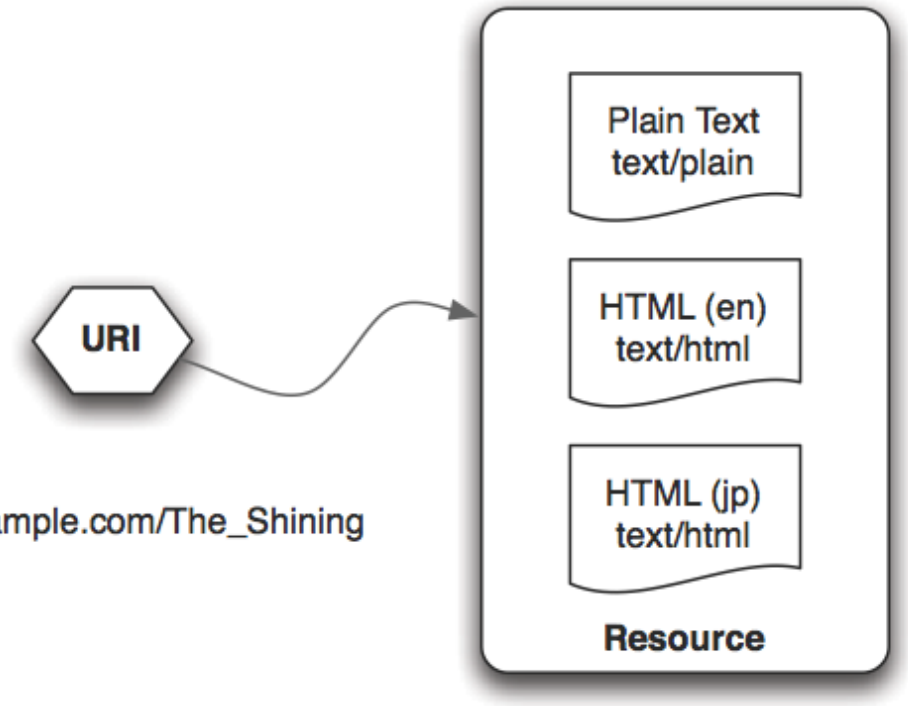

2Mendeley
HASTAC
Imported From Fi
Docs
Kalender
Mehr
oogle
Deutsch
he
Auf gut Glück!
Google
http://www.google.com/
ggplot
Google Scholar
Apple
Yahoo!
YouTube
Wikipedia
News (69)
Popular
2CiteULike
2Mendeley
Google
+You
Search
Images
Maps
Play
YouTube
News
Gmail
Documents
Calendar
More
Google
Google Search
I'm Feeling Lucky
Advertising Programs
Business Solutions

Content Negotiation

- **A URI is not a filename**
- Content negotiation refers to the practice of **making available multiple representations via the same URI**
- *"A resource representation encodes information about resource state"* [W3C Recommendation: Web Architecture (vol. 1) http://www.w3.org/TR/webarch/]
- **Basic Idea:** serve the best variant of the representation of a resource based on
  - What variants are available on the web server
  - What the user clients can accept and with what preference (accept headers of http request message)

Content Negotiation

- Main application areas
  - **format negotiation**: serve images in different formats (png, svg, etc.)
  - **language negotiation**: serve representations of a resource in several languages



[based on: Bernhard Haslhofer, U Vienna, Austria]

Content types: Internet Media Type

- Originally called a MIME type (Multipurpose Internet Mail Extensions)
- Defined in RFC 2046 & managed by IANA
- Two parts: a type, a subtype and optionally parameters:
  - image/svg+xml
  - text/html
  - text/plain
  - application/pdf

[based on: Carl Lagoze INFO/CS 4301, Fall 2011]

Content Negotiation Implementation

- **Server-driven Negotiation:** selection of the best representation for a response is made by an algorithm located at the server
  - disadvantage: server cannot guess the "best" representation for all possible clients
- **Agent-driven Negotiation:** selection of the best representation for a response is performed by the user agent after receiving an initial response from the server
  - Selection is based on a list of the available representations of the response, each representation identified by its own URI
  - Selection may be performed automatically or manually
  - disadvantage: needs a second request
- **Transparent Negotiation:** a combination of both server-driven and agent-driven negotiation

Possibilities defined in RFC 2616, Section 12

Content Negotiation & Resource Representations

- One **cannot** determine the type of a resource representation by inspecting a URI for that resource.
  - For example: http://example.com/page.html provides no guarantee that representations of the identified resource will be served with the Internet media type "text/html"
  - the URI owner is free to configure the server to return a representation using PNG or any other data format

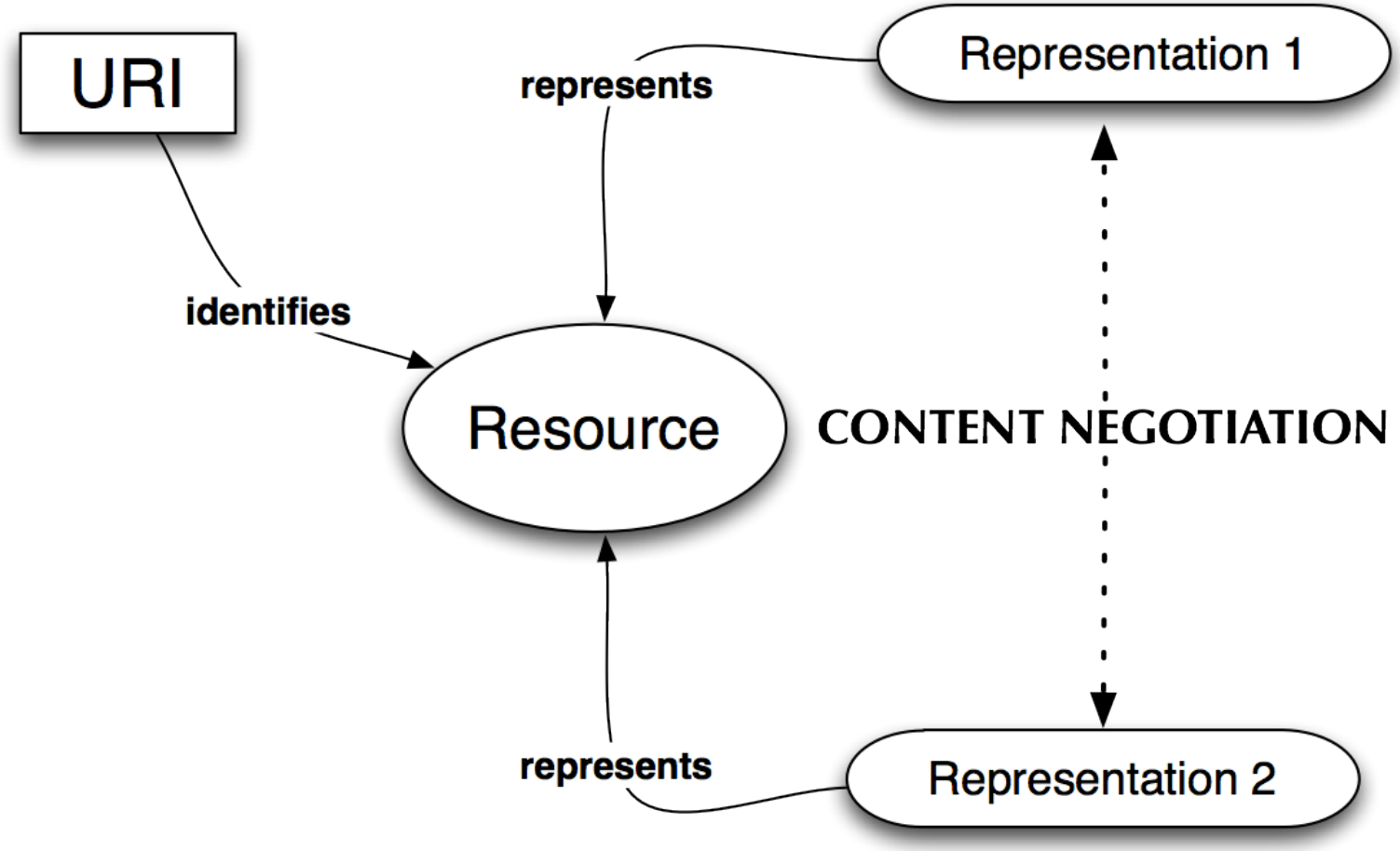
[based on: Bernhard Haslhofer, U Vienna, Austria]

Content Negotiation

Think about it:

The web is completely ephemeral, based on representations that are time and agent dependent.

Refined View of The Web Architecture



Warning: overuse of content negotiation can be bad for the web's health

Time Berners-Lee - "Cool URIs"

[based on Carl Lagoze INFO/CS 4301, Fall 2011]

Representation Management

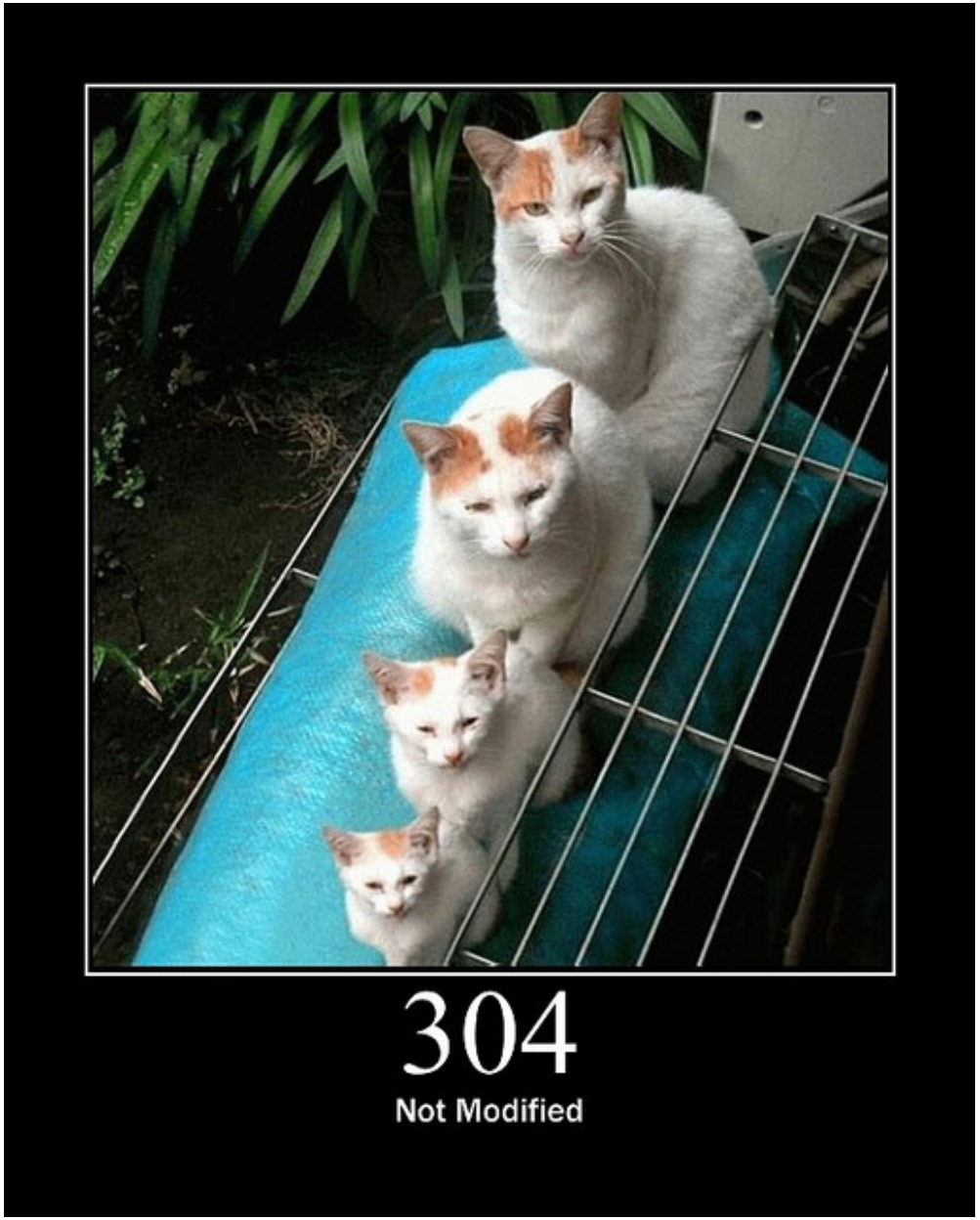
- URI persistence
  - once associated with a resource, a URI should continue indefinitely to refer to that resource
  - is a matter of policy commitment on the part of the URI owner
- HTTP helps to manage URI persistence via 3xx status codes
  - E.g. 301 Moved Permanently, 303 See Other, 305 Use Proxy, 307 Temporary Redirect
- Content Negotiation also supports consistency: a site manager is not required to define new URIs when adding support for a new format specification

[based on: Bernhard Haslhofer, U Vienna, Austria]

WRAP-UP

Web Architecture Essentials

- Three core components
  - Identification, interaction, standardized documents formats (html, xml: next week)
- Uniform standards connecting 'everything'
  - Uniform identification (URI)
  - Uniform interface (http), coordination by status codes
- Content negotiation
  - One web resource per URI
  - Several representations per web resource possible



See you Thursday
for http
in action & some
useful tools

http://www.flickr.com/photos/girliemac/sets/72157628409467125