Web Engineering

Prof. Dr. Dr. h.c. mult. Gerhard Krüger, Albrecht Schmidt

Universität Karlsruhe
Fakultät für Informatik
Institut für Telematik

Wintersemester 2000/2001

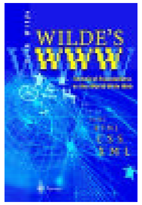
Organisation

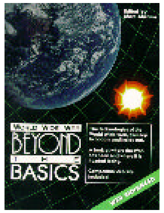
- □ Art der Veranstaltung: Vorlesung, 2 SWS
- □ Dozenten: Prof. Dr. Dr. h.c. mult. Gerhard Krüger
  Dipl. Inf. Albrecht Schmidt, MSc
- □ Ort: Raum -101 im Informatikneubau
- □ Zeit:Freitags von 8.00-9.30 Uhr
- □ Beginn: 20.10.1999
- □ Prüfbar: Ja, 2 SWS, Informatik und Informationswirtschaft
- □ Sprache: Vorlesung in Deutsch, Folien in Englisch

- □ WWW: http://www.teco.uni-karlsruhe.de/lehre/webe/
- □ Email: albrecht@teco.uni-karlsruhe.de
- □ Mailingliste?

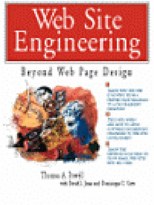
Books

Erik Wilde. Wilde's WWW.
Springer 1998.





Marc Abrams (Editor).
World Wide Web Beyond the Basics.
Prentice Hall 1998.



Thomas A. Powell.
Web Site Engineering.
Prentice Hall 1998.

Further Information

- □ FAQ = Frequently Asked Question
  - questions and the answers that cover basics on a topic
  - e.g. on programming, software setup and usage, etc.
- □ RFC = Request For Comment
  - Internet Standards
  - e.g. protocols, languages, cryptography, etc.
- □ White Papers
  - often provided by companies (from advertisement to technical paper)
  - describing protocols, architecture, systems, products, etc.
- □ WWW
  - W3C - www.w3.org (the www consortium)
  - catalogs (z.B. www.yahoo.com, web.de)
  - search engines (z.B. www.northernlight.com, www.altavista.com www.alltheweb.com, www.google.com)
  - links provided at www.teco.uni-karlsruhe.de/lehre/webe

You will gain ...

- a systematic understanding of the phenomenon WWW
- an in-depth understanding of the technical foundations of the world wide web
- an overview on the WWW as information and communication system as well as a business platform
- the ability to systematically select technologies and design WWW-applications

Web Engineering

Chapter 1: Introduction and Overview

What is the World Wide Web?

- Definitions in literature
  - „an internet-wide distributed hypermedia information retrieval system“ [Liu et al. 1994]
  - „the World Wide Web is a global, seamless environment in which all information (text, images, audio, video, computational services) that is accessible from the Internet and can be accessed in a consistent and simple way by using a standard set of naming and access conventions“ [WebMaster Magazine 1996]
  - „the World Wide Web (known as "WWW', "Web" or "W3") is the universe of network-accessible information, the embodiment of human knowledge“ [W3C 1999]
- In this course we will
  - look at the Web from different angles
  - show the „The Big Picture“

Ideas and Goals of Web

- finding information using a uniform addressing method
- uniform access (read and write) using a standard user interface not bound to a specific system
- display, visualize and share content (hypermedia documents) over different computer platforms
- integrate external information sources (e.g. legacy software, databases)
- support transactions as foundation for interactive applications (Client/Server)
- everyone can add information to the WWW
- inherent distribution

History

- □ 1945 Memex (Vannevar Bush)
- □ 1961 Packet switching (Leonard Kleinrock)
- □ 1965 Terms Hypertext und Hypermedia (TedNelson)
- □ 1969 ARPANET (with 4 points)
- □ 1974 TCP (Vinton Cerf , Bob Kahn; replaced NCP 1982)
- □ 1981 Xanadu (TedNelson)
- □ 1983 Term Internet
- □ 1989 World Wide Web (Berners-Lee, Cailliau; Release 1991)
- □ 1993 Mosaic Browser (Web has 341634% annual growth rate)
- □ 1995 Web has higher transfer volume than FTP, Sun releases JAVA
- □ ... The WWW becomes vital to some/many businesses
- □ ... M-Commerce, viruses, NAPSTER, domain name hijackings, denial of service attacks, cyberwar, intra-day trading, ...

About Statistics

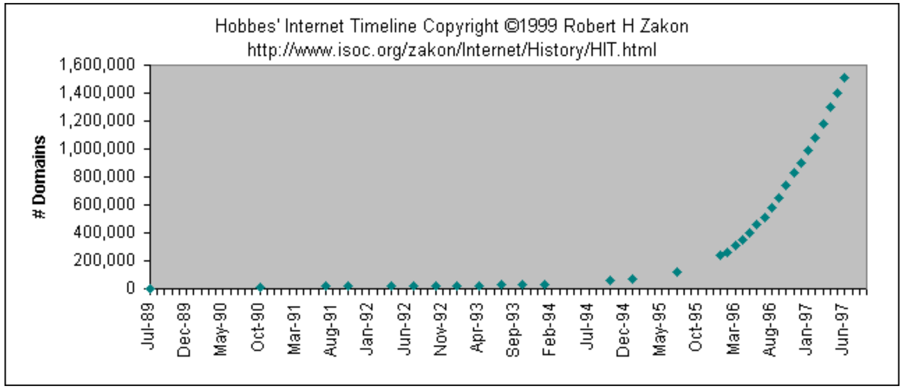
- □ All numbers that are out are estimates!

„The Internet is distributed by nature. This is its strongest feature, since no single entity is in control ...“

[Marc Abrams (Editor). World Wide Web Beyond the Basics. Prentice Hall 1998. Seite 40]

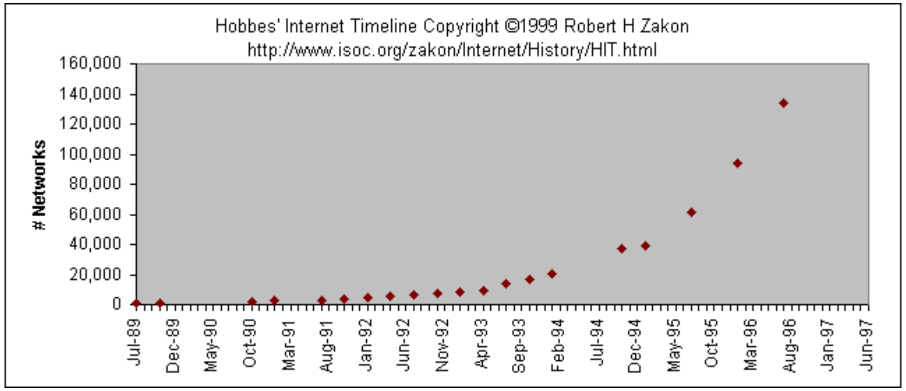
Growth of the Internet I

- □ Number of domains (from [Hobbes' Internet Timeline v4.0, 1999])



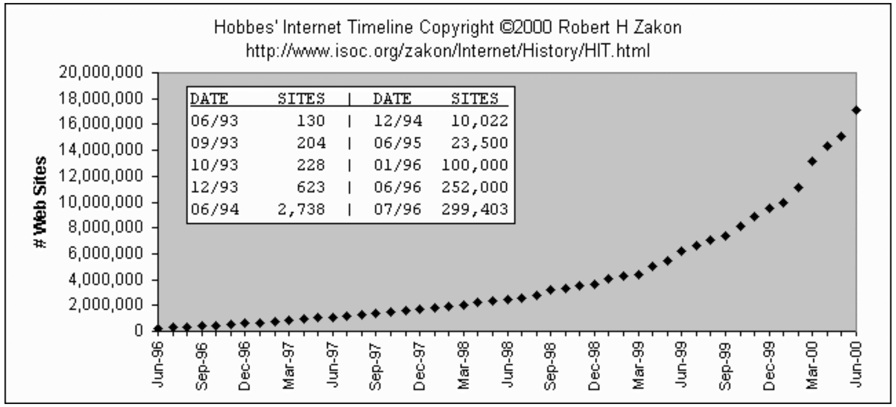
Growth of the Internet II

- □ Number of networks (from [Hobbes' Internet Timeline v4.0, 1999])



Growth of the Web

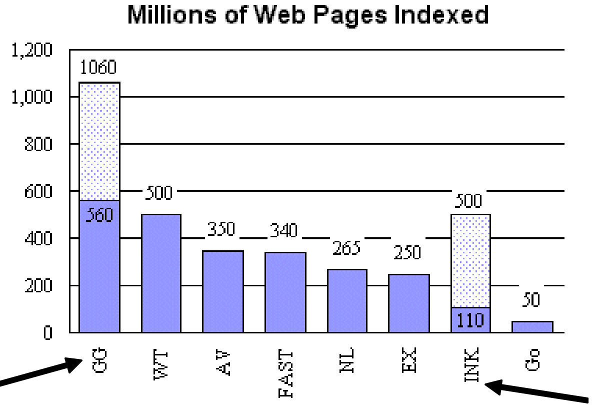
□ Number of Web Sites (from [Hobbes' Internet Timeline v5.1, 2000])



DATE	SITES	DATE	SITES
06/93	130	12/94	10,022
09/93	204	06/95	23,500
10/93	228	01/96	100,000
12/93	623	06/96	252,000
06/94	2,738	07/96	299,403

Size of the Web I (07.07.00)

□ Pages indexed in search engines



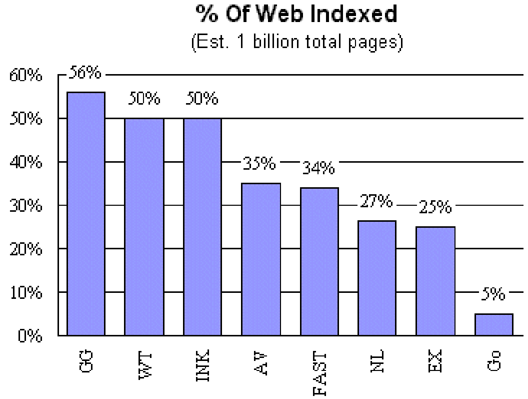
uses link information

two databases

source (19/10/00): http://www.searchenginewatch.com/reports/sizes.html

Size of the Web II (07.07.00)

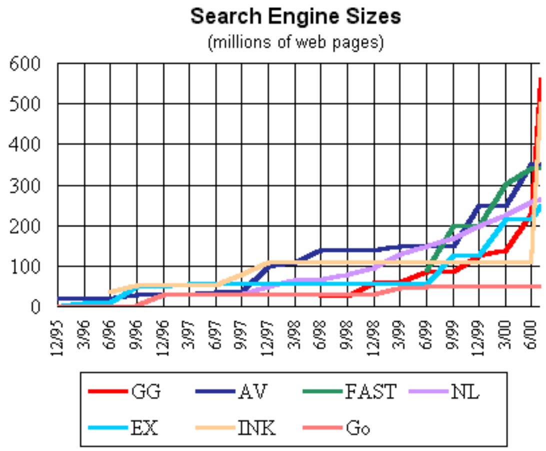
□ Number of pages on the web estimated by www.searchenginewatch.com: over 1000 million



source (19/10/00): http://www.searchenginewatch.com/reports/sizes.html

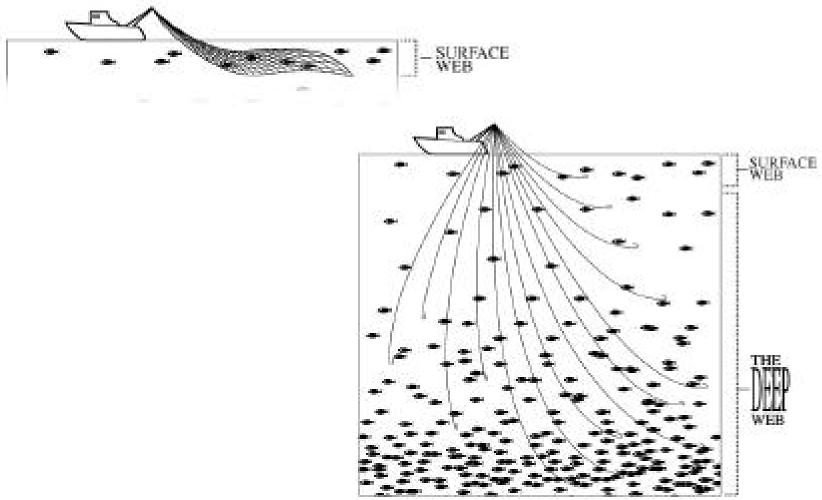
Growth over time (07.07.00)

□ Search Engine Sizes Over Time



source (19/10/00): http://www.searchenginewatch.com/reports/sizes.html

The Deep Web - A different estimate I



source (19/10/00): http://www.completeplanet.com

The Deep Web - A different estimate II

- □ Deep Web sources store their content in searchable databases that only produce results dynamically in response to a direct request.
- □ Public information on the deep Web is currently 400 to 550 times larger than the commonly defined World Wide Web
- □ The deep Web contains 7,500 terabytes of information, compared to 19 terabytes of information in the surface Web
- □ The deep Web contains nearly 550 billion individual documents compared to the 1 billion of the surface Web
- □ More than an estimated 100,000 deep Web sites presently exist
- □ 60 of the largest deep Web sites collectively contain about 750 terabytes of information — sufficient by themselves to exceed the size of the surface Web by 40 times
- □ A full 95% of the deep Web is publicly accessible information — not subject to fees or subscriptions.

source (19/10/00): http://www.completeplanet.com

What do we need for a distributed system to share document?

- □ How are documents encoded?
  - content
  - semantics
  - presentation
- □ How documents are identified?
  - Where is data held?
  - How can data be accessed?
- □ How are the document transmitted/transported to the user?

The Web Approach

- □ Document format
  - Hypertext Markup Language, HTML
    - □ Document Type Definition (DTD)
      Standardized General Markup Language (SGML)
- □ Mechanism for identification
  - Uniform Resource Identifier, URI
    - □ use as Uniform Resource Name, URN
    - □ use as Uniform Resource Locator, URL
- □ Transfer protocol
  - Hypertext Transfer Protocol, HTTP
    - □ ASCII-coded Request-Reply protocol using TCP/IP

Web Terms I

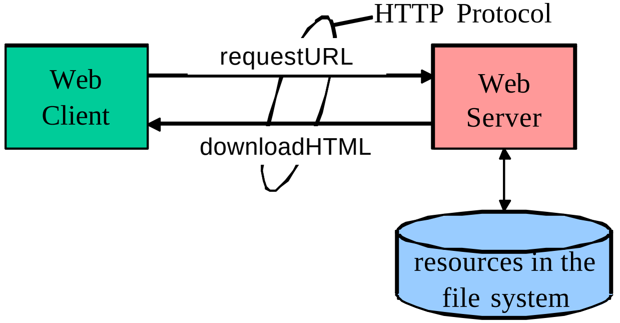
- User
  - Human user of the World Wide Web
- Client
  - Software, sends HTTP-requests to a Web server
- Browser
  - Client that can visualize HTML + ...
- Server
  - Software, that waits for HTTP-requests and answers with HTTP-replies
- Site
  - A collection of web pages (usually belonging to one organization)

Web Terms II

- Page
  - A single HTML page
    - May contain other document types (e.g. Images, sounds, ...)
- Homepage
  - Entry page to interconnected pages that have a common content, e.g.
    - homepage of a company or organization
    - homepage on a specific topic
    - personalhomepage
- Portal
  - specific entry page to the web, e.g.
    - portal of a provider (e.g. AOL, t-online, ...)
    - portal of content/new provider, (e.g.cnn.com, ...)
    - portal of search engines/catalogs (z.B. yahoo, lycos, ...)

Architecture and Protocols

- client-server architecture
- synchronous communication model (request/response)
- resources
  - Unit that is communicated between Client and Server
  - static or dynamic

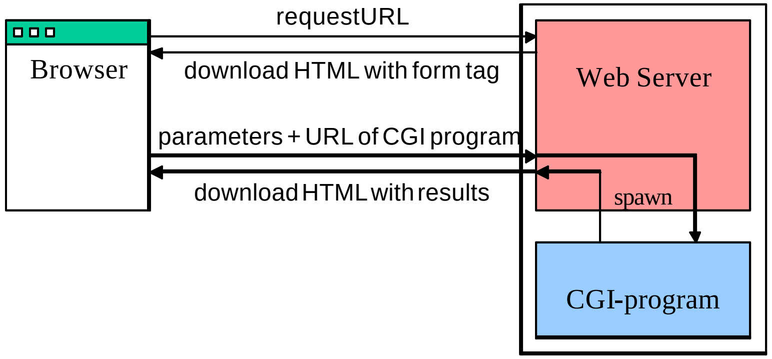


Resources in the World Wide Web I

- Structure of the documents exchanged
  - HTML
  - MIME-types denote embedded non-HTML parts
- Visualization on the screen
  - Client parses HTML und visualizes the content
  - non-HTML is displayed by
    - thebrowser
    - client extension
    - plug-In
    - helperapplication

Resources in the World Wide Web II

- □ connecting external sources
  - Common Gateway Interface, CGI
- □ Forms in HTML to sent input parameters


Browser
requestURL
download HTML with form tag
parameters + URL of CGI program
download HTML with results
Web Server
spawn
CGI-program

Web vs. Internet

- □ In the Internet heterogeneous networks are interlinked
  (ISO layer 1 und 2)
  - Abstraction of networks with IP
    - IP is packet oriented
  - Different transfer protocols
    - TCP, UDP, RTP, ...
  - various application protocols
    - telnet, FTP, NNTP, SMTP, HTTP, ...
- □ Web is one of the services in the Internet

Web vs. Multimedia

- □ Multimedia is when discreet and continuous data is used concurrently
- □ Multimedia documents
  - HTML supports embedding of graphics, animation, video and audio
  - Using extensions other formats are available
    - VRML, RealAudio, RealVideo, Shockwave, ...
- □ Multimedia communication
  - HTTP can be used to transfer any resources
  - other protocols for streaming Audio and Video
- □ The Web can be used to create, transport, and show multimedia documents over different platforms

Further Topics I

- □ The Internet, technical foundation:
  - Client and server, architecture of the Web
  - Technical foundation and protocols (TCP/IP, HTTP)
  - Services (Mail, News, FTP, WWW)
  - Names und resources (DNS), URL, URI
  - Security and proxy server
- □ The Web seen as an information system
  - Information, media, types (MIME)
  - Usage of media types used in the WWW (HTML, GIF, JPG, PNG)
  - Organizing and structuring information, Hypertext
  - Describing information (Markup, SGML, XML, HTML)
  - Access to Information / addressing (search, Navigation, ...)
- □ Architecture I
  - Distributed systems (advantages, requirements, implementation)
  - WWW as a application platform
  - 2-tier, 3-tier, multi-tierarchitectures
  - Middleware, agents, objects, Corba/COM in the WWW

Further Topics II

- □ Architecture II (Server)
  - Functionality of a Web servers, minimized server in C
  - Advanced concepts for servers (multi-threading, Server Side Scripting, CGI)
  - Concepts in standard Web servers (Jigsaw, Apache, IIE)
  - Customized server (streaming video/audio, control, gateways)
- □ Architecture III (Browser)
  - Functionality of a Web browser, minimized browser (example telnet)
  - Advanced concepts for a web browser (e.g. multi-threading)
  - Concepts of standard web browser (Amaya, IE5, Mozilla)
  - Customized browsers and clients (Web radio, Web-TV, robots)
- □ programming I
  - Design of distributed applications
  - Technology survey
  - criteria(server-side, client-side, static, dynamic)
  - Session concept (cookies, applets, ActiveX)
  - Integration of legacy systems

Further Topics III

- □ Programming II (statistic resources)
  - Request-replay procedure using statistic resources
  - HTML, CSS, forms
  - JavaScript/Jscript
  - Java applets
  - ActiveX
  - Helper application und plug-Ins
  - Security for statistic resources
- □ programming III (dynamic resources)
  - Request-replay procedure using dynamic resources
  - Server Side Include (SSI)
  - Common Gateway Interface (CGI)
  - Server Side Scripting (Active Server Pages),
  - Servlets
  - Server extensions (NSAPI, ISAPI)
  - Security for dynamic resources

Further Topics IV

- □ Management of web applications
  - Requirements and concepts, differences to standard software
  - Project planning, building of applications and project management
  - Support and maintenance
  - Analyzing user behavior
  - Provider models, analyzing costs
- □ Lifecycle of WWW applications
  - Comparison with software engineering models
  - Methods
  - Tools
- □ Selected Applications and advanced topics
  - Infrastructure as WWW-applications: search engines and catalog
  - Implementing a online-shop
  - Mobile web access - WAP and WML
  - Control and management using the WWW
  - Commercial systems with a WWW front-end

Refercences Chapter 1

- □ Liu, C., Peek, J., Jones, R., Buus, B., and Nye, A.; 1994. *Managing Internet Information Services*. O´Reilly, Sebastopol.
- □ WebMaster Magazine; 1996. *Overview of the World Wide Web*. http://www.cio.com/WebMaster/sem2_home.html
- □ W3C, World Wide Web Consortium; 1999. *About the World Wide Web*. http://www.w3.org/WWW/
- □ Zakon R.H.; 1999. *Hobbes' Internet Timeline v4.0*. RFC 2235, http://info.internet.isi.edu/in-notes/rfc/files/rfc2235.txt oder http://www.isoc.org/zakon/Internet/History/HIT.html
- □ Michael K. Bergman. *The Deep Web: Surfacing Hidden Value.* White Paper. BrightPlanet.com LLC. http://www.completeplanet.com/Tutorials/DeepWeb/index.asp

Web Engineering

Chapter 2: Foundation - Identifiers and Protocols

Table of Contents

1. Uniform resource identifier

2. Transfer protocols

Uniform Identifiers

□ It must be possible to identify resources
- by Name
- by Address resp. Location

□ any resource in the Internet should be identified
- Web pages, FTP-Resources, Mailboxes, Directories, interactive services

➔ Requirements:
the identification mechanism should be
- extensible,
- complete,
- printable (to be represented as string of 7-bit characters)

Uniform Resource Identifier (URI)

□ Syntax for identifiers [RFC1630]
`<uri> ::= <scheme>":"<scheme-specific-part>`

□ `<scheme>`
name of the scheme

□ `<scheme-specific-part>`
identifier in a format that is according to the `scheme`

□ URIs are:

▪ Names:	Uniform Resource Name (URN)
▪ Locations/Addresses:	Uniform Resource Locator (URL)
▪ Meta information:	Uniform Resource Characteristic (URC)

Reserved characters

- □ For all types of URIs
- □ The percent sign **("%", ASCII 25 hex)**
  - ▪ Escapecharacter
- □ Hierarchical forms **("/", ASCII 2F hex)**
  - ▪ delimitingof substrings whoserelationshipishierarchical
  - ▪ not the same as the / in Unix File system!
- □ Hash - fragment delimiter **("#", ASCII 23 hex)**
  - ▪ Identifies a fragment in a resource
- □ Query Delimiter **("?", ASCII 3F hex)**
  - ▪ to delimit the boundary between the URI of a queryable object

Uniform Resource Name (URN)

- □ Unifying all naming
- □ URNs serve as persistent, location-independent, resource identifiers.
- □ URN [RFC 1737, RFC 2141] (`<scheme> ::= "urn"`)

```
<urn> ::= "urn:" <nid> ":" <nss>
```

- □ `nid` = Namespace Identifier
- □ `nss` = Namespace Specific String

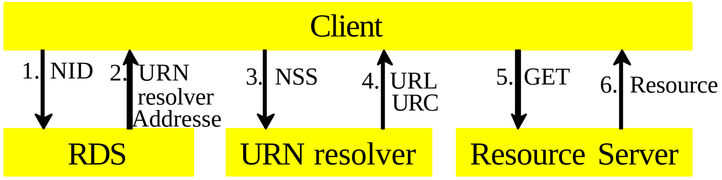
URN Properties

- □ global scope and uniqueness
- □ persistence
- □ scalable
- □ legacy support
- □ extensible
- □ independent
- □ resolvable

URN - Resolution

- □ Infrastructure for URNs is still experimental
  - ▪ Resolver Discovery Service (RDS)
  - ▪ Name service, name resolution (URN resolver)
  - ▪ Result of the resolution is a URL or a URC
  - ▪ See: RFC 1737, 2276

```
<urn> ::= "urn:" <nid> ":" <nss>
```



Uniform Resource Locator (URL)

- □ Unifying all addresses
- □ URL `scheme` definition [RFC1738]
  - ▪ **`http`** - **`https`** - **`ftp`** - **`news`**
  - ▪ **`nntp`** - **`mailto`** - **`telnet`** - **`ldap`** and more ...
- □ followed by `scheme-specific-part`
  Definition in a general format

  ```
  ["//"] [user [":"password] "@"] host [":"port] ["/"url-path]
  ```
- □ Definitions are maintained by the
  Internet Assigned Numbers Authority (IANA)
- □ URLs can also be relative [RFC 1808]

HTTP URL

HTTP URL **`scheme-specific-part`**

```
<http_URL> =
   "http://"[user[":"password]"@"]<host>[":"<port>][<abs_path>]

<host>     = <A legal Internet host domain name or IP
              address (in dotted-decimal form),
              as defined by Section 2.1 of RFC 1123>

<port>     = *DIGIT

<abs_path> =
   "/"[<path>][";"<params>]["?"<query>]["#"<fragment>]

<path>     = <fsegment> *( "/" <segment> )
```

Vergleich URN vs. URL

Property	*URN*	*URL*	
scope	global	abs. URL: global	rel. URL: local
global unique	yes	abs. URL: yes	rel. URL: no
persistent	yes	no	
scalable	yes	yes	
Legacy support	yes	limited	
resolution	still open	partly using DNS	

Table of Contents

1. Uniform resource identifier

2. Transfer protocols

Transfer protocol

- □ The identified resources must be transferred
- □ Client-Server Architecture based on
  - Request-Reply Protocol
  - Transactions
- □ Design goals
  - Simple (easy to implement)
  - Lightweight (little processing power required)
  - fast
- □ Hypertext Transfer Protocol, HTTP
  - based on TCP/IP
  - Request are idempotent,
  - Basic communication is state less
  - ASCII coded

Transmission Control Protocol (TCP) I

Motivation (September <u>1981</u>), RFC793

„Computer communication systems are playing an increasingly important role in military, government, and civilian environments. This document focuses its attention primarily on military computer communication requirements, especially robustness in the presence of communication unreliability and availability in the presence of congestion,…

As strategic and tactical computer communication networks are developed and deployed, it is essential to provide means of interconnecting them and to provide standard interprocess communication protocols which can support a broad range of applications. ….

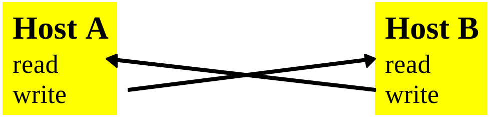
TCP is a connection-oriented, end-to-end reliable protocol designed to fit into a layered hierarchy of protocols which support multi-network applications. The TCP provides for reliable inter-process communication between pairs of processes in host computers attached to distinct but interconnected computer communication networks. Very few assumptions are made as to the reliability of the communication protocols below the TCP layer. TCP assumes it can obtain a simple, potentially unreliable datagram service from the lower level protocols.“

Transmission Control Protocol (TCP) II

- □ Design goals
  - Reliable connections over an unreliable network
- □ properties
  - Connection oriented
  - a stream of data sent on a TCP connection is delivered reliably and in order at the destination.
  - flow control mechanisms
  - based on IP
- □ Identification of sender and receiver
  - IP-Address
  - Port (16 Bit)
- □ Further information
  - TCP: RFC 793, RFC 1122, RFC 1323
  - Well Known Ports: RFC 1700

Socket Connection (TCP)

- □ bi-directional connection
- □ read and write like in a file, byte stream
- □ identified by IP -address and port number
- □ interface for the application programmer, provided as functions in a library, as control, or as classes in a lot of programming languages
  - C: socket(...), <sys/socket.h>
  - Java: class Socket in java.net.*
  - VisualBasic: Winsock Control



Communication with a Web Server

- □ HTTP is based on TCP –
  to experiment with the protocol telnet can be used.

```
> telnet 129.13.170.1 80[RETURN]
  GET /index.html HTTP/1.0[RETURN]
  [RETURN]
```

```
> telnet www.teco.edu 80[RETURN]
  HEAD /index.html HTTP/1.0[RETURN]
  [RETURN]
```