



Netzbasierte Informationssysteme **Die Architektur des Web I**

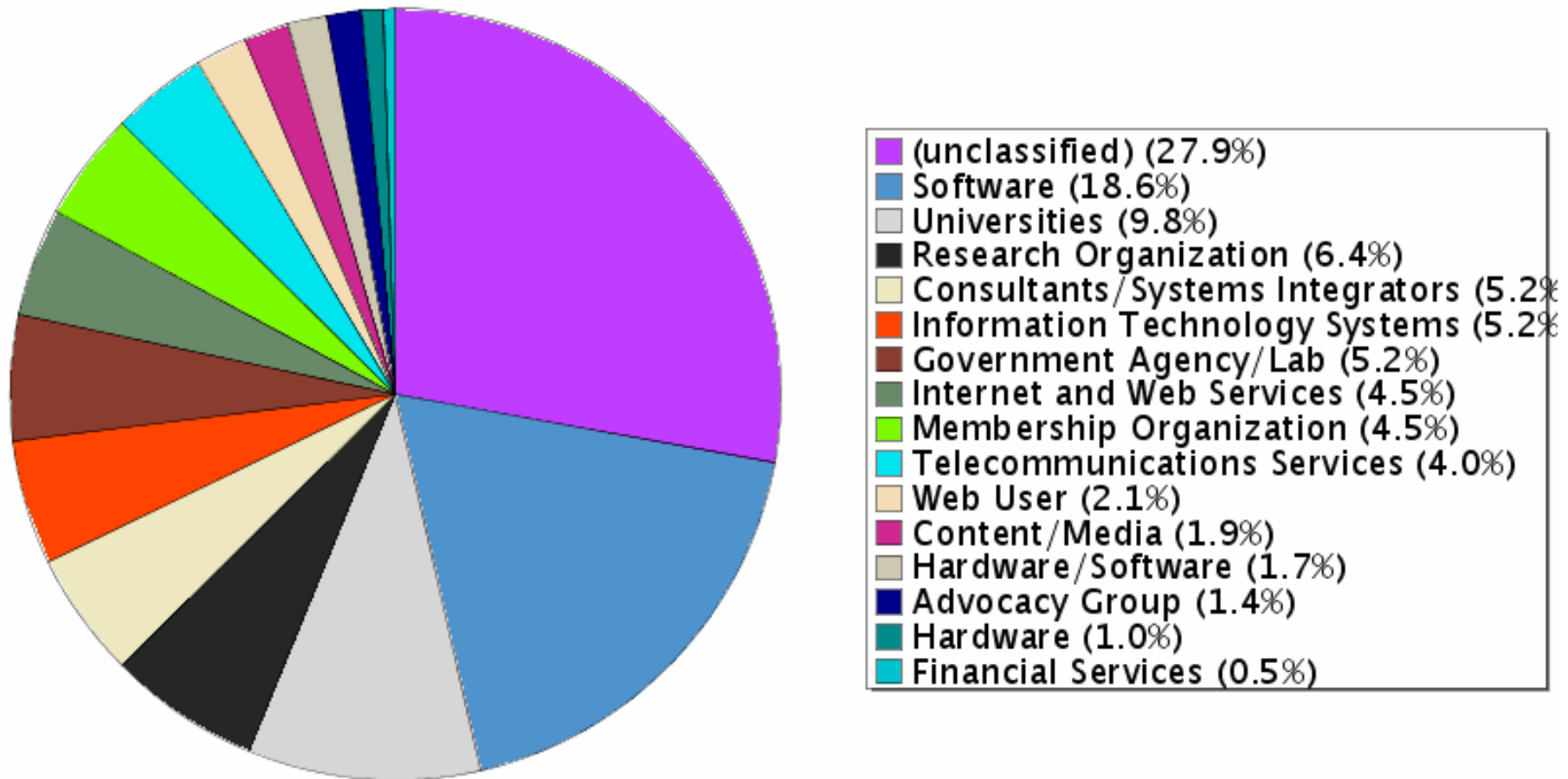
Prof. Dr.-Ing. Robert Tolksdorf
Freie Universität Berlin
Institut für Informatik
Netzbasierte Informationssysteme
mailto: tolk@inf.fu-berlin.de
<http://www.robert-tolksdorf.de>

- Beginnend um 1992 entstanden
- Zunächst viele ad-hoc Entscheidungen
- Erst ab 2000 liegen ernstzunehmende systematische Arbeiten zum Design der Web-Architektur vor:
 - Roy Fielding. Architectural Styles and the Design of Network-based Software Architectures. PhD thesis, University of California, Irvine. 2000. <http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm>
 - Ian Jacobs, Norman Walsh (Eds.) Architecture of the World Wide Web, Volume One. W3C Recommendation 15 December 2004. <http://www.w3.org/TR/webarch>
- Das World Wide Web Konsortium nimmt diese als Leitlinie für die weitere Standardisierung

Die Standardisierung des Web

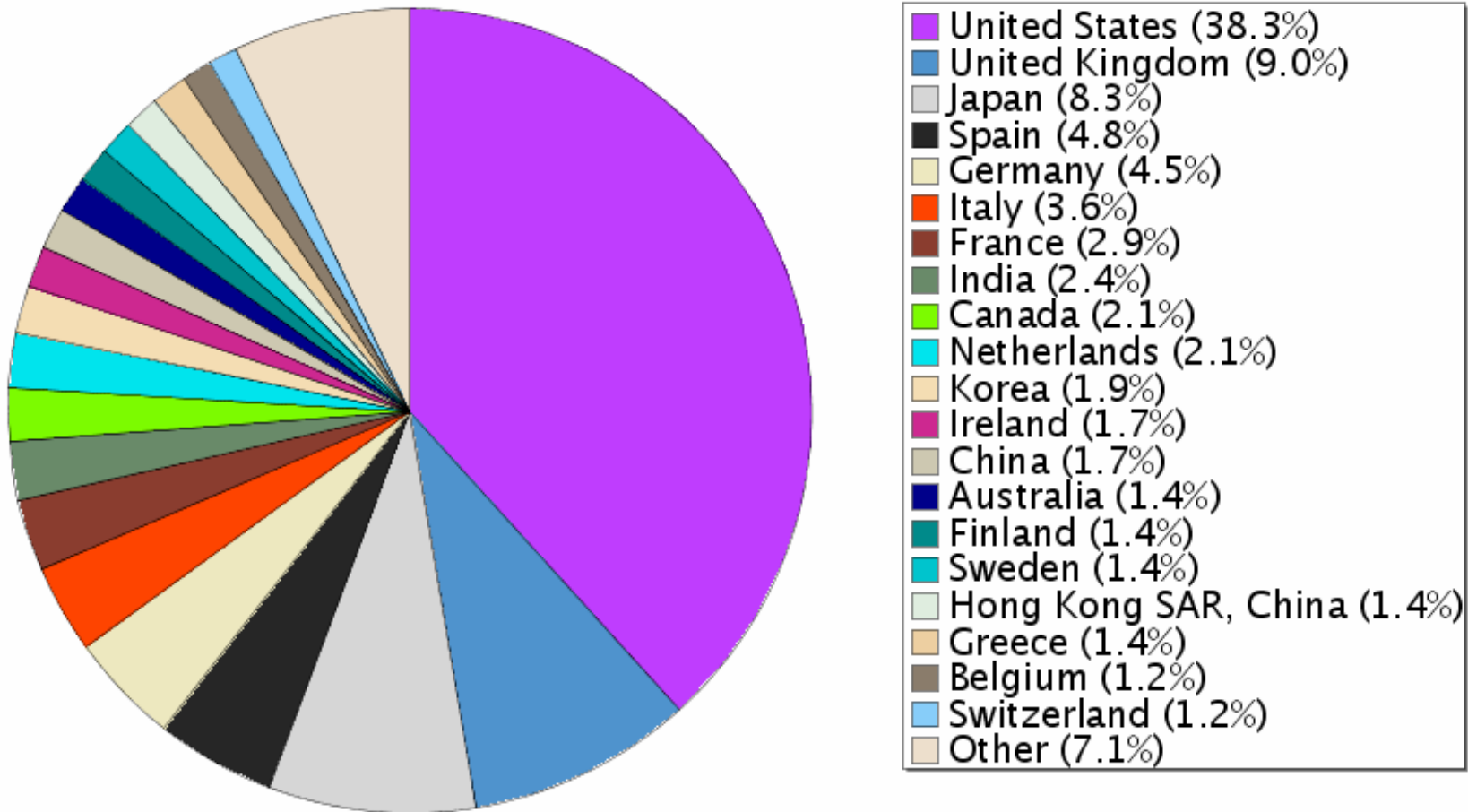
- Das World Wide Web Consortium W3C treibt die Entwicklung des Web voran:
 - „W3C's mission is: To lead the World Wide Web to its full potential by developing protocols and guidelines that ensure long-term growth for the Web“
[<http://www.w3.org/Consortium/>]
- Prinzipien
 - Konsens
 - Offenheit
 - Lizenzfreiheit
- Ca. 400 Mitglieder

Die Standardisierung des Web



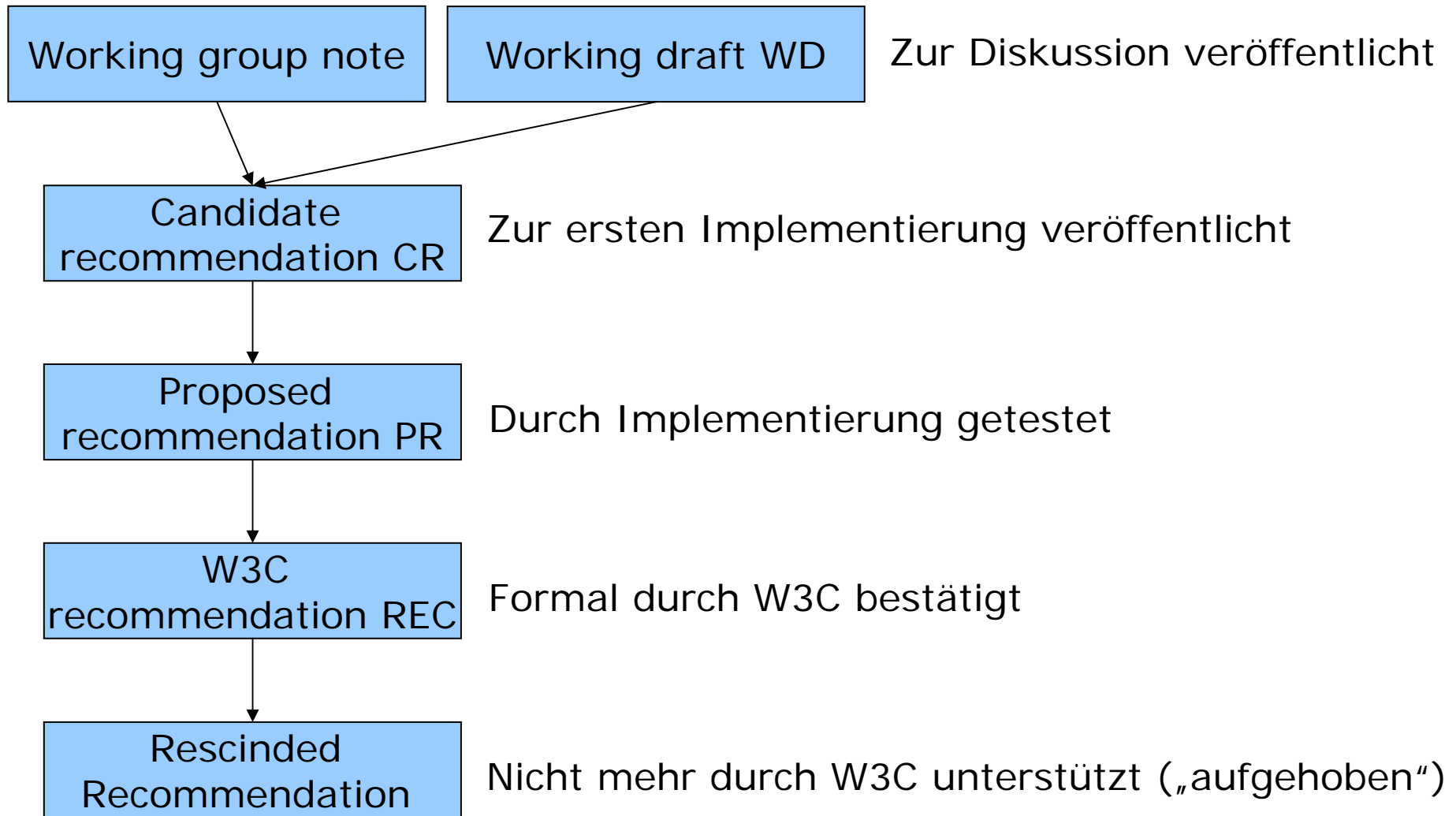
[<http://www.w3.org/Consortium/org>]

Die Standardisierung des Web

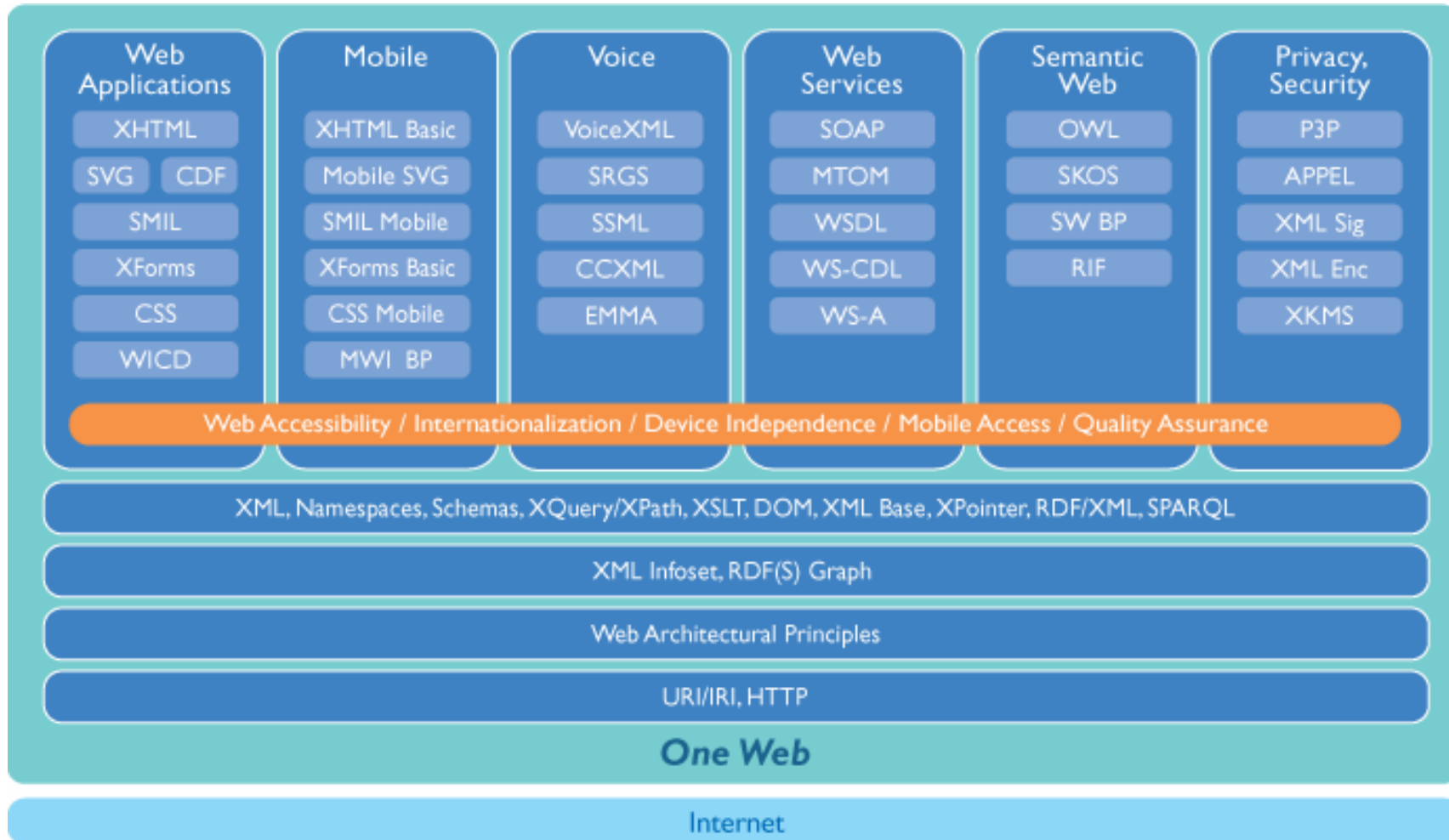


[<http://www.w3.org/Consortium/org>]

Prozess der Standardentwicklung



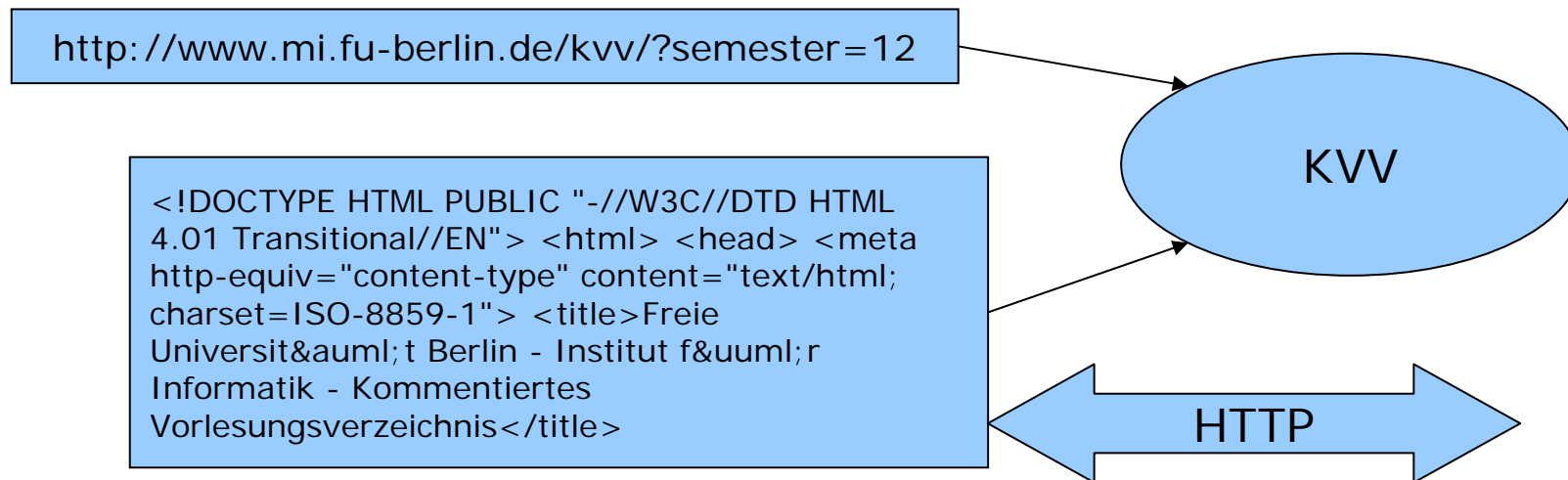
Der W3C Technologie-Stack



[<http://www.w3.org/Consortium/technology>]

- Aber: Welche Prinzipien stecken dahinter?

- Drei Grundlagen der Web Architektur:
 - *Identifikation* durch Uniform Resource Identifiers URI
 - *Interaktion* durch Protokolle wie HTTP
 - *Datenformate* wie HTML oder XML
- (Informations-)Ressourcen werden identifiziert und durch Interaktion sind sie durch Übermittlung einer Nachricht in einer Repräsentation zugänglich





Identifikation URIs

- Uniform Resource Identifier URI: „A Uniform Resource Identifier (URI) is a compact string of characters for identifying an abstract or physical resource“ [RFC 2396]
- Request For Comments-Dokumente (RFC) definieren alle technischen Aspekte des Internet
 - RFC 1738 :
T. Berners-Lee, L. Masinter, und M. McCahill. Uniform Resource Locators (URL). RFC 1738, Internet Engineering Task Force, December 1994.
- Internet Engineering Taskforce IETF erstellt RFCs <http://www.ietf.org/rfc.html>
- Standardisierungsprozess ist als RFC standardisiert: The Tao of IETF: A Novice's Guide to the Internet Engineering Task Force, RFC 3160, August 2001

- Uniform Resource Identifier URI: „A Uniform Resource Identifier (URI) is a compact string of characters for identifying an abstract or physical resource“ [RFC 2396]
- Lediglich Syntax: *schema:schemaspezifischer_Teil*
- URI-Schema typisiert URIs:
 - <ftp://ftp.is.co.za/rfc/rfc1808.txt>
 - <gopher://spinaltap.micro.umn.edu/00/Weather/Los%20Angeles>
 - <http://www.math.uio.no/faq/compression-faq/part1.html#sec1>
 - <mailto:mduerst@ifi.unizh.ch>
 - <news:comp.infosystems.www.servers.unix>
 - <telnet://melvyl.ucop.edu/>
 - <urn:isbn:n-nn-nnnnnn-n>
 - <fon:+49-838-0>

- Uniform Resource Locator URL: „[...]a compact string representation for a resource available via the Internet.“ [RFC 1738]
 - Ist ein URI, dessen Schema auf die Zugreifbarkeit der Ressource im Netz hinweist
 - z.B. `ftp://ftp.is.co.za/rfc/rfc1808.txt`
- Uniform Resource Name URN: „[...] intended to serve as persistent, location-independent, resource identifiers and are designed to make it easy to map other namespaces“ [RFC 2141]
 - Ist eher URI, der Eigenschaft der Ressource beschreibt
 - `urn:isbn:n-nn-nnnnnn-n`
 - URN-Namensraum strukturiert URNs (isbn,...)

Registrierte URI Schemas

[<http://www.iana.org/assignments/uri-schemes.html>]

URI Scheme	Description	Reference
acap	application configuration access protocol	[RFC2244]
cid	content identifier	[RFC2392]
crid	TV-Anytime Content Reference Identifier	[RFC4078]
data	data	[RFC2397]
dav	dav	[RFC2518]
dict	dictionary service protocol	[RFC2229]
dns	Domain Name System	[RFC4501]
fax	fax	[RFC2806]
file	Host-specific file names	[RFC1738]
ftp	File Transfer Protocol	[RFC1738]
go	go	[RFC3368]
gopher	The Gopher Protocol	[RFC4266]
h323	H.323	[RFC3508]
http	Hypertext Transfer Protocol	[RFC2616]
https	Hypertext Transfer Protocol Secure	[RFC2818]
im	Instant Messaging	[RFC3860]
imap	internet message access protocol	[RFC2192]
info	Information Assets with Identifiers in Public Namespaces	[RFC-vandesompe- info-uri-04.txt]
ipp	Internet Printing Protocol	[RFC3510]

Registrierte URI Schemas

[<http://www.iana.org/assignments/uri-schemes.html>]

URI Scheme	Description	Reference
iris.beep	iris.beep	[RFC3983]
ldap	Lightweight Directory Access Protocol	[RFC-ietf-ldapbis-url-09.txt]
mailto	Electronic mail address	[RFC2368]
mid	message identifier	[RFC2392]
modem	modem	[RFC2806]
mtqp	Message Tracking Query Protocol	[RFC3887]
mupdate	Mailbox Update (MUPDATE) Protocol	[RFC3656]
news	USENET news	[RFC1738]
nfs	network file system protocol	[RFC2224]
nntp	USENET news using NNTP access	[RFC1738]
opaquelocktoken	opaquelocktoken	[RFC2518]
pop	Post Office Protocol v3	[RFC2384]
pres	Presence	[RFC3859]
rtsp	real time streaming protocol	[RFC2326]
service	service location	[RFC2609]
sip	session initiation protocol	[RFC3261]
sips	secure session initiation protocol	[RFC3261]
snmp	Simple Network Management Protocol	[RFC4088]

Registrierte URI Schemas

[<http://www.iana.org/assignments/uri-schemes.html>]

URI Scheme	Description	Reference
snmp	Simple Network Management Protocol	[RFC4088]
soap.beep	soap.beep	[RFC3288]
soap.beeps	soap.beeps	[RFC3288]
tag	tag	[RFC4151]
tel	telephone	[RFC2806]
telnet	Reference to interactive sessions	[RFC4248]
tftp	Trivial File Transfer Protocol	[RFC3617]
tip	Transaction Internet Protocol	[RFC2371]
urn	Uniform Resource Names (click for registry)	[RFC2141]
vemmi	versatile multimedia interface	[RFC2122]
xmlrpc.beep	xmlrpc.beep	[RFC3529]
xmlrpc.beeps	xmlrpc.beeps	[RFC3529]
xmpp	Extensible Messaging and Presence Protocol	[RFC4622]
z39.50r	Z39.50 Retrieval	[RFC2056]
z39.50s	Z39.50 Session	[RFC2056]

- `snmp://example.com/bridge1;800002b804616263`
- `tel: +358-555-1234567`
- `modem: +3585551234567;type=v32b?7e1;type=v110`

URN Namensräume

[<http://www.iana.org/assignments/urn-namespaces>]

IETF	[RFC2648]	liberty	[RFC3622]
PIN	[RFC3043]	IPTC	[RFC3937]
ISSN	[RFC3044]	UUID	[RFC4122]
OID	[RFC3061]	UCI	[RFC4179]
NEWSML	[RFC3085]	CLEI	[RFC4152]
OASIS	[RFC3121]	tva	[RFC4195]
XMLORG	[RFC3120]	fdc	[RFC4198]
publicid	[RFC3151]	ISAN	[RFC4246]
ISBN	[RFC3187]	NZL	[RFC4350]
NBN	[RFC3188]	oma	[RFC4358]
WEB3D	[RFC3541]	IVIS	[RFC4617]
MPEG	[RFC3614]	S1000D	[RFC-rushing-s1000d-urn-00.txt]
mace	[RFC3613]		
fipa	[RFC3616]		
swift	[RFC3615]		

- URN: ISBN: 3-5401-4009-3

- Prinzip:
 - Grundlegende Regel der Systemgestaltung
 - "separation of concerns", "generic interface", "self-descriptive syntax," "visible semantics," "network effect" (Metcalfe's Law), and Amdahl's Law: "The speed of a system is limited by its slowest component."
- Beschränkung:
 - Entwurfsentscheidungen, die Verhalten oder Interaktionsmöglichkeiten einschränken um ein erwünschtes Ziel zu sichern
- Gute Praxis:
 - Anwendungsweise (im Rahmen der Beschränkungen), die den Wert des Systems erhöht

- Prinzip: Global Identifiers
 - URIs haben sind global eindeutige Bezeichner
 - Damit können sie auch global verwendet werden
 - Netzwerkeffekt: Je mehr das URI Schema genutzt wird umso wertvoller wird es
- Gute Praxis: Identify with URIs
 - Ressourcen sollte durch URIs zu bezeichnen sein
 - Damit erhöhen Anbieter den Wert des Web

- Beschränkung: URIs Identify a Single Resource
 - Unterschiedliche Ressourcen sollen mit unterschiedlichen URIs bezeichnet werden
- Gute Praxis: Avoiding URI aliases
 - Nicht unterschiedliche URIs für die gleiche Ressource verwenden
- Gute Praxis: Consistent URI usage
 - Erhaltene URIs identisch weiterverwenden

- Gute Praxis: Reuse URI schemes
 - Vorhandene URI Schemas verwenden um Kosten für neue zu sparen
- Gute Praxis: URI opacity
 - Nicht aus URI auf Inhalt schliessen
 - Hinter <http://www.foobar.com/docs> kann ein PDF-Dokument stehen
 - <http://www.foobar.com/index.html> kann ein Film sein



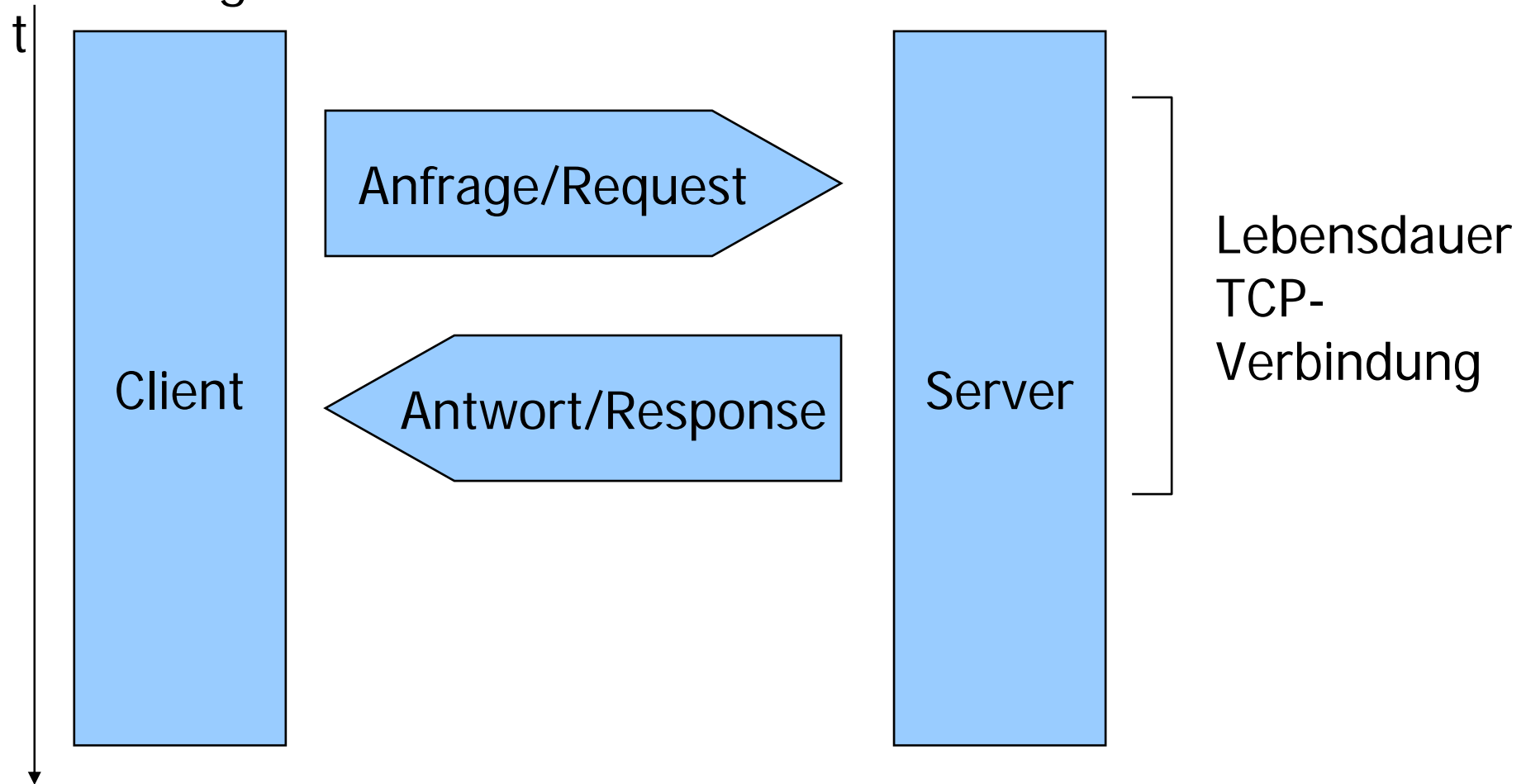
Interaktion HTTP

Hypertext Transfer Protocol

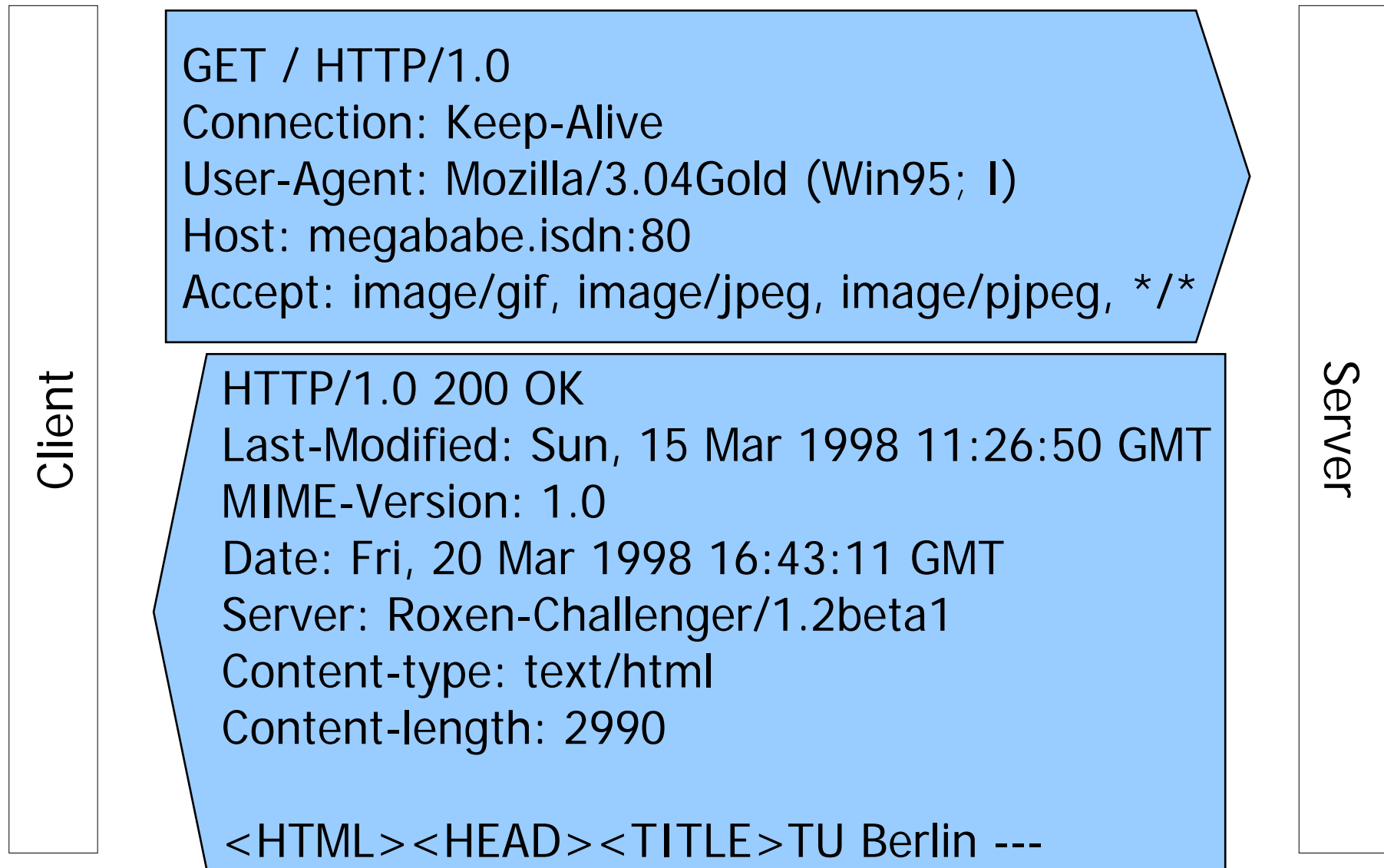
- Aufgabe:
Transfer von Informationen zwischen Web-Servern und Clients
- Port:
80 ist für HTTP reserviert
- Transportprotokoll:
TCP
- Protokoll:
R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach und T. Berners-Lee. *Hypertext Transfer Protocol - HTTP/1.1*. RFC 2616,
<http://www.ietf.org/rfc/rfc2616.txt>

HTTP

- Zustandsloses Protokoll
- Anfrage mit Antwort beantwortet



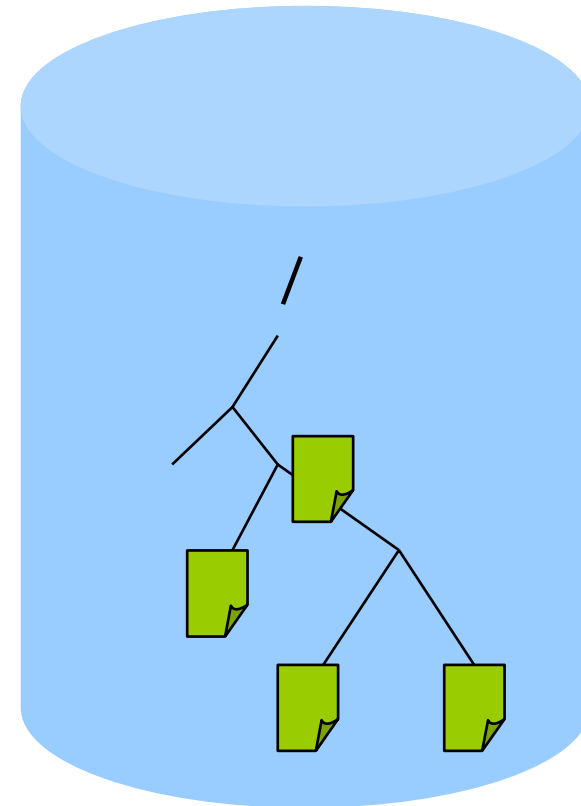
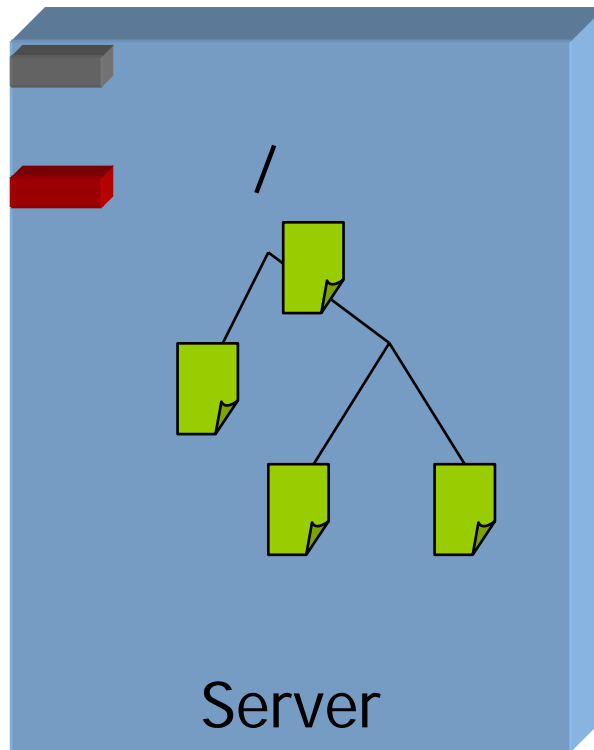
Beispiel: HTTP Protokoll



Aufbau Web-Server

- Web-Server wartet auf Verbindungen
- Beantwortet Nachfragen nach Ressourcen bzgl. des Web-Server Verzeichnisbaums mit Dateien des verwendeten Dateisystembaums

Port 80



Aufbau Anfrage

- Anfrage besteht aus
 - Anfragemethode
 - Anfragebeschreibung durch Kopfzeilen
 - Allgemeine Beschreibungen
 - Anfragespezifische Beschreibungen
 - Beschreibung eventuell beiliegenden Inhalts
 - Leerzeile
 - Eventueller Inhalt
- Beispiel:

```
GET / HTTP/1.0  
Connection: Keep-Alive  
User-Agent: Mozilla/3.04Gold (Win95; I)  
Host: megababe.isdn:80  
Accept: image/gif, image/jpeg, image/pjpeg, */*
```

Anfragen in HTTP

- Format Anfragemethode: *Methode Ressource* HTTP/x.y
- Resource ist
 - Absoluter Pfad im Server-Verzeichnisbaum
 - Voll-qualifizierte URL bei Anfrage an Proxy (s.u.)
 - *, Authority bei bestimmten Methoden
- GET Methode
 - Anforderung einer Informationseinheit vom Server
 - GET /Style/CSS/ HTTP/1.1 an Server www.w3.org
 - Beantwortet mit Antwortcode, Kopfzeilen, Inhalt

- Date: Tue, 15 Nov 1994 08:12:31 GMT
Datum des Abschickens der Anfrage im RFC 1123 Format
- Connection: close
Verbindung nach Ergebnisübermittlung abbauen
- Cache-Control: *Direktive*
Steuert das Caching von Anfragen und Antworten
 - no-cache: Antwort darf nicht zur Beantwortung anderer Anfragen genutzt werden
 - no-store: Antwort- oder Anfragemitteilungen dürfen nicht gespeichert werden
 - weitere: max-age, max-stale, min-fresh, no-transform, only-if-cached, public, private, must-revalidate, proxy-revalidate, s-maxage
- Pragma: no-cache
Entspricht Cache-Control: no-cache
- ...

- Transfer-Encoding: *Encoding*
Wie die Mitteilung für den Transfer kodiert wurde
 - chunked: Mitteilung in Teilen geschickt, Zeichenanzahl in initialer Hexzahl

```
>java HttpClient11 focus.msn.de
java HttpClient11 focus.msn.de
HTTP/1.1 200 OK
Date: Fri, 25 Nov 2005 13:20:01 GMT
Server: Apache
set-cookie: NGUserID=11329248012594; path=/; domain=.msn.de;
  expires=fri, 10-aug-2012 16:48:59 gmt
Transfer-Encoding: chunked
Content-Type: text/html

2e96
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
  <html> <head>
  <title>FOCUS Online in Kooperation mit MSN Homepage</title>
  <meta
```
 - identity: Mitteilung unkodiert geschickt
 - gzip, compress, deflate: Komprimierte Übertragung

Anfrage Kopfzeilen

- Host: *Name*
Aus der URL ermittelter Name des Rechners von dem angefordert wird. *Einziger Pflichtkopfzeile in HTTP 1.1*
- If-Modified-Since: *Datum*
Änderung der Informationseinheit seit *Datum*
 - Ja: 200 und Inhalt schicken
 - Nein: 304 und Inhalt nicht schicken
- If-Unmodified-Since: *Datum*
Änderung der Informationseinheit seit Datum
 - Ja: 412 und nicht verarbeiten
 - Nein: Normal verarbeiten (als sei If-Unmodified-Since: nicht vorhanden)

Inhalts-Kopfzeilen

- Content-Encoding: *Kodierung*
Kodierung des Inhalts
 - binary, 8bit, 7bit, quoted-printable, base64, ...
- Content-Transfer-Encoding: *Kodierung*
Transferkodierung
 - gzip, compress, deflate, chunked, identity
- Content-Type: *Medienart*
Medientyp des Inhalts
 - text/html, image/gif, ..
- Content-Language: *Sprachkürzel*
Sprache des Inhalts
 - de, en, en-US
- Content-Length: *Länge*
Länge des Inhalts in Byte
- Content-Range: *Range*
Beschreibung des Ausschnitts bei Teilanforderung

Inhalts-Kopfzeilen

- Content-Location: *URI*
Verweis auf eigentlichen Inhalt
- Content-MD5: *MD5Checksum*
Message Digest für Inhalt zur Integritätsprüfung
- Expires: *Datum*
Kann nach Datum aus Caches gelöscht werden
- Last-Modified: *Datum*
Letzte Änderung

Inhaltstypen

- Per HTTP können beliebige Inhalte transportiert werden, nicht nur HTML
- Multipurpose Internet Mail Extensions MIME (RFC 2045, RFC 2046) definiert ein Schema zur eindeutigen Benennung durch einen Inhaltstypen
- In HTTP in Kopfzeile Content-Type
- Format: *Typ/Untertyp*
 - text/html
 - image/jpeg
 - vnd.motorola.video

```
HTTP/1.0 200 OK
Last-Modified: Sun, 15 Mar 1998 11:26:50 GMT
MIME-Version: 1.0
Date: Fri, 20 Mar 1998 16:43:11 GMT
Server: Roxen-Challenger/1.2beta1
Content-type: text/html
Content-length: 2990
```

```
<HTML><HEAD><TITLE>TU Berlin ---
```

MIME Typen

- Acht Typen:
 - text: Text
 - text/plain, text/html, text/rtf, text/vnd.latex-z
 - image: Grafiken
 - image/png, vnd.microsoft.icon
 - video: Bewegtbilder
 - video/mpeg, video/quicktime, video/vnd.vivo
 - audio: Audiodaten
 - audio/G726-16 , audio/vnd.nokia.mobile-xmf
 - application: binäre und/oder anwendungsspezifische Daten
 - application/EDIFACT, application/vnd.ms-powerpoint
 - multipart: mehrteilige Daten
 - multipart/mixed
 - message: Nachrichten
 - message/rfc822
 - model: Daten
 - model/vrml

MIME Typen

- MIME-Typen werden von der Internet Corporation for Assigned Names and Numbers IANA verwaltet
- <http://www.iana.org/assignments/media-types/>
- Verarbeiten eines bestimmten Medientyps nach Erhalt:
 - Teil der Anwendung (siehe auch: `javax.mail.internet.MimeMessage`)
 - eventuell Unterstützung durch Betriebssystem
- Ermittlung des MIME-Typs für eine Datei:
 - Ableitung aus Endung (`javax.activation.MimetypesFileTypeMap`)
 - Ableitung aus Inhalt der Datei

- Gute Praxis: Reuse representation formats
 - Protokolle sollten MIME als Typisierung ausgetauschter Datenströme verwenden
- Gute Praxis: Available representation
 - Unter einer URI sollten sich Repräsentationen der bezeichneten Ressource abrufbar sein