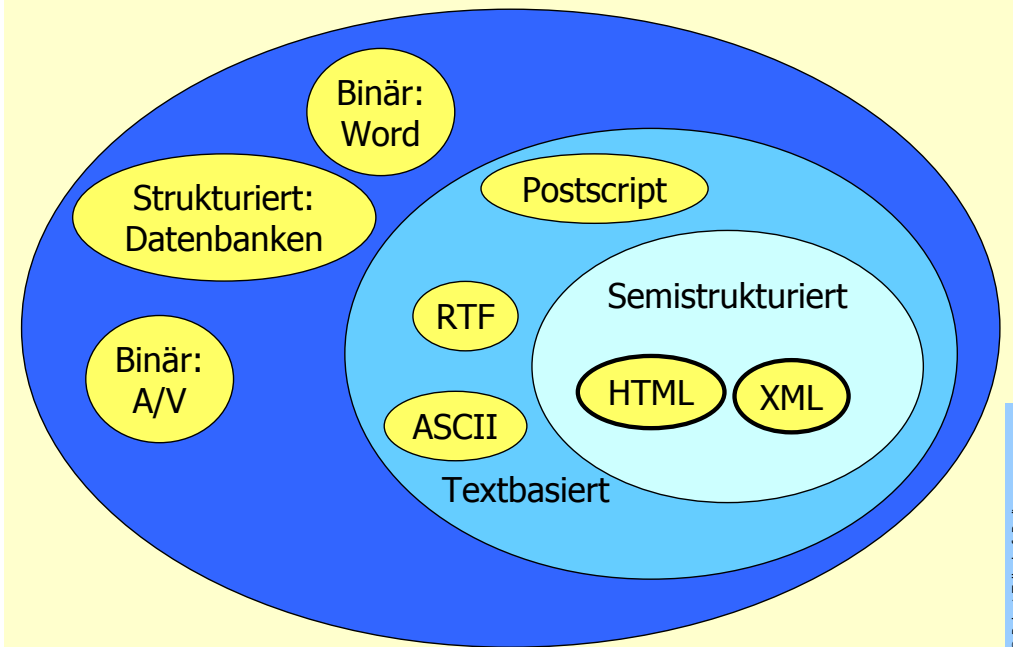


Vorlesung Netzbasierte Informationssysteme (WS 2004/05) Zusammenfassung

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

[1] © Robert Tolksdorf, Berlin

Daten im Netz



[2] © Robert Tolksdorf, Berlin

HTML

[3] © Robert Tolksdorf, Berlin

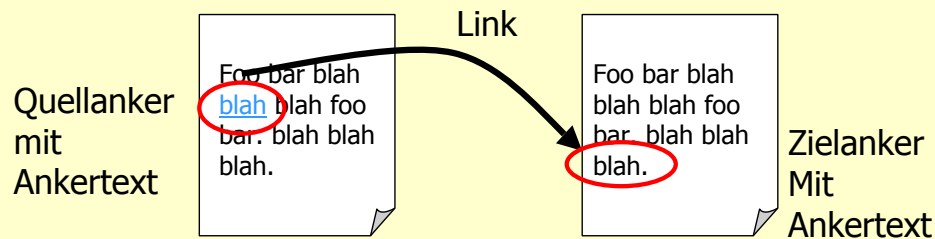
Hypertext Markup Language

- Dominierende Sprache zur Auszeichnung von Dokumenten im Internet
- Definiert vom World Wide Web Consortium, W3C:
 - MIT
 - ERCIM
 - Keio University
- Jedes Informationssystem im Netz muss:
 - HTML Informationen integrieren können
 - HTML Ausgaben erzeugen
 - Mit HTML-Mitteln mit Nutzern interagieren

[4] © Robert Tolksdorf, Berlin

Hypertext Markup Language

- Konzepte:
 - Informationen werden als Dokumente aufgefasst
 - Dokumenteninhalte werden als Klartext dargestellt
 - Dokumententeile werden durch Tags ausgezeichnet
 - Inhaltlich (<h1>Einleitung</h1>, wichtig)
 - Gestalterisch (wichtig)
 - Dokumente werden durch Links zu einem Hypertext verbunden (dem Web)



[5] © Robert Tolksdorf, Berlin

HTML

- Sprache umfaßt
 - Elemente (wie <h1>)
<h1>Neue Vorlesungen</h1>

<hr>
 - Attribute (wie bei <hr height="3">)
 - Entitäten (wie & ;)
 - Grammatikalische Regeln über Elemente (<html> ist Startsymbol, darin die Elemente <head> und <body>)

[6] © Robert Tolksdorf, Berlin

Dokumentenadressen - URLs

- Uniform Resource Locator definiert eine Syntax für eindeutige Bezeichner im Internet
- Internet Dienste sind (zumeist) definiert durch
 - Aufgabe
 - Portnummer auf dem der Dienst angeboten wird
 - Transportprotokoll (TCP oder/und UDP)
 - Protokoll
- Z.B.: Web Dienst
 - Übertragen von HTML Seiten
 - Port 80
 - TCP
 - HTTP
- Z.B.: Usenet Dienst
 - Übertragen von News
 - Port 119
 - TCP
 - NNTP

[7] © Robert Tolksdorf, Berlin

URL

- Uniform Resource Locators sind syntaktische Vereinheitlichung von Dienstbezeichnungen:
http://grunge.cs.tu-berlin.de:8000/
ftp://ftp.cs.tu-berlin.de/pub/net/www
mailto:tolk@cs.tu-berlin.de
 - Form:
http://grunge.cs.tu-berlin.de:8000/resource/data.html#top
-
- Bedeutung ist von Protokoll abhängig, URL ist nur als Syntax definiert

[8] © Robert Tolksdorf, Berlin

XML: Sprache zur Definition von Auszeichnungssprachen

Auszeichnungssprachen

- Kann es eine universelle Auszeichnungssprache geben?
 - Alle visuellen und sonstigen Möglichkeiten aller Ausgabegeräte müßten durch Tags steuerbar sein
 - Alle semantischen Konzepte aller Domänen müßten durch Tags repräsentierbar sein
 - Alle notwendigen Granularitäten der Auszeichnung müßten unterstützt werden:
 - `<ADRESSE>...</ADRESSE>`
 - `<ADRESSE><STRASSE>...</STRASSE><ORT>...</ORT></ADRESSE>`
 - `<ADRESSE><STRASSE>...</STRASSE><ORT><PLZ>...</PLZ><ORTSNAME>...</ORTSNAME></ORT></ADRESSE>`
- Nein: Anwendungsspezifische Auszeichnung nötig

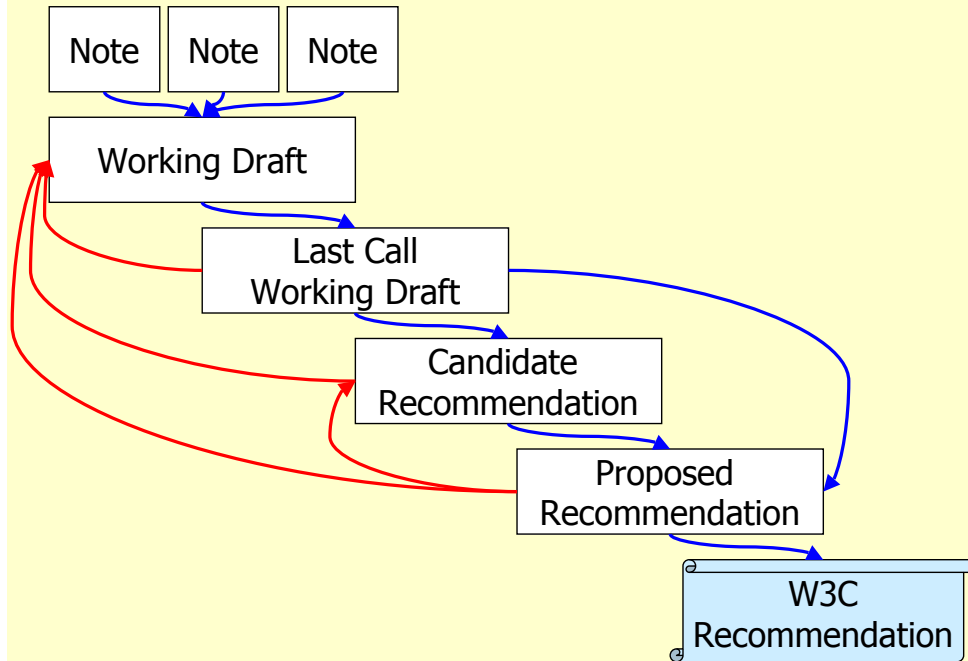
XML Q&A

- *Was ist XML?*
Die Extensible Markup Language ist die Definition einer Untermenge von SGML, mit der man einfach Auszeichnungssprachen definieren kann
- *Woher kommt XML?*
XML ist ein Standard des World Wide Web Konsortiums W3C
- *Was macht man mit XML?*
Anwendungsspezifische Auszeichnungssprachen definieren und standardisieren
- *Was ist der Vorteil von XML-basierten Auszeichnungssprachen?*
Standardisierung ermöglicht Datenaustausch

Document Type Definition

- Eine XML-basierte Sprache wird durch eine XML-DTD (*Document Type Definition*) definiert
- Eine DTD enthält
 - Definitionen gültiger Elemente (Tags) der Sprache
 - Definitionen von Attributen der Elemente und deren Typen
 - Definitionen von Kürzeln (Entitäten)
 - Grammatikregeln

Lifecycle The W3C Recommendation Track



[17] © Robert Tolksdorf, Berlin

Internet

[18] © Robert Tolksdorf, Berlin

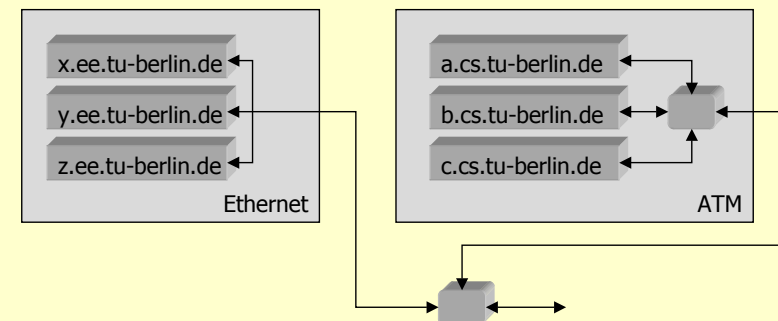
Was ist das Internet

- Eine weltweiter *Verbund von Rechnern*, die über Netze Daten austauschen können.
 - Hardware-bezogene Sicht
 - Zusammenschalten von lokalen Netzen zum Internet
 - Dabei notwendige Verarbeitung von Datenpaketen
- Eine *Protokollfamilie*
 - Netzbezogene Sicht
 - Protokollspezifikationen
- Ein *offenes System*, in dem Dienste genutzt und angeboten werden können.
 - Nutzungs- und anwendungsbezogen
 - Beschreibt die Anwendungsmöglichkeiten des Internet

[19] © Robert Tolksdorf, Berlin

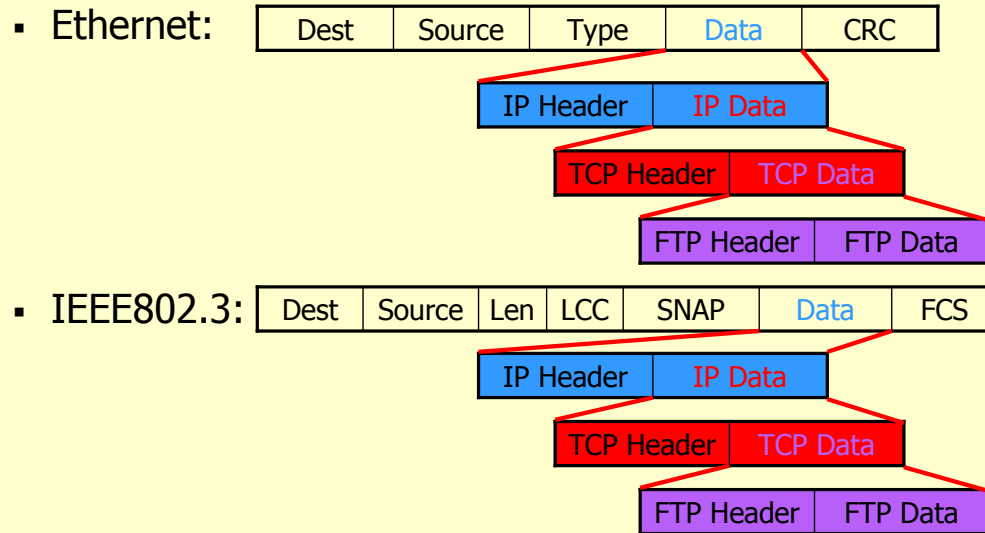
Internet als vernetzter Rechnerverbund

- Das Internet Protokoll IP ermöglicht Internetworking durch Etablierung eines Datenformats und Transportprotokollen, die auf unterschiedlichen Datenverbindungen aufgesetzt werden können



[20] © Robert Tolksdorf, Berlin

Enveloping / Encapsulating



- Fragmentation / Reassembly von IP Paketen

[21] © Robert Tolksdorf, Berlin

IP Adressen

- Aktuell 32 Bit: eg. 130.149.27.12
- Abbildung je nach Medium auf die MAC (Media Access Control), die physikalische Netzadresse
 - ARP, RARP
- Netzwerkmaske definiert, was im lokalen Netz ist, und was nach außen geht
- Netzmaske 255.255.0.0 ->
 - 130.149.0.0 bis 130.149.255.255 sind lokales Netz
 - alles andere muß über einen Router laufen
- Routing: Weiterleiten von Paketen in andere Netzwerke

[22] © Robert Tolksdorf, Berlin

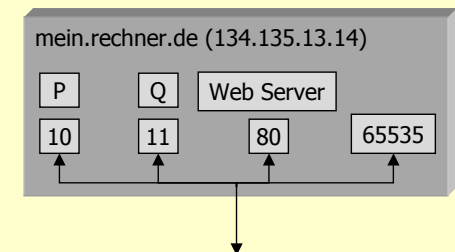
IP Namen

- Internetadresse (IP Adresse) bezeichnet einen Rechner eindeutig
 - als Nummer
130.149.27.12
 - als Name
grunge.cs.tu-berlin.de
- Dienste wie DNS (=Domain Name Service) bilden Namen und Nummern aufeinander ab

[23] © Robert Tolksdorf, Berlin

Transport Protokolle

- Drei Protokolle zum Datentransport
 - UDP: Ein Paket (Datagramm) von Rechner A nach Rechner B schaffen
 - TCP: Pakete werden *geordnet* und *zuverlässig* über eine Verbindung transportiert
- Ports als Kommunikationsadresse
 - Ein *Port* ist ein logischer Netzanschluß, benannt von 0 bis 65535
- Socket ist Endpunkt einer Verbindung



[24] © Robert Tolksdorf, Berlin

Sockets

- Sockets sind die Kommunikationsendpunkte einer Internet-Verbindung
- Die Server Seite:
 - Ein Prozeß „lauscht“ auf einem Rechner an einem Port auf Verbindungswünsche („listen“)
 - Bei einem Verbindungswunsch erzeugt er einen Kommunikationssocket („accept“)
 - Der Kommunikationssocket hat eine andere Nummer als der Verbindungswunschsocket!
- Die Client Seite:
 - Melden des Verbindungswunsches („connect“)
 - „Einstecken“ in den Kommunikationssocket
- Protokollkommunikation:
Zeichen über Kommunikationssocket schicken

[25] © Robert Tolksdorf, Berlin

Internet als dienstorientiertes offenes System

- Internet Dienste sind (zumeist) definiert durch
 - Aufgabe
 - Portnummer auf dem der Dienst angeboten wird
 - Transportprotokoll (TCP oder/und UDP)
 - Protokoll
- Z.B.: Web Dienst
 - Übertragen von HTML Seiten
 - Port 80
 - TCP
 - HTTP
- Z.B.: Usenet Dienst
 - Übertragen von News
 - Port 119
 - TCP
 - NNTP

[26] © Robert Tolksdorf, Berlin

Beispiel: HTTP Protokoll



[27] © Robert Tolksdorf, Berlin

HTTP (Überblick)

[28] © Robert Tolksdorf, Berlin

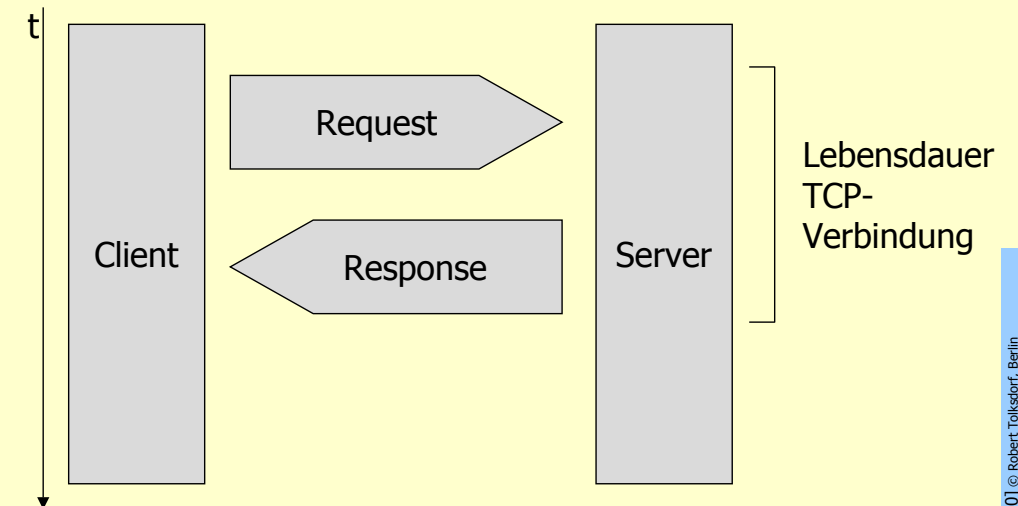
Hypertext Transfer Protocol

- Aufgabe:
Transfer von Informationen zwischen Web-Servern und Clients
- Port:
80 ist für HTTP reserviert
- Transportprotokoll:
TCP (leider)
- 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.w3.org/Protocols/rfc2616/rfc2616.txt>

[29] © Robert Tolksdorf, Berlin

HTTP

- Zustandsloses Protokoll
- Request mit Response beantwortet



[30] © Robert Tolksdorf, Berlin

Aufbau Request

- Request besteht aus
 - Request
 - Request-Beschreibung durch Header
 - Allgemeine Beschreibungen
 - Request-spezifische Beschreibungen
 - Beschreibung eventuell beiliegenden Inhalts
- 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, */*
```

[31] © Robert Tolksdorf, Berlin

Aufbau Response

- Response besteht aus
 - Antwort-Code
 - Response-Beschreibung durch Header
 - Allgemeine Beschreibungen
 - Response-spezifische Beschreibungen
 - Beschreibung eventuell beiliegenden Inhalts
- Beispiel:

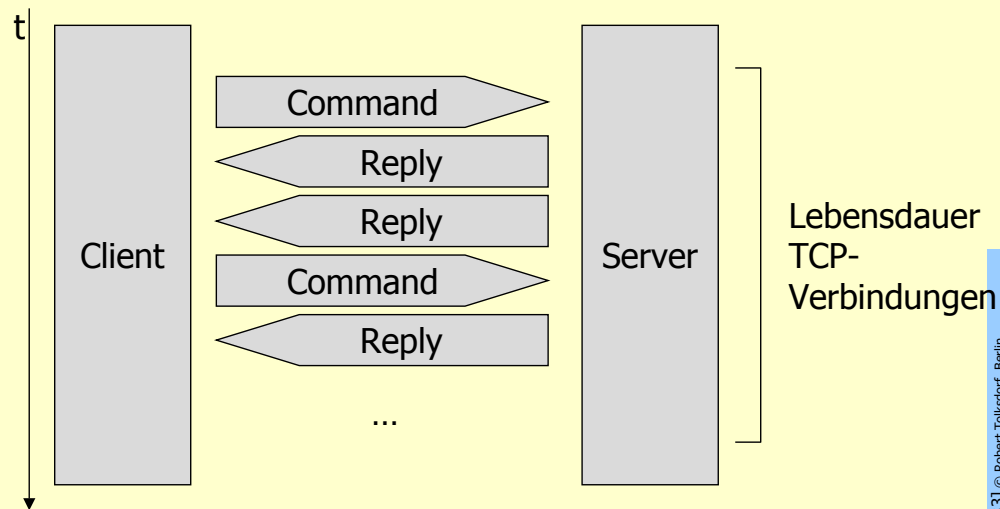
```
HTTP/1.0 200 OK
Last-Modified: Sun, 15 Mar 1998 11:26:50 GMT
MIME-Version: 1.0
Date: Fri, 20 Mar 1998 16:43:11 GMT
Server: Roxen-Challenger/1.2beta1
Content-type: text/html
Content-length: 2990

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

[32] © Robert Tolksdorf, Berlin

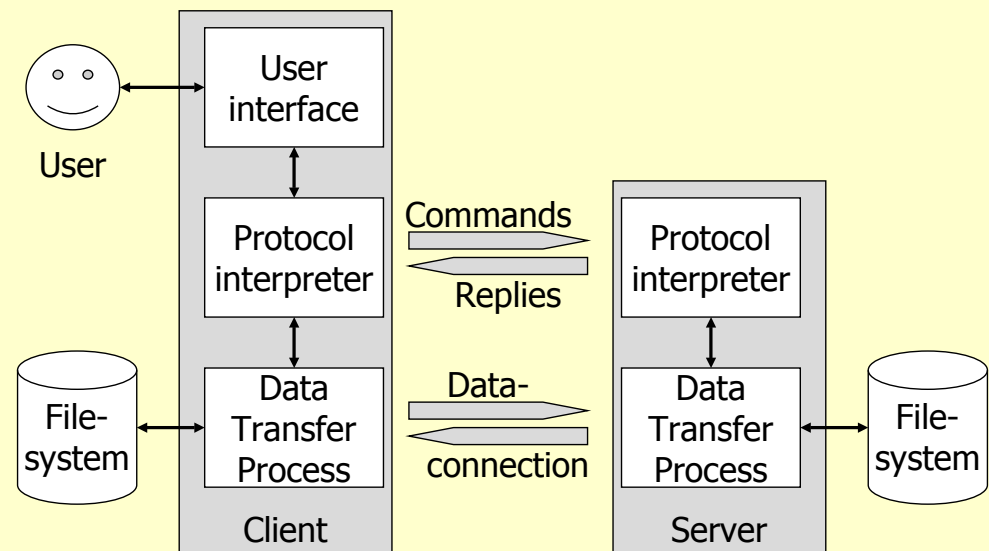
FTP

- Zustandshaltiges Protokoll
- Request mit Response beantwortet



FTP

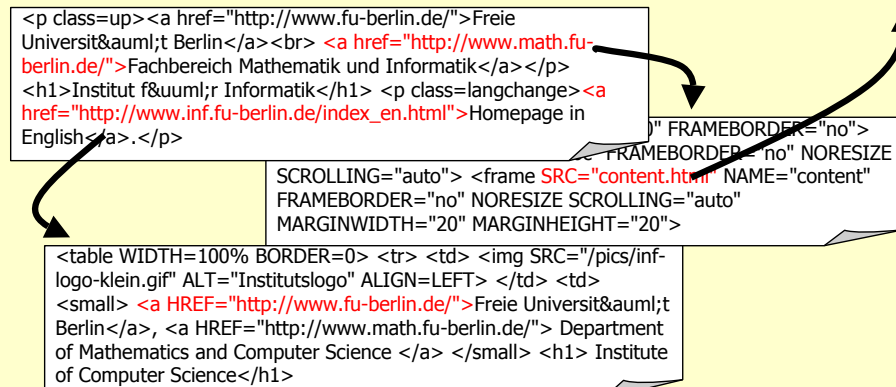
- Modell:



Crawling

Crawling Algorithmus

- Das Web als traversierbarer Graph von Seiten die über Links als Kanten verbunden sind
 - `<a>`, `<link>`, `<meta>`, ``, `<object>`, `<frameset>`
 - FTP-Server, Adressen in nicht-HTML Dokumenten
 - ..

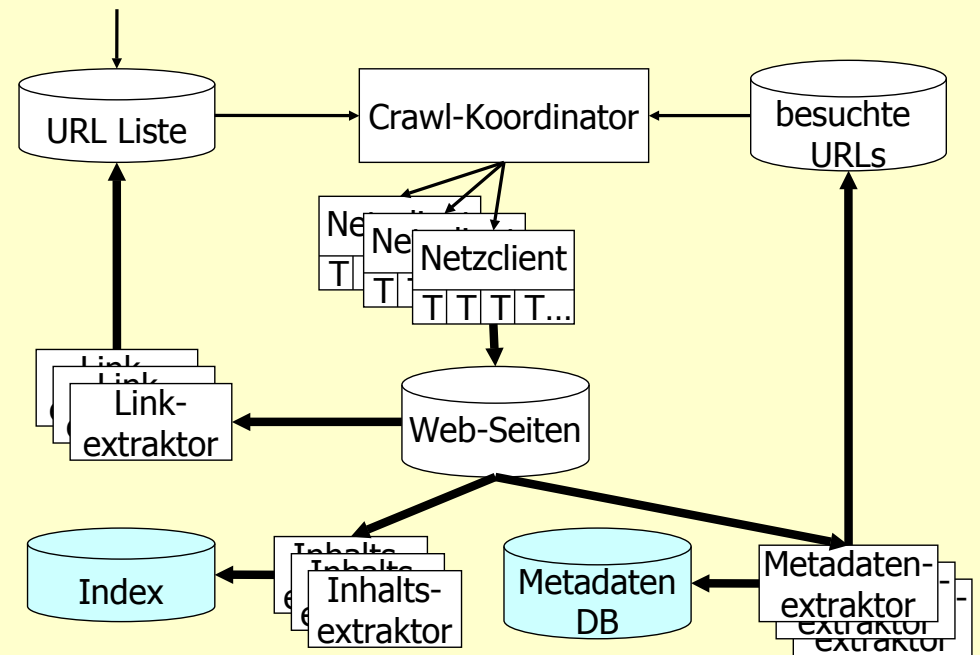


Crawling Algorithmus

1. URL-Liste initial füllen
2. Nehme URL aus Liste und teste
 - schon besucht?
 - passender Medientyp (html/ps/pdf/gif/...)?
 - andere Kriterien (Ort/...)?
3. hole Seite
4. extrahiere URLs und schreibe sie in URL-Liste
5. extrahiere und indexiere Seiteninhalt
6. extrahiere und speichere Metadaten
7. gehe nach 2

[37] © Robert Tolksdorf, Berlin

Einfache Architektur



[38] © Robert Tolksdorf, Berlin

Robots Exclusion Protokoll

- Definiert einen Mechanismus mit dem ein Server festlegt, ob er von einem Crawler besucht werden will
- Daten /robots.txt auf Server
- <http://www.inf.fu-berlin.de/robots.txt>:

```
# robots.txt for http://www.inf.fu-berlin.de/  
User-agent: *  
Disallow: /tec/net/  
Disallow: /tec/rechner/  
Disallow: /tec/software/packages/  
Disallow: /cgi-bin/  
User-agent: MOMspider/1.00  
Disallow: /cgi-bin/  
Disallow: /tec/software/packages/
```

[39] © Robert Tolksdorf, Berlin

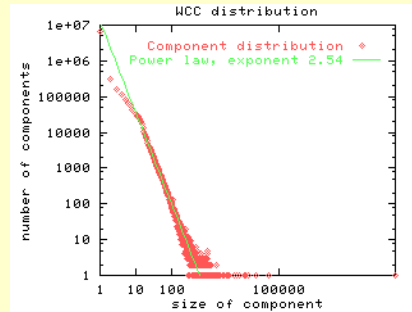
Masse (Maße) des Web

Nach: Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, Janet Wiener. Graph structure in the Web. Proc. 9th International World Wide Web Conference, 2000.

[40] © Robert Tolksdorf, Berlin

Komponenten im ungerichteten Graphen

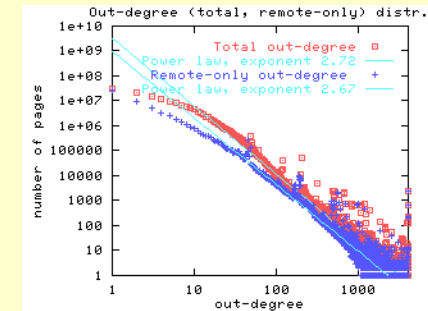
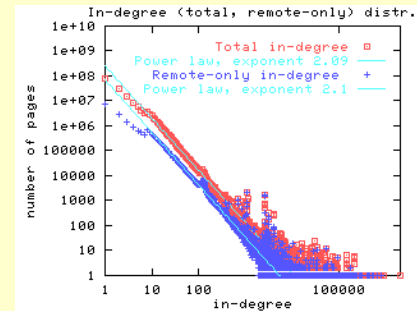
- Ungerichteter Graph (V,E) mit Kanten als $\{u,v\}$
- Pfad: $(u,u_1), (u_1,u_2), \dots, (u_k,v), \{u,v\} \Rightarrow (u,v), (v,u)$
- Komponente: Menge von Knoten, so dass für Knoten u und v im Graphen ein Pfad von u nach v existiert
- Eine große Komponenten mit 186m Knoten (91%)
- Verteilung der Größen der Komponenten hat Powerlaw mit $\frac{1}{n^{2,54}}$



[41] © Robert Tolksdorf, Berlin

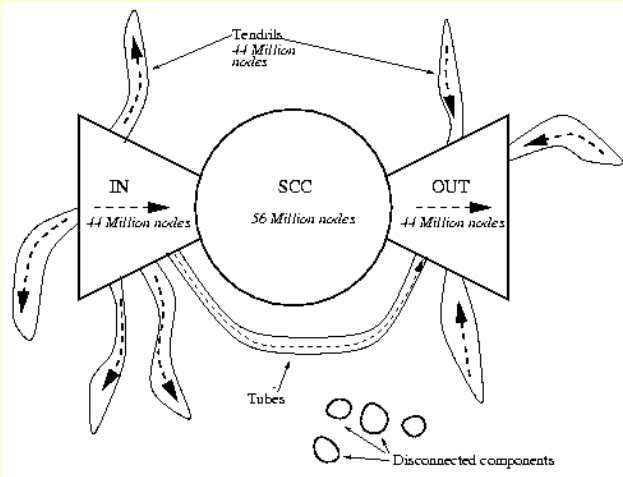
Messung in- und out-Degree

- Web: Gerichteter Graph (V,E) , Knoten V und Kanten E , Kante ist Paar (u,v) als Verbindung von u nach v
- in-degree: $|\{(u,v_1)\dots(u,v_k)\}|$, out-degree: $|\{(v_1,u)\dots(v_k,u)\}|$
- Anteil der Seiten mit in-degree i proportional zu $\frac{1}{i^{2,1}}$
- Anteil der Seiten mit out-degree i proportional zu $\frac{1}{i^{2,72}}$

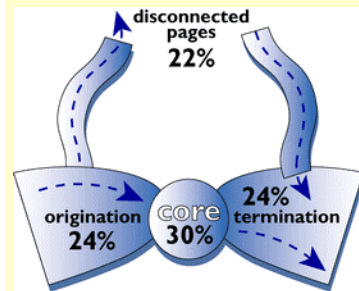


[42] © Robert Tolksdorf, Berlin

Struktur des Web



"Bow tie":



Region	SCC	IN	OUT	Tendrils	Disc.	Total
Grösse	56463993	43343168	43166185	43797944	16777756	203549046
Anteil	28%	21%	21%	22%	8%	100%

[43] © Robert Tolksdorf, Berlin

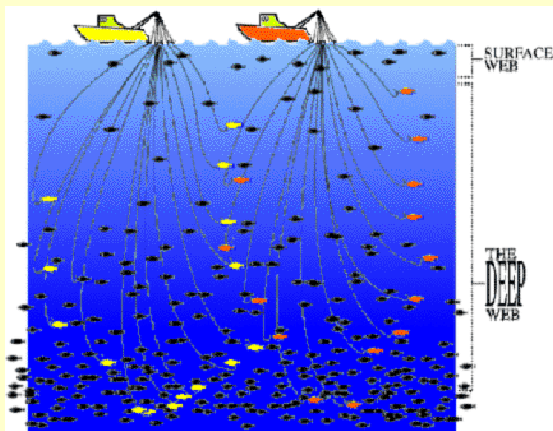
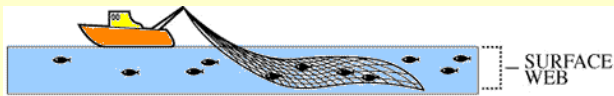
"Deep Web" Problematik

Nach: Michael K. Bergman. The Deep Web: Surfacing Hidden Value. The Journal of Electronic Publishing August, 2001 Volume 7, Issue 1 und <http://www.brightplanet.com/deepcontent/tutorials/DeepWeb>

[44] © Robert Tolksdorf, Berlin

"Deep Web"-Argumentation

- Traversierung des Web über Links führt nur zu einem Bruchteil der Informationen
- "Deep Web" wird von Datenbankinhalten gebildet
- Umfang 400-500 mal größer als "normales" Web
- 500Mrd Dokumente vs. 1 Mrd Dokumente
- Zugriff aber nur durch Datenbank-anfragen möglich



[45] © Robert Tolksdorf, Berlin

Information Retrieval

[46] © Robert Tolksdorf, Berlin

Information Retrieval

- Aufgabe des Information Retrievals: Technologien bereitstellen, die für eine Anfrage relevante Dokumente aus einer Sammlung von Dokumenten herausuchen
- Dokumente werden üblicherweise so vorverarbeitet und repräsentiert dass Anfrage einfach zu beantworten sind
- Bei Suchmaschinen üblich:
 - Volltextindex gesammelter Seiten erstellen
 - Anfragen an den Volltextindex weiterleiten
 - Ergebnisse ordnen
 - Verweise auf Ursprungsdokumente an Nutzer ausliefern

[47] © Robert Tolksdorf, Berlin

Terme in IR

- Zwei Arten von Termen in IR
- *Objektive* Terme:
 - Außerhalb des eigentlichen Inhalts
 - Beispiele: Autorennamen, URL etc.
 - Einfach und klar zuzuordnen
- *Nichtobjektive* Terme / *Inhaltsterme*:
 - Beschreiben Informationen des Dokumenteninhalts
 - Schwierig zuzuordnen
 - Hauptaufgabe des Indexing

[48] © Robert Tolksdorf, Berlin

Masse für Termzuordnung

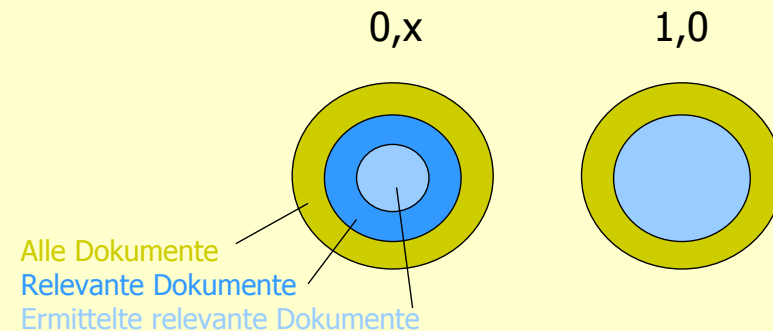
- *Indexing exhaustivity*:
 - Grad zu dem Inhalt durch Indexing erfasst wird
 - Hohe Ausschöpfung: Viele Terme zugeordnet
 - Geringe Ausschöpfung: Weniger Terme zugeordnet
- *Term specificity*:
 - „Breite“ der Terme beim Indexen
 - Breite Terme erfassen viele relevante und viele irrelevante Dokumente bei einer Anfrage
 - „Enge“ Terme erfassen weniger Dokumente und viele relevante nicht

[49] © Robert Tolksdorf, Berlin

Recall und Precision

- *Recall (Nachweisquote)*:
Wie gut findet das System relevante Dokument wieder?

$$\text{recall} = \frac{\text{Anzahl ermittelte relevante Dokumente}}{\text{Anzahl relevante Dokumente}}$$

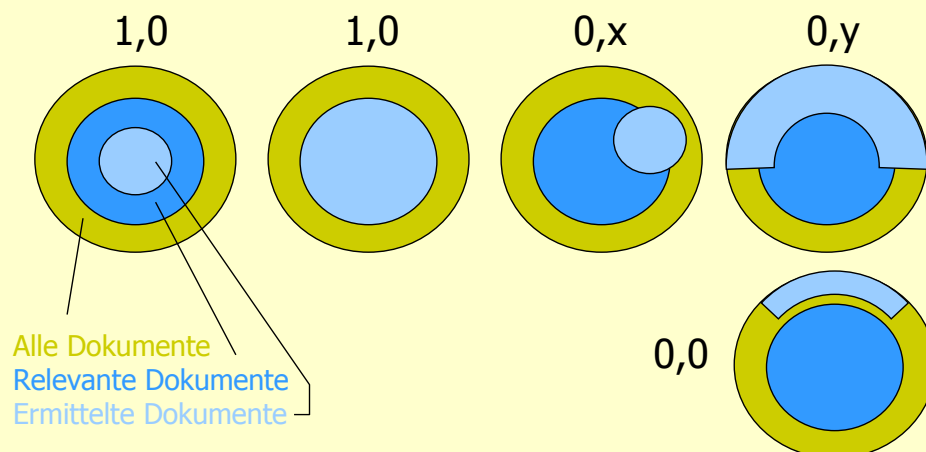


[50] © Robert Tolksdorf, Berlin

Recall und Precision

- *Precision (Präzision)*: Wie gut ist die Antwortmenge

$$\text{precision} = \frac{\text{Anzahl ermittelte relevante Dokumente}}{\text{Anzahl ermittelte Dokumente}}$$



[51] © Robert Tolksdorf, Berlin

IR Modelle

- Vier Eigenschaften
 - Repräsentation von Dokumenten und Anfragen
 - Feststellung der Relevanz eines Dokuments zu einer Anfrage
 - Ordnung der Ergebnismenge
 - Beachtung von Relevanz-Feedback durch Nutzer
- Vier Klassen von Modellen
 - Mengentheoretisch
 - Algebraisch
 - Stochastisch
 - Mischformen

[52] © Robert Tolksdorf, Berlin

Mengentheoretisch / Boolesches Retrieval

- $T = („heute“, „ist“, „dienstag“, „vorlesung“, „nicht“)$
- Dokumente d_1 : „heute ist dienstag“, d_2 : „heute ist vorlesung“, d_3 : „dienstag ist vorlesung“

	heute	ist	dienstag	vorlesung	nicht
d_1	1	1	1	0	0
d_2	1	1	0	1	0
d_3	0	1	1	1	0

	ist	dienstag AND vorlesung	heute OR dienstag	NOT vorlesung
d_1	1	1	1	1
d_2	1	0	1	0
d_3	1	1	1	0

[53] © Robert Tolksdorf, Berlin

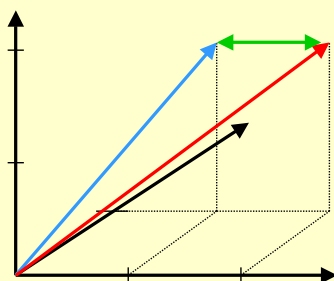
Algebraisch / Vector Space Modell

- Dokumente und Anfragen repräsentiert durch Vector in einem n-dimensionalen Raum
- Dimensionen durch Terme gegeben
- Gewichtet und normalisiert
- Relevanz durch Ähnlichkeitsmaß von Anfrage und Dokument gegeben und geordnet
- Sehr einfaches Modell
- Ausdrucksmächtigkeit boolescher Ausdrücke nicht vorhanden

[54] © Robert Tolksdorf, Berlin

Ähnlichkeit im Vektorraum

- $d_j = (w_{1,j}, \dots, w_{n,j})$ als Dokument
- $q = (q_1, \dots, q_n)$ ist Anfrage
- Terme sind gewichtet
- d_j und q sind Punkte in einem n-dimensionalen Raum
- Ähnlichkeitsmaß als Abstand zwischen den Punkten (Euklidische Distanz)



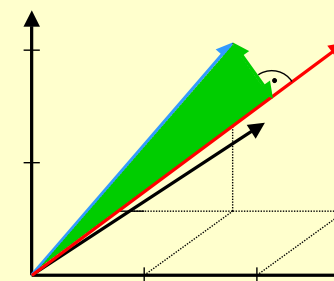
$$sim(d_j, q) = \sqrt{\sum |w_{i,j} - q_i|}$$

- Problem: Je mehr Terme, je größer der Abstand
- Dokumente haben mehr Terme als Anfragen
- Abstand immer groß

[55] © Robert Tolksdorf, Berlin

Ähnlichkeit im Vektorraum

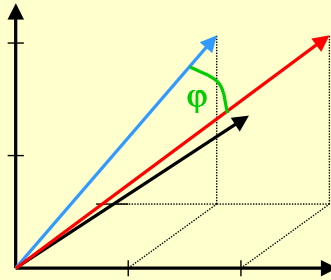
- Skalarprodukt als Ähnlichkeitsmaß:
- $$sim(d_j, q) = w_{1,j} * q_1 + \dots + w_{n,j} * q_n$$
- Entspricht Coordinate Match bei Gewichten 0 und 1
 - Problem: Je mehr Terme, je größer der Vektor, je größer das Skalarprodukt



[56] © Robert Tolksdorf, Berlin

Ähnlichkeit im Vektorraum

- Cosinusmaß: Nimmt Unterschied der Richtung der Vektoren, also den Winkel zwischen Dokument und Anfrage



$$\begin{aligned} \cos\varphi * |d_j| * |q| &= w_{1,j} * q_1 + \dots + w_{n,j} * q_n \\ \text{sim}(d_j, q) &= \cos\varphi \\ &= \frac{d_j \bullet q}{|d_j| * |q|} \\ &= \frac{\sum w_{i,j} * q_i}{\sqrt{\sum w_{i,j}^2} * \sqrt{\sum q_i^2}} \end{aligned}$$

[57] © Robert Tolksdorf, Berlin

Manuelles Indexieren

- Indexierer ordnen Dokumente per Hand in Kategorien ein oder bestimmen Indexterme

Service	Editoren	Kategorien	Links...	Datum
Open Directory	36000	361000	2.6 million	04/01
LookSmart	200	200000	2.5 million	08/01
Yahoo	>100	n/a	1.5 to 1.8 million	8/00

[<http://www.searchenginewatch.com/reports/directories.html>]

- Yahoo: In „beste“ Kategorie einordnen
- National Library of Medicine: Einordnung in so viele Kategorien des Medical Subject Headings (MeSH) Katalogs wie möglich
- Skalierungsproblem
- Spezialisierung als Ausweg

[58] © Robert Tolksdorf, Berlin

Automatisches Indexieren

- Notwendig wegen Web-Grösse
 - Crawling
 - Indexing
 - Anfragemanagement

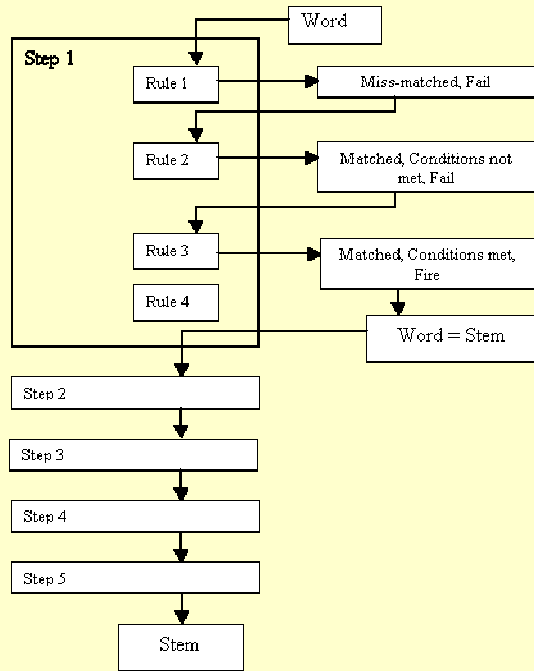
[59] © Robert Tolksdorf, Berlin

Normalisierung

- Dokumentenvorbereitung
 - Parsieren/Entfernen von HTML
 - Ermittlung indexierungsrelevanter Informationen
 - alt Attribut bei
 - <meta>-Tags
 - lang Attribut
 - ...
 - Umgang mit Zeichenkodierungen
 - Entitäten expandieren
 - Interpunktion entfernen
- Aufteilen in Token
- Stop words entfernen
- Stemming

[60] © Robert Tolksdorf, Berlin

Porter Stemmer



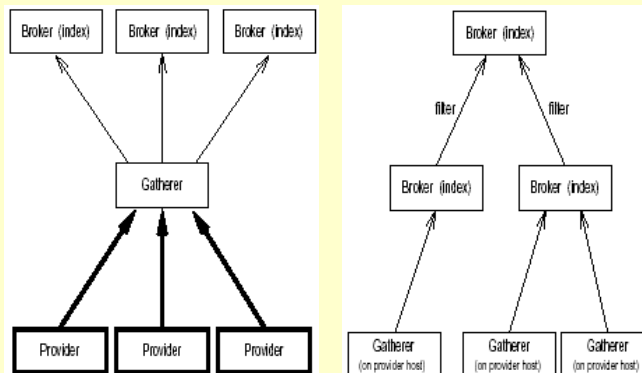
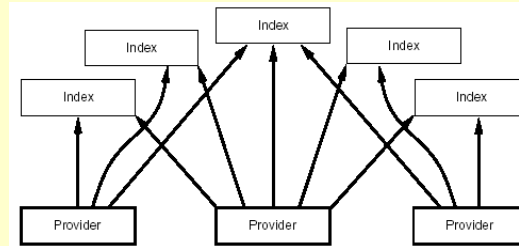
[61] © Robert Tolksdorf, Berlin

Collaborative Indexing

[62] © Robert Tolksdorf, Berlin

Harvest System

- Indexe lesen jeweils komplette Information aus Servern aus
- Harvest-System:
 - Gatherer extrahieren Quellen
 - Broker stellen Index her
 - Broker stellen beantworten Anfragen
 - Kaskadierte Gatherer



[63] © Robert Tolksdorf, Berlin

Multimedia Indexing

[64] © Robert Tolksdorf, Berlin

Multimedia Indexing

- Große Anteile der im Netz verfügbaren Informationen sind kein Text und können nicht in einem Volltextindex gespeichert werden
- Beispiele:
 - Bilder – Fotos, Zeichnungen
 - Audio – Musik, Sprache
 - Video – Filme, Nachrichten
- Oftmals ist Information in mehreren Medien verteilt
 - Beispiel: Nachrichten, Bild und Ton
- Problem
 - Ermittlung von Indextermen
 - Entwurf von Ähnlichkeitsmaßen

[65] © Robert Tolksdorf, Berlin

Audioanfragen

- Anfragbare Eigenschaften von Audio:
 - Ähnliche Sounds („wie eine Elefantenherde“, „Applaus“)
 - Akustische Eigenschaften (Laut, tief, schnell)
 - Subjektive Beschreibung (sanft)
 - Onomatopoeia („eine der zentral verwendeten Tropen im Chat ist die Onomatopoeia“) Lautmalerei (buzz, bumm, klingeling)

[66] © Robert Tolksdorf, Berlin

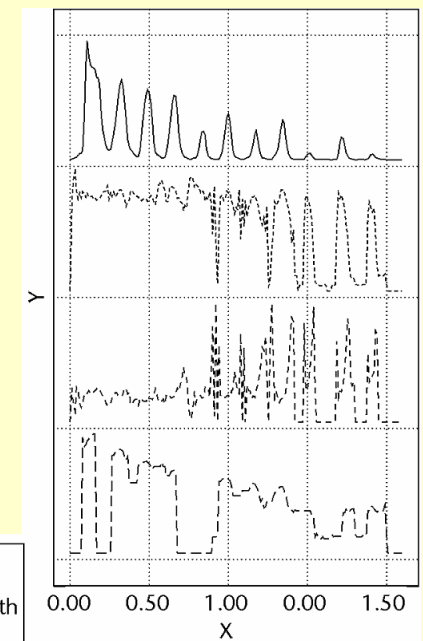
Anfragemöglichkeiten

- Stufen von Inhalts-Retrieval von Audio
 - Numerische Anfrage nach Sample-Daten (Text: Stringvergleich)
 - Enthaltensein bestimmter Audioteile unabhängig von deren Geschwindigkeit, Höhe etc. (Text: Ähnlichkeit von Zeichenketten)
 - Vorhandensein bestimmter Maße (Text: tf, etc)
 - Inhalt des Audiostücks (Text: Semantische Suche)

[67] © Robert Tolksdorf, Berlin

Ansatz

- Akustische Maße ermittelt
- Als N-Vektor repräsentiert
- Dimensionen
 - Lautstärke
 - Takt
 - Helligkeit (Brightness)
 - Bandbreite
 - Harmonie (Abweichung vom harmonischen Spektrum)



[68] © Robert Tolksdorf, Berlin

Ansatz

- Maße variieren über die Zeit, Dynamik erfasst durch:
 - Mittelmaße
 - Varianz
 - Autokorrelation
 - ...
 - Gewichtet durch Lautstärke

Property	Mean	Variance	Autocorrelation
Loudness	-54.4112	221.451	0.938929
Pitch	4.21221	0.151228	0.524042
Brightness	5.78007	0.0817046	0.690073
Bandwidth	0.272099	0.0169697	0.519198

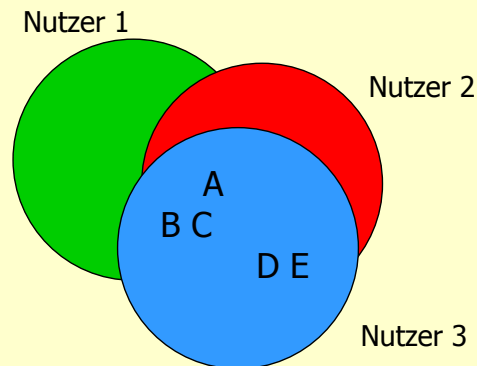
[69] © Robert Tolksdorf, Berlin

Collaborative Filtering

[70] © Robert Tolksdorf, Berlin

Grundidee

- Ähnliche Nutzer haben ähnliche Vorlieben
- Vorlieben eines Nutzers können genutzt um einem anderen einen Vorschlag zu machen
- Beispiel: A...E sind Produkte, Informationen etc.
- Nutzer 1, 2 und 3 ähneln sich, da sie alle A, B und C mögen/haben
- Für Nutzer 1 sind wahrscheinlich auch D und E relevant



[71] © Robert Tolksdorf, Berlin

PageRank

[72] © Robert Tolksdorf, Berlin

PageRank

- PageRank Verfahren: Bewertung aller Web-Seiten nach ihrer relativen Wichtigkeit
- Kerntechnologie von Google
- Viele Links auf eine Seite legen nahe, dass die Seite wichtig ist (www.yahoo.com etc)
- Setzen eines Links ist Einschätzung der Wichtigkeit der referenzierten Seite (ähnlich einem Zitat)
- Links von wichtigen Seiten erhöhen Wichtigkeit
- Eine Seite hat hohen PageRank, wenn die PageRanks der Seiten, die auf sie verweisen, hoch sind
 - Viele mittelwichtige Seiten verweisen
 - Wenige hochwichtige Seiten verweisen

[73] © Robert Tolksdorf, Berlin

PageRank

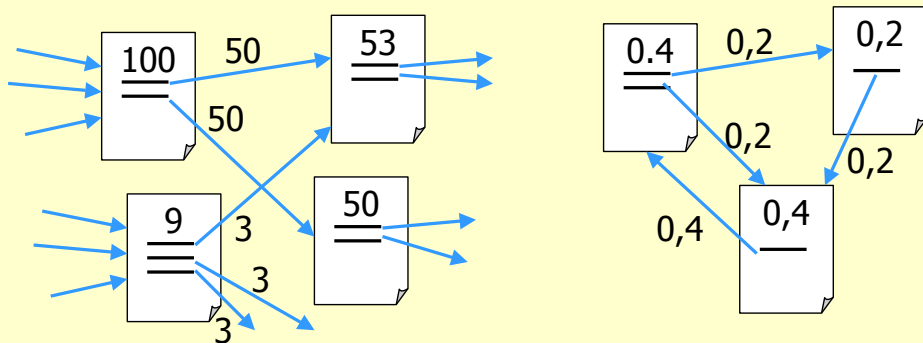
- Web-Seite u
- F_u : Seiten, auf die u verweist
- B_u : Seiten, die auf u verweisen
- $N_u = |F_u|$: Anzahl der Links von u (out-degree)
- c : Faktor zur Normalisierung
- Vereinfachter PageRank $R(u)$ der Seiten u :

$$R(u) = c * \sum_{v \in B_u} \frac{R(v)}{N_v}$$

- c normalisiert die Summe aller PageRanks zu 1

[74] © Robert Tolksdorf, Berlin

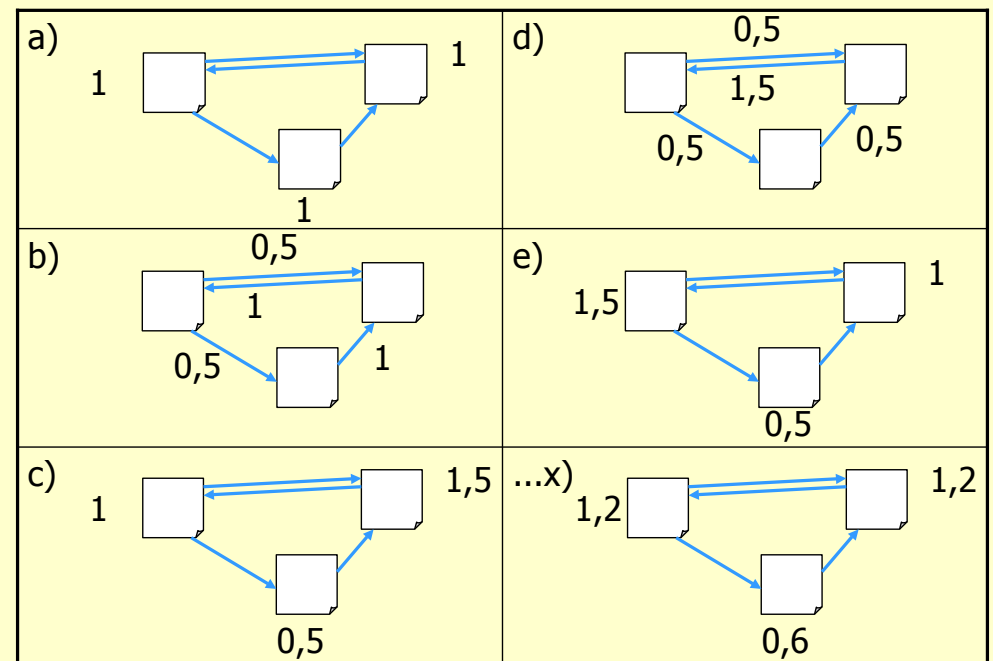
Verteilung der Ranks



- Ist iterativ errechenbar

[75] © Robert Tolksdorf, Berlin

Beispieliterationen



[76] © Robert Tolksdorf, Berlin

Autoritäten im Netz / HITS

Nach: Jon M. Kleinberg: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM 46(5): 604-632 (1999)
<http://www.cs.cornell.edu/home/kleinber/auth.pdf>

Autoritative Quellen

- Autoritative Antworten als Relevanzfilter
- Wie findet man autoritativ Seiten?
 - www.harvard.edu für "Harvard"?
 - Suche nach "Search engine" für Google?
 - Suche nach "automobile manufacturers" für Honda?
- Fachliche Autorität ist nicht ausschließlich endogene Eigenschaft sondern in grossem Mass exogen
- Schreiben Hyperlinks schreiben dem Ziel eine fachliche Autorität zu?

Art der Links

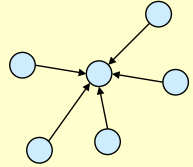
- Nicht alle Links verfolgen gleiche Absicht
 - Hinweis auf wichtige Seiten
 - Navigationsunterstützung
 - Anzeigen
- Popularität vs. Autorität
 - Viele Seiten die einen Begriff enthalten und oft verlinkt werden sind nicht autoritativ
 - Seiten die allgemeinen Inhalt haben und oft verlinkt werden sind nicht für alle Themen autoritativ (eg. www.yahoo.com)

Analyseziel

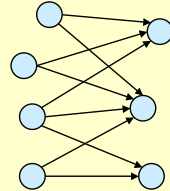
- Anfrage σ
- Ziel: Analyse der Linkstruktur in einem Teil des Web um autoritative Quellen zu finden
- Q_σ (alle Seiten, die σ enthalten)?
 - Zu gross
 - Autoritative Quellen eventuell nicht enthalten
- Gesucht: S_σ so dass:
 - S_σ vergleichsweise klein ist (i)
-> Aufwand begrenzt
 - S_σ viele relevante Seiten enthält (ii)
-> gute Autoritäten auffindbar
 - S_σ die meisten oder viele Autoritäten enthält (iii)

Autoritäten und Hubs

- Autoritative Quellen in G_σ sollten sich dadurch auszeichnen, dass die Mengen der Seiten, die auf sie zeigen überlappen
- *Hubs* verweisen auf mehrere Autoritäten



grosser in-degree



Hubs Autoritäten

- Guter Hub zeigt auf gute Autoritäten
- Gute Autorität wird von vielen guten Hubs verwiesen
- (PageRank ermittelt nur Autoritäten)

[81] © Robert Tolksdorf, Berlin

Metasuchmaschinen

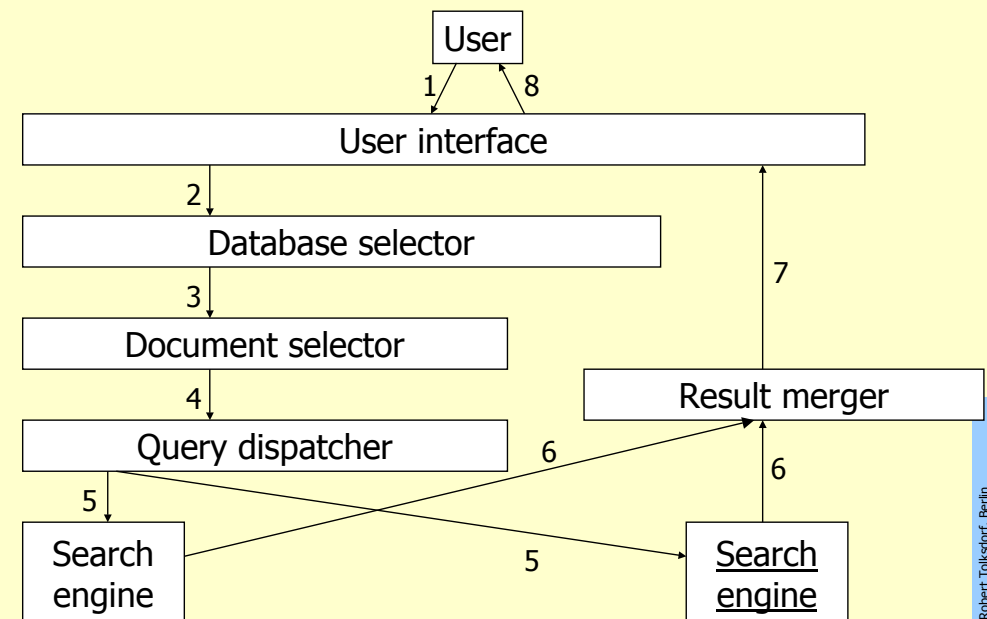
[82] © Robert Tolksdorf, Berlin

Vorteile

- Bessere Abdeckung des Suchraums Web: Suchmaschinen decken immer nur einen Teil des Web aus, Metasuchmaschinen deren Summe
- Bessere Skalierbarkeit
Suche wird auf Suchmaschinen verteilt
- Bessere automatische Absuche mehrerer Suchmaschinen
Keine manuellen Abfragen notwendig
- Effektivität der Suche verbessern
Spezialsuchmaschinen clustern Dokumente thematisch

[83] © Robert Tolksdorf, Berlin

Architektur von Suchmaschinen



[84] © Robert Tolksdorf, Berlin

Nutzung und Nutzer von Web-Sites

Nutzungs-/Nutzerinformationen

- Nutzer von Web-Sites sind für den Server anonym
 - Keine Identifikation des tatsächlichen Rechners: Proxies, Caches, private Netze, dynamische IP-Nummern
 - Keine Identifikation des Nutzerprozesses: Mehrbenutzerrechner, Proxies, Caches
 - Keine Identifikation des Nutzers: Account-Informationen lokal
- Informationen über Nutzer sind aber nützlich
 - Personalisierung
 - Optimierung des Angebots
 - Grundlage des Geschäftsmodells

Messgrößen

- Auf Basis von Logfiles lassen sich verschiedene Aussagen über die Nutzung einer Site treffen
- Insbesondere sind diese Aussagen Basis für die Preisfindung der Werbewirtschaft
- Diese Aussagen sind von unterschiedlicher Güte

Messgrößen

- Hits
 - Anzahl der Abrufe von Informationen
 - Summe der Anzahl der Requests mit 200 und 304 Antwort
 - Nicht sehr aussagekräftig, weil nicht jede Datei eigenständige Informationseinheit
- Pageviews/Page Impressions
 - Anzahl der abgerufenen HTML-Seiten
 - Anzahl der Hits mit HTML Dateien als Antwort
 - Beschränkt auf einen Medientyp

Messgrößen

- Visits / Sessions
 - Zusammenhängende Abrufe in einem Zeitraum
 - Navigationspfade aus Logfile
 - Nicht zuverlässig identifizierbar
 - Problem: Wann ist Visit beendet?
- Heuristiken
 - Zeitorientiert:
 - Gesamtdauer einer Visit ist nach oben begrenzt
 - Verweildauer auf einer Seite ist nach oben begrenzt
 - Navigationsorientiert
 - Topologische Begrenzung: Sitzungsende, wenn Seite nicht von vorherigen Seiten aus erreicht werden konnte
 - Begrenzung durch Referrer: Sitzungsende, wenn Seite nicht durch Navigation von vorheriger Seite erreicht wurde

[89] © Robert Tolksdorf, Berlin

Messgrößen

- Unique Visitors
 - Abrufe von gleicher IP Adressen als 1 Besucher gezählt
 - Objektiv nicht aussagefähig (Proxies, Dynamische IP Adressen, etc.)
- AdImpressions / Clickthroughs
 - Klick auf Werbebanner
 - Messbar beim Werbekunden
 - Quelle durch Referer ermittelbar
 - Abrechnung
 - Preis nach Attraktivität des Werbeträgers: Pageviews und Visits als Maß
 - Preis nach Effizienz des Werbemittels: Clickthroughs als Maß

[90] © Robert Tolksdorf, Berlin

Messgrößen

- Viewtime
 - Dauer des Verweilens auf einem Angebot
 - Kaum aus Logfile messbar
 - Klientenseitige Unterstützung notwendig (z.B. Skripting)
 - Sitzt der Nutzer vor dem Rechner?
- Durch zusätzliche direkte Befragung ermittelbar:
 - Qualified visits: Bestätigte Besuche
 - Regionale Herkunft
 - Alter, Geschlecht etc.
 - Interessen
 - Akzeptanz

[91] © Robert Tolksdorf, Berlin

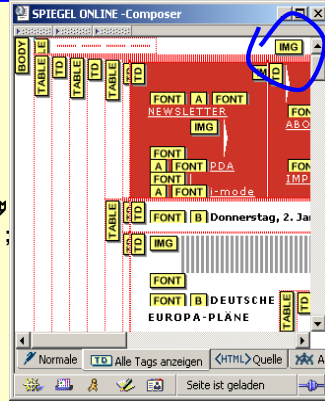
Wer misst?

- Serverbetreiber nach eigenem Verfahren und eigener Auswertung
- Serverbetreiber oder Externer nach standardisiertem Verfahren und Auswertung
 - Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern e.V." (IVW) (<http://www.ivwonline.de/>)
 - Messung z.B. durch transparente Grafiken („IVW-Pixel“) auf Seiten
 - ``
 - ``
 - Lösen Messung aus
 - IVW Zahlen sind Grundlage für Preisgestaltung

[92] © Robert Tolksdorf, Berlin

Aus www.spiegel.de/index.html

```
<body bgcolor="#ffffff" text="#000000"
link="#b20a15" vlink="#b20a15" alink="#ff0000"
marginheight="0" marginwidth="4" leftmargin="4"
topmargin="0" rightmargin="4" bottommargin="0">
<!-- IWV VERSION="1.2" -->
<script language="JavaScript">
<!--
var IWV="http://spiegel.iwvbox.de/cgi-bin/iwv/CP/
spiegel;/home/c-18/be-PB64-ag9tZXBhZ2UvY2VudGvy";
document.write('<IMG SRC="'+IWV+'?r='+
escape(document.referrer)+'" WIDTH="1,"
HEIGHT="1" BORDER="0" ALIGN="RIGHT">');
// -->
</script>
```

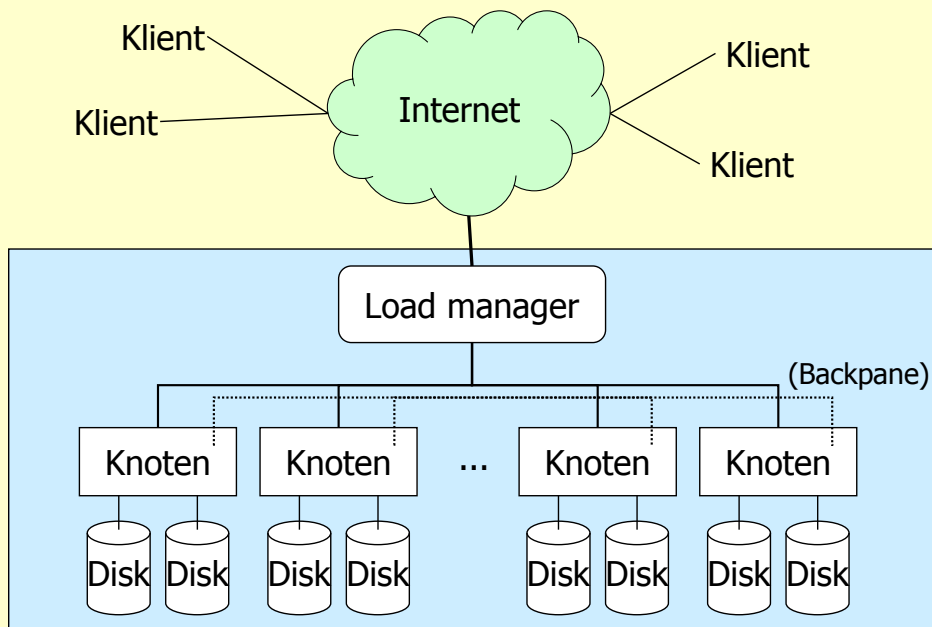


```
<noscript>
<IMG SRC="http://spiegel.iwvbox.de/cgi-bin/iwv/CP/spiegel;/home/c-18/be-
PB64-ag9tZXBhZ2UvY2VudGvy" WIDTH="1" HEIGHT="1" BORDER="0" ALIGN="RIGHT">
</noscript>
<!-- /IWV -->
<!-- IWV VERSION="prev" -->

<!-- /IWV -->
```

Betriebsaspekte sehr grosser Dienste

Grundlegendes Modell großer Server



Load management

- Layer-4-switches
 - Hardware die TCP "versteht"
 - Switch leitet Pakete aufgrund von TCP-Dienste-Feld und Portnummern an unterschiedliche Server weiter
- Layer-7-switches
 - Hardware die HTTP "versteht"
 - Können URLs mit Netzbandbreite parsieren und leiten Pakete entsprechend weiter
- Meistens als Paar vorhanden
- >20Gbits/s Durchsatz
- Automatisches Monitoring von Knoten

Verfügbarkeit

- Zentrale Anforderung an grosse Dienste:

Verfügbarkeit (Availability)

- Gemessen in "Neunern":
 - Vier Neuner: 0,9999 Verfügbarkeit (<60 Sek. Ausfall/Woche)
 - Fünf Neuner: 0,99999
- Ähnlich geleitete Systeme:
 - Telefonsystem
 - Zugverkehr
 - Wasserversorgung

[97] © Robert Tolksdorf, Berlin

Yield und Harvest

- Weiteres Maß: Yield – vielen Anfrageergebnisse?

$$\text{yield} = \frac{\text{bearbeitete Anfragen}}{\text{gestellte Anfragen}}$$

- entspricht Nutzererfahrung
- gewichtet Uptime-Sekunden

- Weiteres Maß: Harvest – welcher Teil der Datenbank ist nutzbar?

$$\text{harvest} = \frac{\text{zugreifbare Daten}}{\text{gesamte Daten}}$$

- Erweiterbar zum Anteil der nutzbaren Dienste

[98] © Robert Tolksdorf, Berlin

Upgrades / Wartung

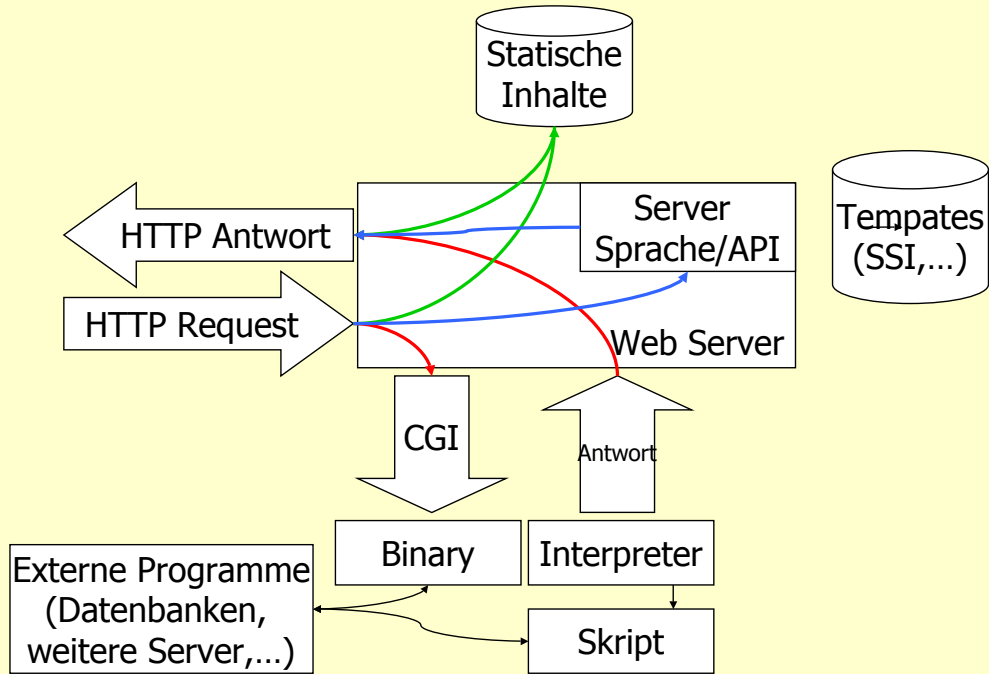
- Schneller Neustart
 - Maschinen werden in neue Konfiguration gebootet
 - Yield geht verloren
 - Optimierte durch geeigneten Zeitpunkt (off-peak)
- Rolling Upgrade
 - Jeweils ein Knoten upgraden
 - Replikation: Yield sinkt minimal
 - Partitionierung: Harvest sinkt
 - Zwei Versionen müssen verträglich koexistieren
- „Big Flip“
 - Halber Cluster wird neu gestartet, danach andere Hälfte
 - Durch Layer-4 Switch transparent nach aussen
 - 50% DQ Verlust

[99] © Robert Tolksdorf, Berlin

Server- und Klientenseitige Verarbeitung Überblick

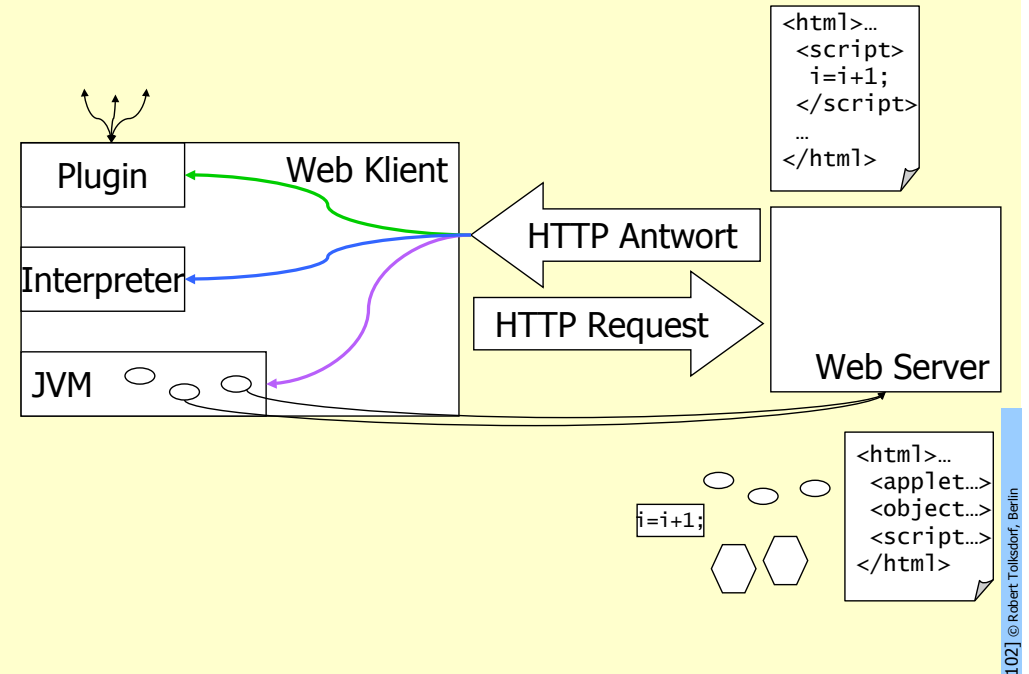
[100] © Robert Tolksdorf, Berlin

Im Überblick



[101] © Robert Tolksdorf, Berlin

Im Überblick



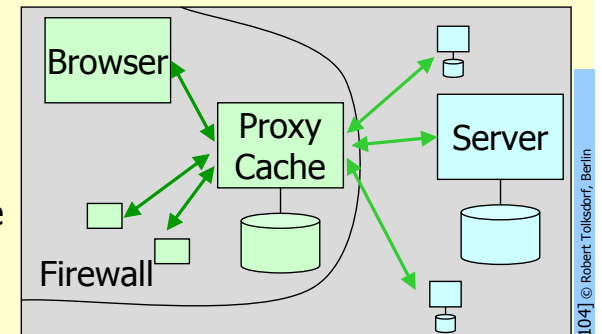
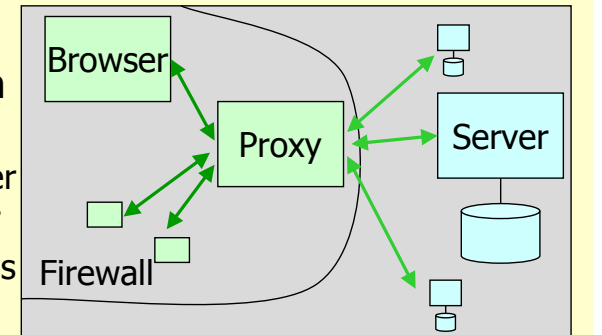
[102] © Robert Tolksdorf, Berlin

Caching im Web

[103] © Robert Tolksdorf, Berlin

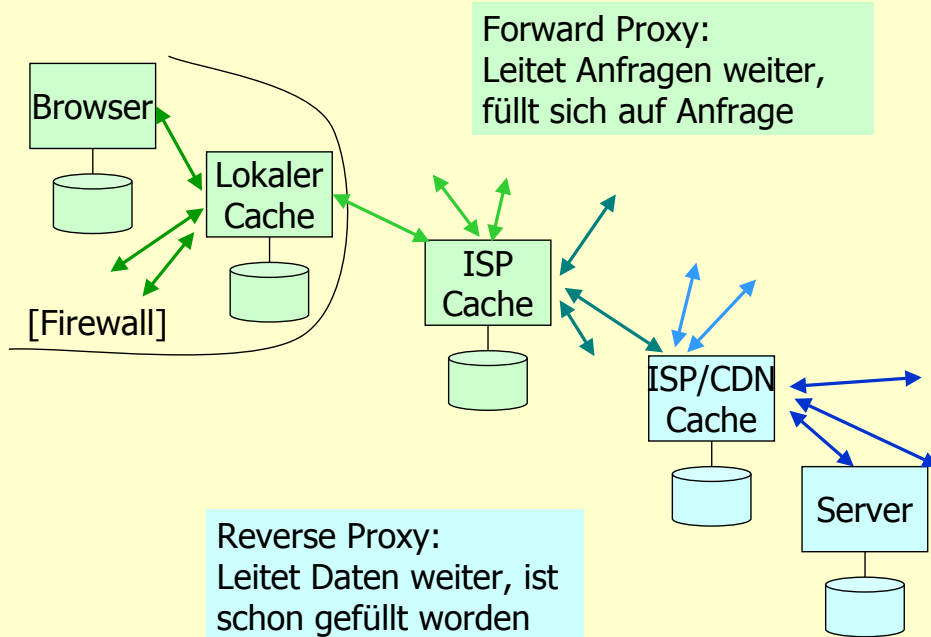
Proxies und Caching

- *Proxy/Stellvertreter* anstelle vom Klienten
 - Leitet HTTP Anfrage von Klienten an Server andere Proxies weiter
 - Tritt für den Server als Klient auf
 - Leitet Antwort an Klienten weiter
- *Proxy Cache*
 - Agiert auch als Cache
 - → Annahme: Zugriffe organisatorisch gruppiert



[104] © Robert Tolksdorf, Berlin

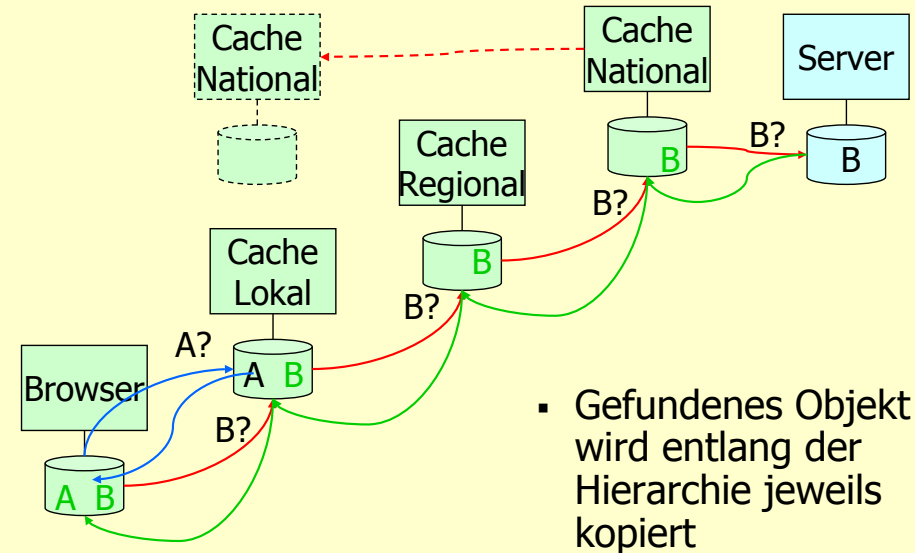
Caches im Web allgemein



[105] © Robert Tolksdorf, Berlin

Cache Architekturen - Hierarchisches Caching

- Hierarchisches Caching: Anfrage wird bei einem Cache-Miss über mehrere Hierarchiestufen weitergereicht

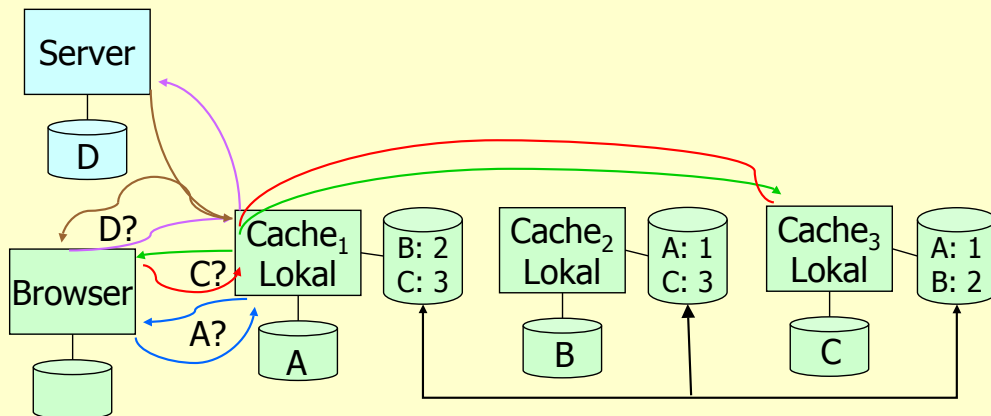


- Gefundenes Objekt wird entlang der Hierarchie jeweils kopiert

[106] © Robert Tolksdorf, Berlin

Cache Architekturen - Verteiltes Caching

- Caches kooperieren
 - wissen, welcher Cache welche Objekt hält (Tabelle, Hashing,...)
 - fragen nach (→ Harvest ICP)
 - Ältere Forschungsprototypen, nicht verbreitet



[107] © Robert Tolksdorf, Berlin

Darstellung von Inhalten

[108] © Robert Tolksdorf, Berlin

Cascading Style Sheets CSS

- ... Mechanismus zur separaten Definition von Stileigenschaften für HTML und XML Dateien (Quelle: <http://www.w3.org/Style/CSS/>)
- Cascading Style Sheets, level 1
 - Ziel: Sprache zur Definition des Darstellungsstils von HTML Dokumenten
 - Status: W3C Recommendation 17 Dec 1996, revised 11 Jan 1999
 - Quelle: <http://www.w3.org/TR/REC-CSS1>
- Cascading Style Sheets, level 2
 - Ziel: Sprache zur Definition des Darstellungsstils von HTML und XML Dokumenten für unterschiedliche Medienarten
 - Status: W3C Recommendation 12-May-1998
 - Quelle: <http://www.w3.org/TR/REC-CSS2>
- Cascading Style Sheets, level 3
 - Ziel: Modularisierte und erweiterte Sprache zur Definition des Darstellungsstils von HTML und XML Dokumenten
 - Status: unterschiedlich, erste Recommendations eventuell April 2003
 - Quelle: <http://www.w3.org/Style/CSS/current-work>

[109] © Robert Tolksdorf, Berlin

CSS Grundidee

- Grundidee:
Zu HTML Tags werden definierte Attribute für Darstellungseigenschaften gesetzt
- CSS-Datei getrennt von HTML-Datei gehalten
- Beispiel: Um Überschriften in grosser blauer Schrift darzustellen:

```
h1 {color: blue; font-size: 22pt; }
```
- CSS definiert
 - Rahmensyntax zur Notation
 - Menge von Attributen
 - Menge von Werten
 - Bedeutung
 - Mechanismen zur Anbindung von Stilinformationen an und in HTML Seiten

[110] © Robert Tolksdorf, Berlin

CSS Anbindung an HTML

- Drei Wege der Einbindung in HTML
 - Mit externem Stylesheet über Verweis im <link>-Tag:

```
<link rel="stylesheet" type="text/css" href="http://www.inf.fu-berlin.de/inst/ag-nbi/include/nbi.css">
```
 - Im HTML Dokument mit dem <style>-Tag:

```
<style>
h1 {color: blue; font-size: 22pt; }
</style>
```


Kompatibel für alte Klienten:

```
<style><!--
h1 {color: blue; font-size: 22pt; }
--></style>
```
 - Bei den einzelnen Elementen im style-Attribut:

```
<h1 style="color: red">Rote Überschrift</h1>
```
- Einbindung innerhalb eines CSS
 - `@import url(http://www.inf.fu-berlin.de/inst/ag-nbi/include/colors.css);`

[111] © Robert Tolksdorf, Berlin

CSS2: Medienarten

- Darstellungsstil ist abhängig vom Ausgabemedium
 - Bildschirm
 - Papier
 - Spache
 - Braille
 - ...
- CSS erlaubt getrennte Stildefinitionen:
...

```
a:link {
  color: #000099;
  text-decoration : none ;
}
@media print {
a:link,a:visited,a:hover,a:active,a:focus {
  text-decoration:none;
  color:blue
}
}
```

[112] © Robert Tolksdorf, Berlin

XML und Darstellung

[113] © Robert Tolksdorf, Berlin

XML+CSS

- Cascading Style Sheets definieren Darstellung von Tags durch Belegen von CSS-Attributen
- Während ursprünglich für HTML entworfen auch für XML nutzbar
- Darstellung vom eigenen Element `<price>` weiss auf schwarz:

```
price {
  color: white;
  background-color: black;
}
```
- CSS Attribute für visuelle oder auditive Ausgabe von Texten geeignet
- www.w3.org/1999/06/REC-xml-stylesheet-19990629

[114] © Robert Tolksdorf, Berlin

XSL

- CSS Vorgehen:
 - HTML enthält Struktur in Inhalt
 - CSS definiert Darstellungseigenschaften
 - Struktur fest
- XSL Standard
 - XML enthält Inhalt
 - XSL transformiert in Darstellungsstruktur (und deren Eigenschaften)

[115] © Robert Tolksdorf, Berlin

XSL Regeln

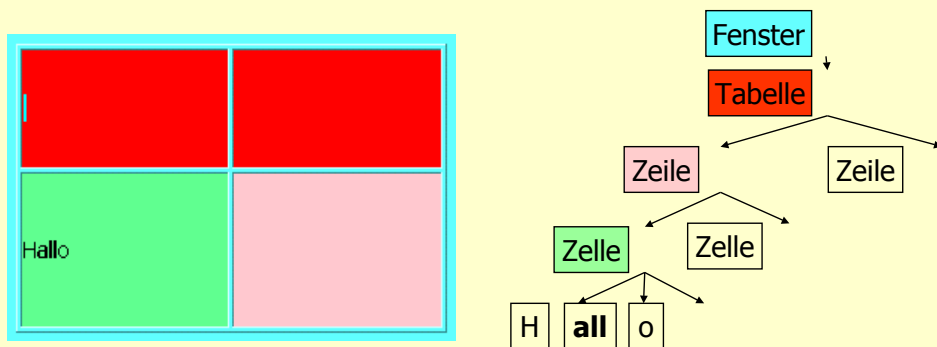
- Eine XSL-Regel definiert Muster und dazugehörige Transformationen
- Sie wird auf alle passenden Knoten im Quelldokument angewandt
- Beispiel:

```
<xsl:template match="ANSCHRIFT">
  <tabelle>
    <xsl:apply-templates select="NAME"/>
  </tabelle>
  <tabelle>
    <xsl:apply-templates select="ORT"/>
  </tabelle>
</xsl:template>
```

[116] © Robert Tolksdorf, Berlin

Formatting Objects

- Bildschirmdarstellung ist auch Baum:



- Formatting Objects definiert Knoten und Attribute als Ziel einer Transformation
- Als "Nebeneffekt": Bildschirmdarstellung, PDF Generierung

[117] © Robert Tolksdorf, Berlin

Mehrsprachigkeit im Web Zeichen, Schriften, Sprachen

[118] © Robert Tolksdorf, Berlin

Internationalisierung

- *Internationalisierung* ist die Planung und Implementierung von Diensten und Produkten so dass sie einfach an lokale Sprachen und Kulturen anpassbar sind, was *Lokalisierung* ist
- Internationalisierung
 - „I18N“ - „I - eighteen letters -N“ – „Internationalization“
 - Voraussetzung für Lokalisierung
 - Beispiele
 - Platzgestaltung in GUIs läßt Raum für Sprachen die mehr Zeichen benötigen
 - Verwendung internationaler Zeichenrepertoires und -codes, z.B. Unicode
 - Vergabe leicht übersetzbarer Beschreibungen für Graphiken
 - Verwendung allgemeinverständlicher Beispiele (Social Security Number ...)
 - Vorausplanung der Übersetzung in Sprachen mit Kodierungen mit mehr als einem Byte pro Zeichen in Software

[119] © Robert Tolksdorf, Berlin

Lokalisierung

- *Lokalisierung* ist die Anpassung eines Produktes oder Dienstes an eine Sprache, Kultur und lokales "look-and-feel" was durch *Internationalisierung* vereinfacht wird
- Lokalisierung
 - „L10N“ – „L - ten letters -N“ – „Localization“
 - Übersetzung
 - Aber auch: Anpassung an Zeitzonen, Währung, Feiertage, Farbkonventionen, Namen, Geschlechterrollen etc.
 - Ziel: Lokalisiertes Produkt oder Dienst soll so aussehen, als sei er/es lokal entwickelt worden

[120] © Robert Tolksdorf, Berlin

Sprachbezeichner

- Sprachen im Internet durch Codes bezeichnet
- Basis nach RFC 3066 (früher 1766)
 - In ISO 639 definierte Kürzel für Sprachen
 - In ISO 3166 definierte Kürzel für Länder
- Format
 - Sprachcode: de en etc.
 - Sprachcode-Ländercode: de-ch en-uk
 - Matching nach Substring am Anfang en passt auf en-us
 - Groß-/Kleinschreibung irrelevant en passt auf En-us und EN
 - Experimentell: x-klingon
(siehe auch <http://www.google.com/intl/xx-klingon/>)
- Nicht perfekt: Lateinamerikanisches Spanisch?

[121] © Robert Tolksdorf, Berlin

Spracheigenschaften in HTML

- Alle HTML Elemente können Sprachbezogene Attribute tragen
 - lang-Attribut: Wert ist Sprachcode
 - Wird vom umgebenden Element „geerbt“
 - Kann jeweils überschrieben werden
 - Default ist durch Content-language HTTP Header gegeben
 - dir-Attribut: (Horizontale) Schreibrichtung der Schrift
 - ltr: Left-to-Right
 - rtl: Right-to-Left
 - Wird vom umgebenden Element „geerbt“
 - Kann jeweils überschrieben werden

[122] © Robert Tolksdorf, Berlin

Spracheigenschaften in CSS

- In CSS2 neue Pseudoklasse :lang

```
:lang(en) {color: red}
:lang(fr) {color: blue}
```

 - Noch nicht implementiert
- In CSS2 Selektorenausdrücke auf Inhalt des lang Attributs

```
*[lang|=en] {color: red}
*[lang|=fr] {color: blue}
```

Ein Absatz mit einem **chaotic** Sprachgebrauch **ridicule**.

```
<p>Ein Absatz mit einem <span lang="en-us">chaotic</span>
Sprachgebrauch <span lang="fr">ridicule</span>.</p>
```
- Eigenschaft direction mit Werten ltr und rtl
- Eigenschaft unicode-bidi
 - Werte normal, embed, bidi-override
 - <bdo>=unicode-bidi: bidi-override

[123] © Robert Tolksdorf, Berlin

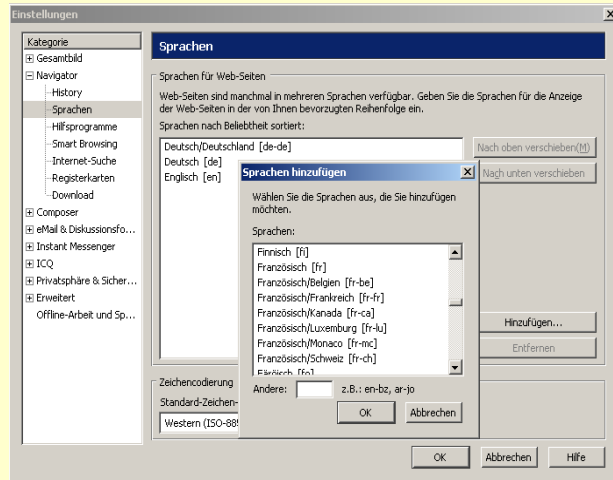
Spracheigenschaften in XML

- Attribut xml:lang in XML definiert, also immer verfügbar
- Bedeutung wie lang in HTML
- In XHTML sowohl xml:lang als auch lang benutzen

[124] © Robert Tolksdorf, Berlin

Sprache in HTTP

- Browser kann Präferenzen im HTTP-Request mitteilen:
GET / HTTP/1.1
AcceptLanguage: en-us;q=0.75,en;q=0.5;*;q=0.25
- q gibt
Priorität an,
* ist
Platzhalter
- Vom Browser
abhängig:



Zeicheneigenschaften

Zeicheneigenschaften

- Zeichenrepertoire (Character Set, Abstract Character Repertoire, ACS)
 - Eine Menge von Zeichen
 - Definiert durch Namen und Beispiele
 - {Pfund (£), Zett (Z), Ypsilon (Y), Herz (♥)}
 - Keine Ordnung, keine Codierung
- Zeichencode (Coded Character Set, CCS)
 - Abbildung(en) Zeichen → Zeichenposition
 - Z → 5A, خ → FEA5 (Khah)
 - z.B. UNICODE, ISO 8859-1

Zeicheneigenschaften

- Zeichenkodierung (Encoding)
 - Character Encoding Form (CEF)
 - Abbildung einer Zeichenfolge auf Strom gleichgroßer Codes
 - z.B.

005A	FEA5
------	------
 - Character Encoding Scheme (CES)
 - Abbildung einer Zeichenfolge auf einen Bytestrom
 - z.B.

5A	00	A5	FE
----	----	----	----
- Zeichensatz
 - Bedeutung unklar, kann Repertoire, Code oder Kodierung meinen
- „charset“
 - meint Encoding!

Metadaten im Web

Ermittlung der Semantik von Dokumenten

- Ermittlung der Bedeutung von Dokumenten:
 - Manuelles Indexing: Manuelle Termvergabe
 - Automatisches Indexing: Automatische Termvergabe auf statistischer Basis
 - Filtering: Indirekt durch Einschätzung der Bedeutung für Nutzer
 - Textverstehen: Computerlinguistische Verfahren
- Explizite Bekanntgabe der Bedeutung von Dokumenten
 - Inhaltsinformationen: Textueller Inhalt
 - Objektive Metainformationen: Datum, Größe...
 - Inhaltliche Metainformationen: Term
- Durch vorgefundene Metainformationen erübrigt sich die Ermittlung von Metainformationen
- Dezentrale Bereitstellung

Syntaktische und semantische Verweise in HTML

- Syntaktisch:
Berlin
- Beziehung durch Link gegeben, aber:
 - Welche inhaltliche Beziehung besteht zwischen Quell- und Zielanker?
 - Was ist die Bedeutung des Verweis?
- Semantische Information:
<p>Ich wohne in Berlin.</p>
- Schema zum gemeinsamen Verständnis ist nötig

Metadaten in HTML: Dublin Core

- "Dublin Core" (http://purl.org/metadata/dublin_core) ist der Versuch, ein verbreitetes Schema für Metadaten zu etablieren:
"The Dublin Core Metadata Initiative is an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models. DDCMI's activities include consensus-driven working groups, global workshops, conferences, standards liaison, and educational efforts to promote widespread acceptance of metadata standards and practices."
- Dublin Core Metadata Element Set, 1.1: Reference Description
 - Quelle: <http://dublincore.org/documents/1999/07/02/dces/>
 - Status: DCMI Recommendation 1999-07-02
- Dublin Core Metadata for Resource Discovery
 - Status: IETF RFC 2413, September 1998
- Encoding Dublin Core Metadata in HTML
 - Status: IETF RFC 2731, December 1999

Dublin Core Elemente

- **Title:** Titel des Dokuments
`<meta name="DC.Title" lang="es" content="La Mesa Verde y la Silla Roja" />`
- **Creator:** Erzeuger des Dokuments
`<meta name="DC.Creator" content="Gogh, Vincent van" />`
- **Contributor:** Jemand, der beigetragen hat
`<meta name="DC.Contributor" content="Curie, Marie">`
- **Publisher:** Der die Resource verfügbar macht
`<meta name="DC.Publisher" content="O'Reilly">`

[133] © Robert Tolksdorf, Berlin

Semantic Web

[134] © Robert Tolksdorf, Berlin

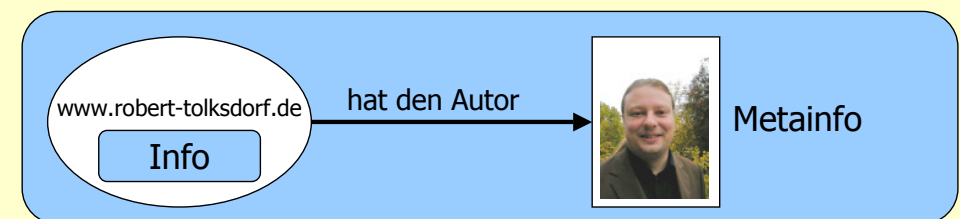
Semantic Web

- "The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"
[Tim Berners-Lee, James Hendler und Ora Lassila: The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, Scientific American, May 17, 2001]
- Explizite Repräsentation von Semantik mit Sprachen
- Genauer:
Weniger Missverständnisse wegen besserem Kontextbezug
- M2M vs. M2H Kommunikation

[135] © Robert Tolksdorf, Berlin

RDF Sätze

- Informationen und Metainformationen:



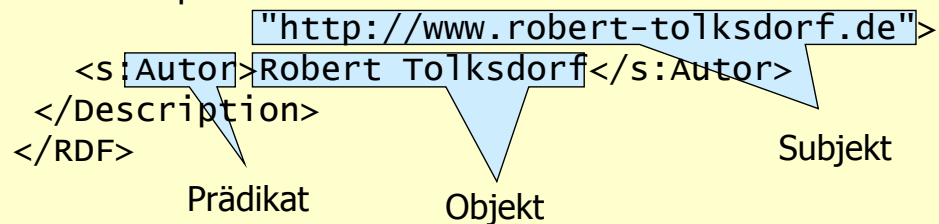
- In RDF als Satz ausgedrückt:

"www.robert-tolksdorf.de	Subjekt
hat den Autor	Prädikat
Robert Tolksdorf"	Objekt

[136] © Robert Tolksdorf, Berlin

In RDF definiert

- ```
<?xml version="1.0"?>
<RDF xmlns=
"http://www.w3.org/1999/02/22-rdf-syntax-ns"
xmlns:s="http://description.de/schema/">
<Description about=
```

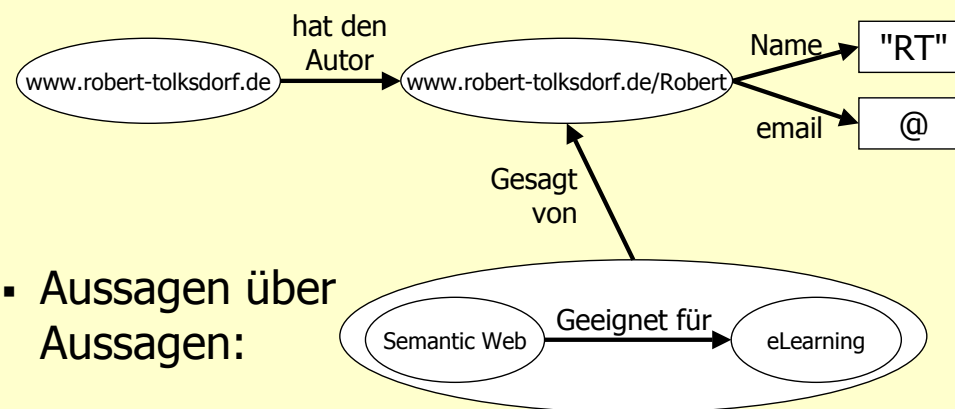


- Aus so explizit gemachten und maschinenverständlich repräsentierten Aussagen können Tools und Dienste inhaltliche Schlüsse ziehen

[137] © Robert Tolksdorf, Berlin

## Verweise auf Ressourcen als Objekte

- Objekte können selber auch Subjekte sein:



- Aussagen über Aussagen:

- Semantic Web: Geflecht aus getypten Beziehungen zwischen Konzepten

[138] © Robert Tolksdorf, Berlin

## RDF Schema

- Mit den grundlegenden RDF Mechanismen lassen sich einfache Aussagen auf vielfältige Weise treffen
- Mit RDF Schema werden einige Typen von Aussagen eingeführt, mit denen Schemas möglich werden, mit denen nützliche Modellierungsaussagen getroffen werden können:
  - "Jede Webseite hat einen Autor"
  - "Webseiten sind elektronische Dokumente"

[139] © Robert Tolksdorf, Berlin

## RDF Schema

- Elektronischen Dokumente bilden eine Klasse:

```
<rdf:Description rdf:ID="electronicDocument">
<rdf:type rdf:resource=
"http://www.w3.org/2000/01/rdf-schema#Class"/>
</rdf:Description>
```
- Web-Seiten sind elektronische Dokumente

```
<rdf:Description rdf:ID="webPage">
<rdf:type rdf:resource=
"http://www.w3.org/2000/01/rdf-schema#Class"/>
<rdfs:subclassOf rdf:resource="#electronicDocument"/>
</rdf:Description>
```
- Web-Seiten haben eine URL

```
<rdf:Property rdf:ID="URL">
<rdfs:domain rdf:resource="#webPage"/>
<rdfs:range rdf:resource=
"http://www.w3.org/2001/XMLSchema#string"/>
</rdf:Property>
```

[140] © Robert Tolksdorf, Berlin

## Nutzbarkeit von Metadaten

- Damit Metadaten nutzbar sind
  - muss der Informationsanbieter sich so ausdrücken, dass Informationsnutzer ihn verstehen
  - muss der Informationsnachfrage so fragen, dass er etwas finden kann
- Gemeinsame Benutzung von Konzepten
- Gemeinsame Sprache
- Ontologie zur Definition einer gemeinsamen Sprache

[141] © Robert Tolksdorf, Berlin

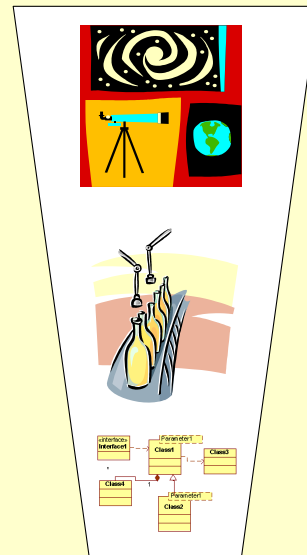
## Ontologie

- Ist Beschreibung einer Wissensdomäne mit
  - Standardisierter Terminologie (Klassen, Axiome etc) für Konzepte
  - Beziehungen zwischen Konzepten
  - Ableitungsregeln
- Ziel:  
Einschränkung der Interpretationsmöglichkeiten von Symbolen
- Dadurch: Gemeinsame Sprache
- Dadurch: Wissensaustausch möglich

[142] © Robert Tolksdorf, Berlin

## Arten von Ontologien

- Top Level Ontologien
  - Domänenüberschreitend
  - Allgemeine Konzepte
  - Person, Mensch, Tätigkeit, Artefakt
- Domain Ontologien
  - Auf bestimmten Bereich bezogen
  - Dozent, Veranstaltungsbesuch, Schein
- Als Ontologien verpackte Daten- und Klassenmodelle



[143] © Robert Tolksdorf, Berlin

## Web Ontology Language OWL

- OWL verfeinert und erweitert Modellierungsmöglichkeiten von RDF
- Es gibt abgesicherte und nicht abgesicherte Webseiten:

```
<owl:Class rdf:ID="protectedPage">
 <rdfs:subClassOf rdf:resource="#webPage"/>
</owl:Class>
```

```
<owl:Class rdf:ID="unprotectedPage">
 <rdfs:subClassOf rdf:resource="#webPage"/>
 <owl:disjointWith rdf:resource="#protectedPage"/>
</owl:Class>
```

[144] © Robert Tolksdorf, Berlin

## Klassenbildung

### ▪ Klasse Car:

```
<owl:Class rdf:ID="Car">
 <rdfs:comment>no car is a
 person</rdfs:comment>
```

### ▪ Ist Unterklasse einer anonymen Klasse und erfüllt auch deren Eigenschaften:

```
<rdfs:subClassOf>
 <owl:Class>
 <owl:complementOf rdf:resource="#Person"/>
 </owl:Class>
</rdfs:subClassOf>
</owl:Class>
```

[145] © Robert Tolksdorf, Berlin

## Beispiel GO

### ▪ Gene Ontology

- Ziel: "The goal of the Gene Ontology Consortium is to produce a dynamic controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing"
- Status: Datensammlung, Public Domain
- Quelle: <http://www.geneontology.org/>

### ▪ Originaldaten:

```
term: myosin
goid: GO:0016459
definition: A protein complex that functions as a molecular motor; uses the energy
of ATP hydrolysis to move actin filaments or to move vesicles or other cargo on
fixed actin filaments; has magnesium-ATPase activity and binds actin.
definition_reference: ISBN:96235764
```

```
term: myosin ATPase
goid: GO:0008570
definition: The hydrolysis of ATP by myosin that provides the energy for
actomyosin contraction.
definition_reference: NC-IUBMB:Proposed Changes to the Enzyme List concerning
ATPases and GTPases
```

[146] © Robert Tolksdorf, Berlin

## GO Oberes Modell

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE go:go>
<go:go xmlns:go="http://www.geneontology.org/xml-dtd/go.dtd#"
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
 <go:version timestamp="Wed May 9 23:55:02 2001" />

 <rdf:RDF>
 <go:term rdf:about="http://www.geneontology.org/go#GO:0003673">
 <go:accession>GO:0003673</go:accession>
 <go:name>Gene_Ontology</go:name>
 <go:definition></go:definition>
 </go:term>

 <go:term rdf:about="http://www.geneontology.org/go#GO:0003674">
 <go:accession>GO:0003674</go:accession>
 <go:name>molecular_function</go:name>
 <go:definition>The action characteristic of a gene
 product.</go:definition>
 <go:part-of rdf:resource="http://www.geneontology.org/go#GO:0003673" />
 <go:dbxref>
 <go:database_symbol>go</go:database_symbol>
 <go:reference>curators</go:reference>
 </go:dbxref>
 </go:term>
```

[147] © Robert Tolksdorf, Berlin

## GO Unteres Modell

```
<go:term rdf:about="http://www.geneontology.org/go#GO:0016209">
 <go:accession>GO:0016209</go:accession>
 <go:name>antioxidant</go:name>
 <go:definition></go:definition>
 <go:isa rdf:resource="http://www.geneontology.org/go#GO:0003674" />
 <go:association>
 <go:evidence evidence_code="ISS">
 <go:dbxref>
 <go:database_symbol>fb</go:database_symbol>
 <go:reference>fbrf0105495</go:reference>
 </go:dbxref>
 </go:evidence>
 <go:gene_product>
 <go:name>CG7217</go:name>
 <go:dbxref>
 <go:database_symbol>fb</go:database_symbol>
 <go:reference>FBgn0038570</go:reference>
 </go:dbxref>
 </go:gene_product>
 </go:association>
 ...
</go:term>
...
```

[148] © Robert Tolksdorf, Berlin

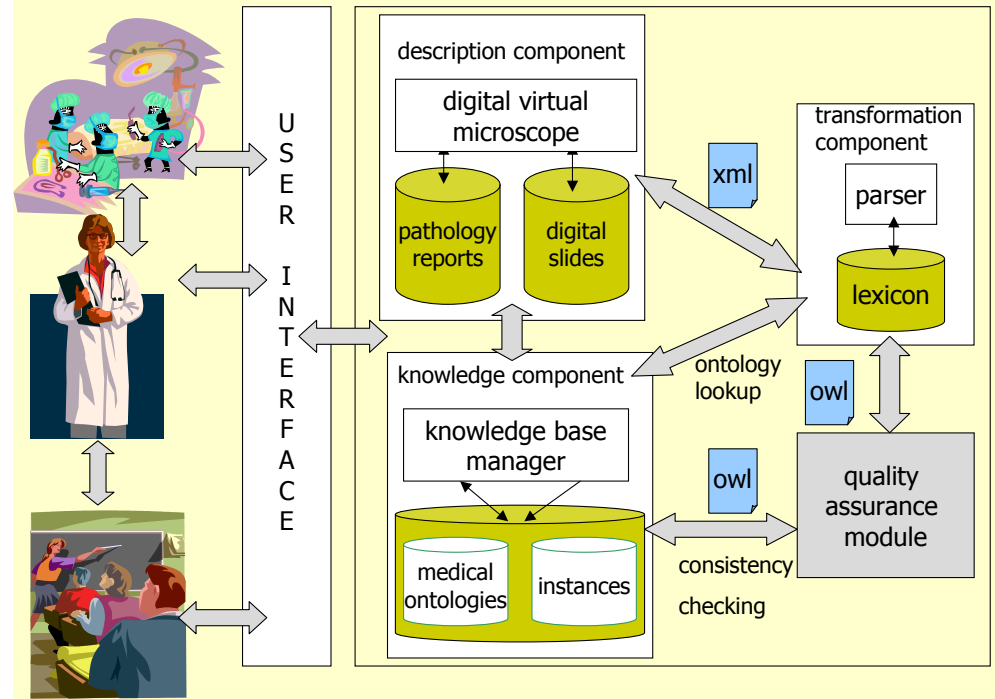
# A Semantic Web for Pathology

## Semantic Web based Information System for Lung Pathology



### Objectives

- re-organization of the available and future expert knowledge for efficient diagnostics and differential diagnostics tasks
  - reuse text and image data for case-based teaching materials
  - minimal invasive usage
  - integration in the available technical infrastructure
  - good precision values (under-diagnosed cases)
- 
- Improve retrieval capabilities of the pathology archive
    - pathology reports are textual representations of the digital images
    - the content of the text and image-based data is represented explicitly
    - use medical ontologies to refine search features
    - use rules to describe diagnostics processes
    - use ontology-based NLP algorithms to extract and represent the content of the pathology reports (semantic annotation)



```

<section><caption>Befund</caption>
<section><caption>Makroskopie</caption>
<paragraph><content>[1]Zwei Gewebszylinder von 15 und 4 mm Laenge[1].</content></paragraph>
</section>
<section><caption>Mikroskopie</caption>
...
<paragraph>
<content>[2]Stanzbiopsate aus Lungengewebe mit deutlicher Stoerung der alveolaren Textur, sowie noch nachweisbar deutlich Verbreiterung der Alveolarsepten, stellenweise Nachweis von Bronchiepithelregeneraten[2]. [3]Restliche Alveolarlumina z.T. durch Fibroblastenproliferate verlegt[3]. [4] Im Interstitium ein gemischt entzuendliches Infiltrat, bestehend aus Plasmazellen und Lymphozyten[4]. [5] Darunter relativ viele CD3-positive kleine und mittelgroesse T-Lymphozyten und CD68-positive Makrophagen[5].</content>
</paragraph>
<section><caption>Kritischer_Bericht</caption>
<paragraph>
<content>[6]Stanzbiopsate aus der Lunge mit Zeichen der organisierenden Pneumonie (klin.Mittellappen)[6].</content>
</paragraph>
<section><caption>Kommentar</caption>

```

## Transformation Component

**LF**

```

[1] card(x1, 2) AND cylinder(x1) AND length(x1, [15,
[2] unspec_plur_det(x2) AND punch_biopsat(x2)
AND from_rel(x2, x3) AND unspec_plur_det(x3)
AND lung_tissue(x3) AND with_rel(x3, x4)
AND def_det(x4) AND disturbance(x4, x5)
AND def_det(x5) AND texture(x5) AND alveolar(x5)
unspec_det(x6) AND extension(x6, x7)
AND def_det_plur(x7)
AND alveolar_septum(x7) AND unspec_det(x8)
AND evidence(x8, x9) AND indef_det(x9)
AND epithelial(x9) AND bronchial(x9)
AND regenerates(x9)
[3] def_det(x10) AND alveolarlumina(x10)
unspec_det_plur(x11) AND fibroblastial_proliferate(x11)
[4] def_det(x12) AND interstitium(x12)
indef_det(x13) AND inflammatory(x13) AND infiltrate(x13)
AND consisting_of_rel(x13, x14)
AND unspec_det_plur(x14)
AND konj(x14, x15, x16) AND plasma_cell(x15)
AND lymphocyte(x16)
[5] indef_det_plur(x17) AND konj(x17, x18, x19)
AND t_lymphocyte(x18)
AND cd68_positive(x19) AND macrophagus(x19)
[6] indef_det_plur(x20) AND punch_biopsate(x20)
AND from_rel(x20, x21) AND def_d
AND lung(x21) AND with_rel(x20, x22)
AND evidence(x22, x23) AND def_d
AND organising(x23) AND pneumonia(x23)

```

**OWL**

```

<Lung_Tissue rdf:ID="lung_tissue_x3">
 <partOf>
 <Lung_C0024109 rdf:ID="lung1">
 <hasSource rdf:resource="#UWDA"/>
 ... properties of the lung ...
 </Lung_C0024109>
 </partOf>
</Lung_Tissue>
<Punch_biopsat rdf:ID="punch_biopsat_x2">
 <from rdf:resource="#lung_tissue_x3"/>
</Punch_biopsat>
<alveola rdf:ID="alveola_x5">
 <hasTexture rdf:datatype="http://.../XMLSchema#string">
 <disturbed />hasTexture>
 <relatedTo rdf:resource="#lung1"/>
</alveola>
<Cylinder rdf:ID="cylinder_x1">
 <length rdf:datatype="http://www.w3.org/2001/
 XMLSchema#float">15.0</length>
 <formOf rdf:resource="#punch_biopsat_x2">
</Cylinder>
<Cylinder rdf:ID="cylinder_x2">
 <length rdf:datatype="http://www.w3.org/2001/
 XMLSchema#float">14.0</length>
 <from rdf:resource="#punch_biopsat_x2">

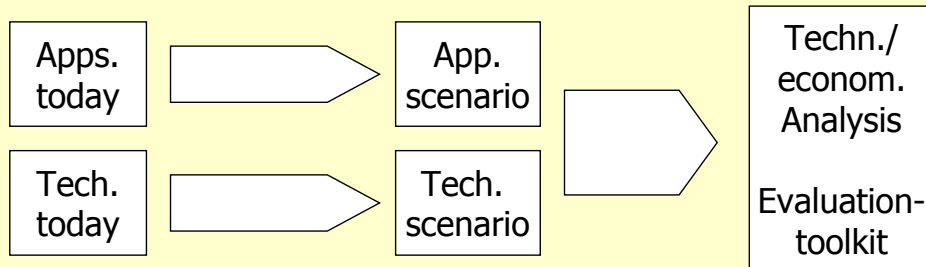
```

transformation component

## KnowledgeNets: Approach

### Business Viewpoint

Predict impacts on markets and value chains  
Estimate realization chances based on participant's interests and business models



### Technological Viewpoint

Derive requirements for the technological infrastructure  
Evaluate current technological developments

[153] © Robert Tolksdorf, Berlin

## Case 1: Economic Implications

- The usage of Semantic Web technologies would lead to an increased market transparency.
- Central Questions: Who would benefit? Is there a win-win situation for all participants?
  - ✓ Customers: Benefit from the increased market transparency.
  - ✗ Manufacturers: Face higher pressure of competition
  - ✗ Merchants: Face higher pressure of competition
- Realization chances depend on the characteristics of the market.
  - Advantage of being found might be more important
  - Call for tenders, labor market

[154] © Robert Tolksdorf, Berlin

## Case 2: Implications

- ✓ The use of concepts for describing jobs
  - makes the descriptions language independent
  - simplifies the global exchange of postings
- ✓ Transparency rises: all job postings accessible through all portals
- ✓ Employers would minimize their publishing costs
- ✓ The use of controlled vocabularies allows better automatic matchmaking between requirements and applications
- ✓ Better machine processability of semantically described postings
- ✗ Job portals would have to change their business models (fees)
- ✗ Additional effort for writing semantic descriptions
- ✗ Applicants have to add RDF descriptions to their application

[155] © Robert Tolksdorf, Berlin