



openHPI: **Web-Technologien** Einführung in das Semantic Web

Prof. Dr. Christoph Meinel

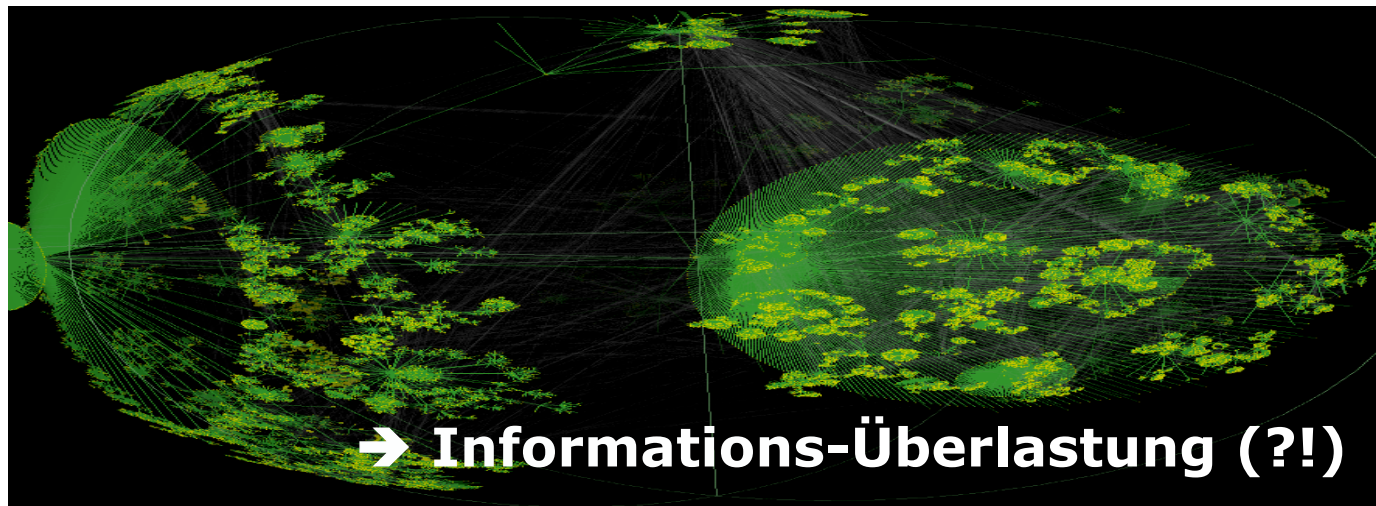
Hasso-Plattner-Institut

Universität Potsdam

Beschränkungen des World Wide Web (1/7)

Erinnerung: Das WWW erlaubt Zugriff auf enorme Menge an Informationen

- Es gibt Milliarden von WWW-Dokumenten – allerdings ist nur ein kleiner Teil von Suchmaschinen indexiert
- Außerdem können Milliarden von Dokumenten im *Deep Web* gefunden werden
- Die Menge der Dokumente im WWW verdoppelt sich alle 6 Monate



Beschränkungen des World Wide Web (2/7)

U.S. Calls on Ferguson to Overhaul Its Justice System
By MATT APUZZO and JOHN ELGON
A report examining the deadly shooting of an unarmed black teenager by a white police officer in Ferguson, Mo., said that the city had engaged in multiple constitutional violations.
177 Comments

Policing, Revenue and Discrimination Went Hand in Hand
By CAMPBELL ROBERTSON
Ferguson, a report finds, acts less like a municipality and more like a business, extracting money from poor blacks that it uses to sustain the city's budget.

Chinese Premier Wen Jiabao's Report

A Mark of Status for Clinton's Inner Circle
By AMY CHOICE and STEVE EDER 12:43 AM ET
A domain name set up for Hillary Rodham Clinton shielded her correspondence and was used for coveted accounts that conferred prestige and closeness to the secretary of state.

Jawbone's Discovery Fills Barren Evolutionary Period
By JOHN NOBLE WILFORD
The 2.8-million-year-old mandible of a Homo habilis predates by at least 400,000 years any previously known Homo fossils.

A Chilling Portrait of Ferguson
By THE EDITORIAL BOARD
The Missouri town must move far more aggressively to correct its dangerous problem of entrenched racism.

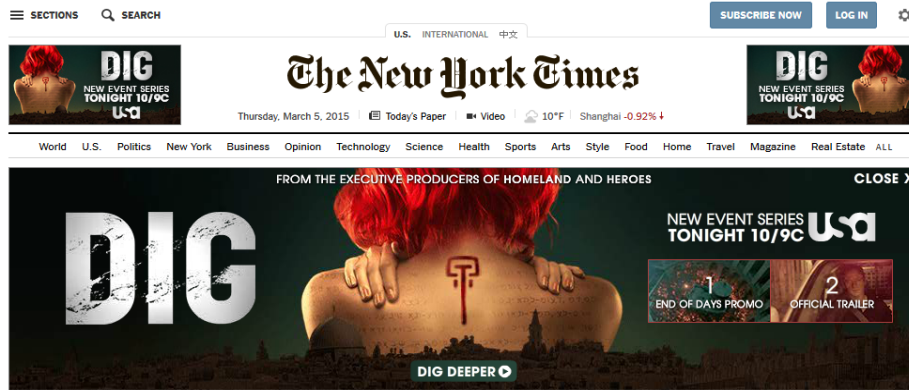
The Feds vs. Ferguson
The Ferguson Police Department and the municipal courts treated citizens like a revenue stream, violating their constitutional rights in the process.

Watching
4m The family of a Canadian pastor says it has received confirmation that he has been detained in North Korea after traveling there on a humanitarian trip.
2h A Singapore court sentenced two Germans to canings and jail time after they admitted spray-painting a train with graffiti.
3h The number of people worth \$30 million or more rose to nearly 173,000 worldwide last year, a study found.

Relevante Informationen?

- Was ist **wichtig**, was nicht?
- Was ist **Information**, was ist Werbung?
- Was ist **Bedeutung** der Information?
- Wie **glaubwürdig** ist die Information?
- Was gehört **zusammen**?
- Welche Teile sind eigentlich **überflüssig**?

Beschränkungen des World Wide Web (3/7)



U.S. Calls on Ferguson to Overhaul Its Justice System

By MATT APUZZO and JOHN ELSON
A report examining the deadly shooting of an unarmed black teenager by a white police officer in Ferguson, Mo., said that the city had engaged in multiple constitutional violations.

177 Comments

Policing, Revenue and Discrimination Went Hand in Hand

By CAMPBELL ROBERTSON
9:50 PM ET
Ferguson, a report finds, acts less like a municipality and more like a business, extracting money from poor blacks that it uses to sustain the city's budget.

RELATED COVERAGE
U.S. Clears Ferguson Officer in Inquiry 8:55 PM ET
Report: What Happened in Ferguson?

Chinese Premier Weighs Options to



Hillary Rodham Clinton on Tuesday in Washington. An aide set up clintonemail.com and registered it to her New York home. Jabin Botsford/The New York Times

A Mark of Status for Clinton's Inner Circle

By AMY CHOIRICE and STEVE EDER 12:43 AM ET
A domain name set up for Hillary Rodham Clinton shielded her correspondence and was used for coveted accounts that conferred prestige and closeness to the secretary of state.

RELATED COVERAGE

- Clinton Asks State Dept. to Release Emails 50 minutes ago
- Democrats Rally to Clinton's Defense After Email Disclosure
- Associated Press Threatens Legal Action Over FOIA Requests

Jawbone's Discovery Fills Barren Evolutionary Period

By JOHN NOBLE WILFORD
The 2.8-million-year-old mandible of a Homo habilis predates by at least 400,000 years any previously known Homo fossils.



A Chilling Portrait of Ferguson

By THE EDITORIAL BOARD
The Missouri town must move far more aggressively to correct its dangerous problem of entrenched racism.

MORE IN OPINION

- Editorial: The Fate of American Health Care
- Op-Ed: Shortchanging New Jersey by Billions
- Taking Note: High Rent, Low Wages

The Opinion Pages

CHARLES M. BLOW

The Feds vs. Ferguson

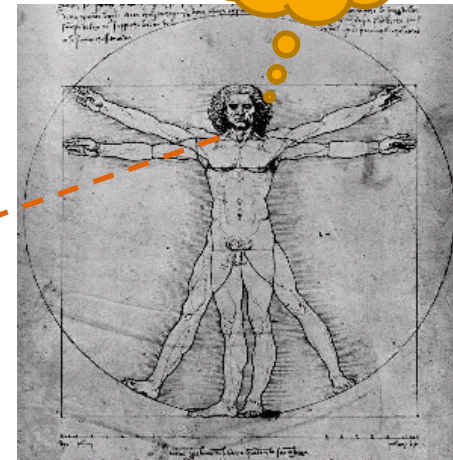
The Ferguson Police Department and the municipal courts treated citizens like a revenue stream, violating their constitutional rights in the process.

- Bruni: Hillary's Messy Habits
- Friedman: What Bibi Didn't Say
- Edsall: Establishment Populism Rising

Watching

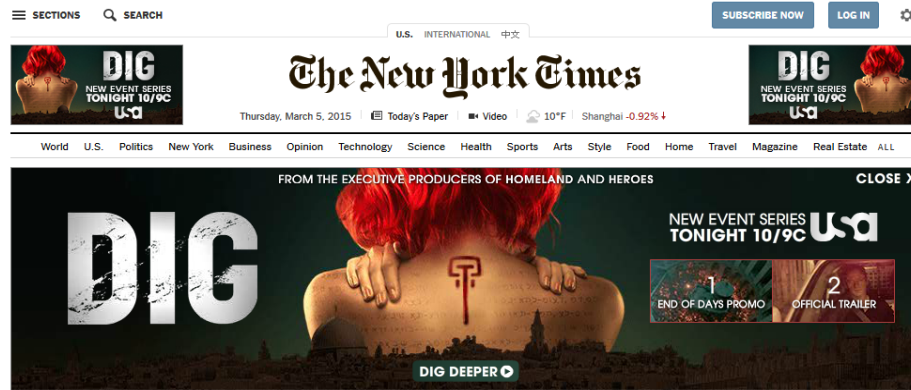
- 4m The family of a Canadian pastor says it has received confirmation that he has been **detained in North Korea** after traveling there on a humanitarian trip. CNN >
- 2h A Singapore court **sentenced two Germans to canings and jail time** after they admitted spray-painting a train with graffiti. Reuters >
- 3h The number of people worth \$30 million or more **rose to nearly 173,000 worldwide** last year, a study found. The Guardian >
- 3h 

Ist diese große Menge an Informationen eigentlich nützlich?



Menschliche Nutzer haben **kontextuelles Wissen / Allgemeinwissen** und können daher die (meisten) Informationen korrekt interpretieren

Beschränkungen des World Wide Web (4/7)



U.S. Calls on Ferguson to Overhaul Its Justice System

By MATT APUZZO and JOHN ELGON
A report examining the deadly shooting of an unarmed black teenager by a white police officer in Ferguson, Mo., said that the city had engaged in multiple constitutional violations.

177 Comments

Policing, Revenue and Discrimination Went Hand in Hand

By CAMPBELL ROBERTSON
9:50 PM ET
Ferguson, a report finds, acts less like a municipality and more like a business, extracting money from poor blacks that it uses to sustain the city's budget.

RELATED COVERAGE
U.S. Clears Ferguson Officer in Inquiry 8:55 PM ET
Report: What Happened in Ferguson?

Chinese Premier Wen Jiabao to



Hillary Rodham Clinton on Tuesday in Washington. An aide set up clintonemail.com and registered it to her New York home. Jabin Botsford/The New York Times

A Mark of Status for Clinton's Inner Circle

By AMY CHOIRICE and STEVE EDER 12:43 AM ET
A domain name set up for Hillary Rodham Clinton shielded her correspondence and was used for coveted accounts that conferred prestige and closeness to the secretary of state.

RELATED COVERAGE

- Clinton Asks State Dept. to Release Emails 50 minutes ago
- Democrats Rally to Clinton's Defense After Email Disclosure
- Associated Press Threatens Legal Action Over FOIA Requests

Jawbone's Discovery Fills Barren Evolutionary Period

By JOHN NOBLE WILFORD
The 2.8-million-year-old mandible of a Homo habilis predates by at least 400,000 years any previously known Homo fossils.



A Chilling Portrait of Ferguson

By THE EDITORIAL BOARD
The Missouri town must move far more aggressively to correct its dangerous problem of entrenched racism.

MORE IN OPINION

- Editorial: The Fate of American Health Care
- Op-Ed: Shortchanging New Jersey by Billions
- Taking Note: High Rent, Low Wages

The Opinion Pages

CHARLES M. BLOW

The Feds vs. Ferguson

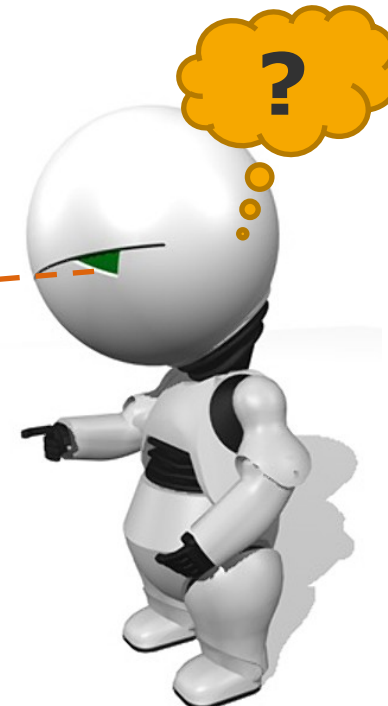
The Ferguson Police Department and the municipal courts treated citizens like a revenue stream, violating their constitutional rights in the process.

- Bruni: Hillary's Messy Habits
- Friedman: What Bibi Didn't Say
- Edsall: Establishment Populism Rising

Watching

- 4m The family of a Canadian pastor says it has received confirmation that he has been **detained in North Korea** after traveling there on a humanitarian trip. CNN >
- 2h A Singapore court **sentenced two Germans to canings** and jail time after they admitted spray-painting a train with graffiti. Reuters >
- 3h The number of people worth \$30 million or more **rose to nearly 173,000 worldwide** last year, a study found. The Guardian >
- 3h 

Ein **Roboter** (Programm) kann nicht zwischen wichtigen und unbedeutenden Informationen unterscheiden, da es **Bedeutung nicht versteht**



Beschränkungen des World Wide Web (5/7)

The screenshot shows the front page of The New York Times. At the top, there's a navigation bar with 'SECTIONS', 'SEARCH', and 'SUBSCRIBE NOW'. Below this, the main header features the newspaper's name and the date 'Thursday, March 5, 2015'. A large advertisement for the TV series 'DIG' is prominently displayed, featuring a woman's back and the text 'NEW EVENT SERIES TONIGHT 10/9C'. Below the ad, there are several article teasers. On the left, 'U.S. Calls on Ferguson to Overhaul Its Justice System' by Matt Apuzzo and John Elgon. In the center, 'A Mark of Status for Clinton's Inner Circle' by Amy Choick and Steve Eder. On the right, 'The Opinion Pages' section with articles like 'A Chilling Portrait of Ferguson' and 'The Feds vs. Ferguson'. At the bottom, there's a 'Watching' section with a video thumbnail and a 'Jawbone's Discovery Fills Barren Evolutionary Period' article.

Roboter braucht Informationen über Bedeutung eines Dokuments und seiner verschiedenen Teile

→ Metadaten

Werbung



Allerdings bietet HTML nur sehr begrenzte Möglichkeiten, Metadaten zu beschreiben

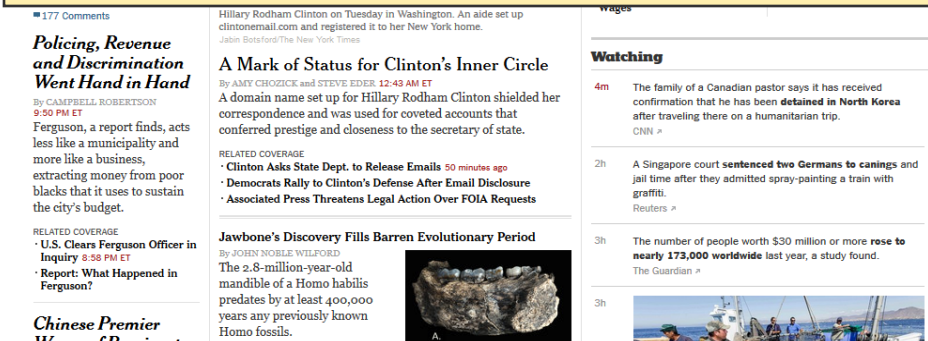
Beschränkungen des World Wide Web (5/7)



Roboter braucht Informationen

```
<meta charset="UTF-8" />
<meta http-equiv="cache-control" content="no-cache" />
<meta http-equiv="content-language" content="en-US" />
<meta http-equiv="content-type" content="text/html; />
<meta name="author" content="Jane Doe" />
<meta name="description" content="...summary of web page..." />
<meta name="robots" content="all" />
<meta name="keywords" content="meta tags meaning" />
```

Alle Metadaten müssen vom Autor der Webseite erstellt werden



Allerdings bietet HTML nur sehr begrenzte Möglichkeiten Metadaten zu beschreiben

Beschränkungen des World Wide Web (7/7)

- WWW wurde ursprünglich **für Menschen entwickelt**
- Technologie hinter dem WWW ist Markup-Sprache HTML
- HTML und CSS beschreiben
 - wie Informationen **strukturiert** und **dargestellt** sind
 - wie Informationen **verlinkt** werden können
 - aber **nicht Bedeutung** der Informationen



Semantik?

Semantik

Semantik ist Teilgebiet der **Linguistik**, dass sich mit

- **Sinn** und
- **Bedeutung**

von **Sprachen** und **linguistischen Symbolen** beschäftigt

Semantik versucht Frage zu beantworten, wie

- Sinn und Bedeutung von komplexen Begriffen abgeleitet werden können von Sinn und Bedeutung einfacher Begriffe

Semantik baut auf **Syntax** auf

- **Kommunikation** benötigt ein allgemeines Verständnis von **Syntax** und **Semantik** der ausgetauschten Sprachsymbole und Zeichenketten

→ Gegenseitiges **Verständnis** ist nur möglich, wenn Sender und Empfänger die **identische Semantik** vereinbart haben

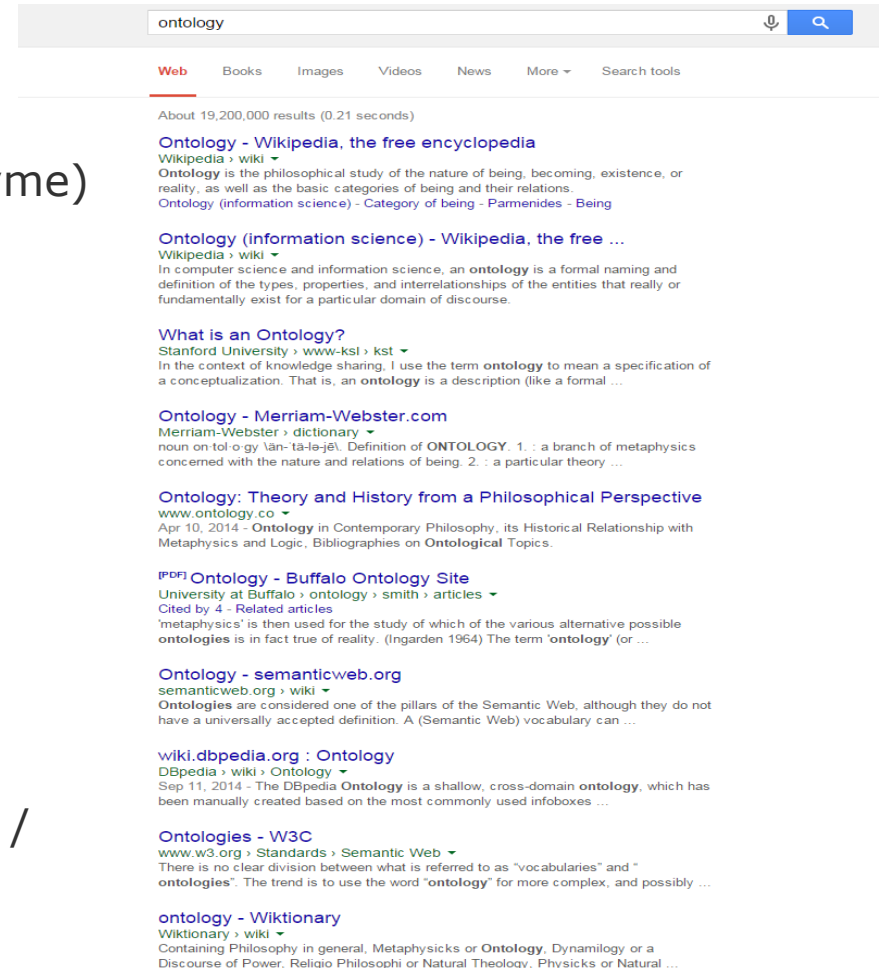
Problembereich 1: Suche nach Informationen

Schlüsselwort-basierte Suche

- Ruft viele **irrelevante** Ergebnisse ab
 - verschiedene Bedeutungen (Homonyme)
 - anderer Kontext
- Kann nicht **alle relevanten** Ergebnisse finden
 - Synonyme
 - Mangel an präzisiertem Kontext

Anforderung:

- ➔ Einführung einer formalen Terminologie / **Ontologie**



Problembereich 1: Suche nach Informationen

Schlüsselwort-basierte Suche

- Ruft viele **irrelevante** Ergebnisse ab
 - Verschiedene Bedeutungen (Homonyme)
 - Anderer Kontext
- Kann nicht die gewünschte Information finden
 - Synonym
 - Mangelt an Kontext

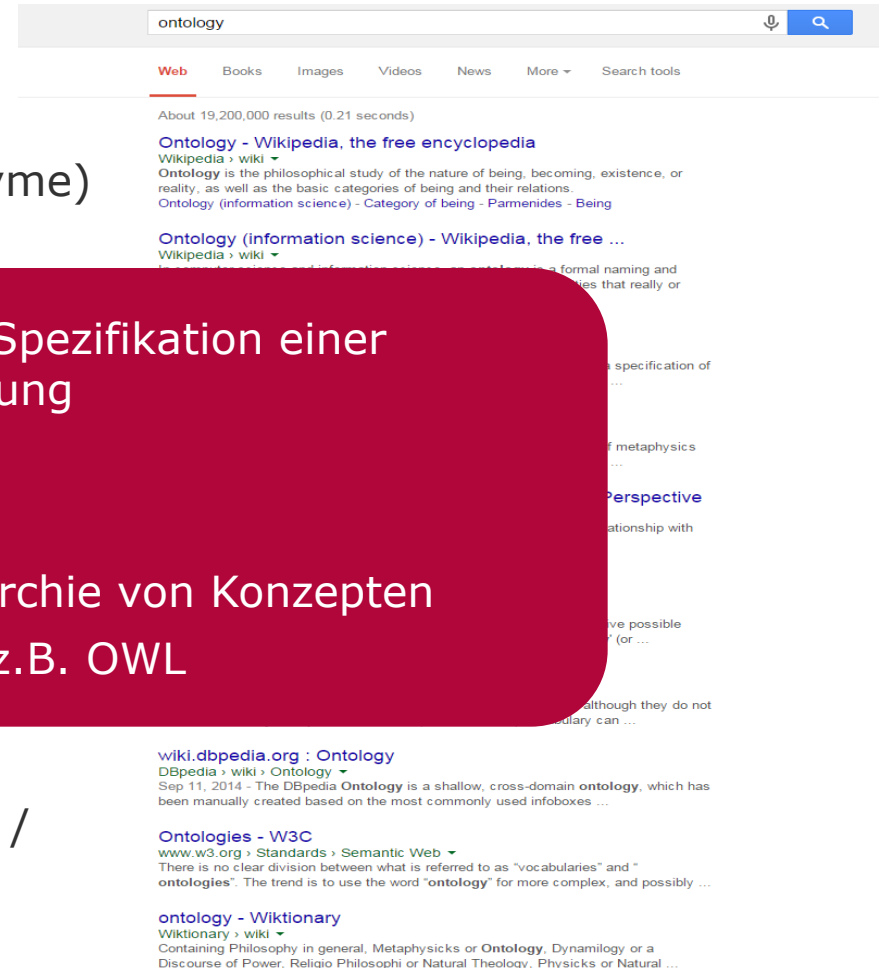
Ontologie: Explizite formale Spezifikation einer gemeinsamen Konzeptualisierung

Eine Ontologie besteht aus:

- **Taxonomie**, d.h. eine Hierarchie von Konzepten
- sprachlichen Beschreibung, z.B. OWL

Anforderung:

- Einführung einer formalen Terminologie /
Ontologie

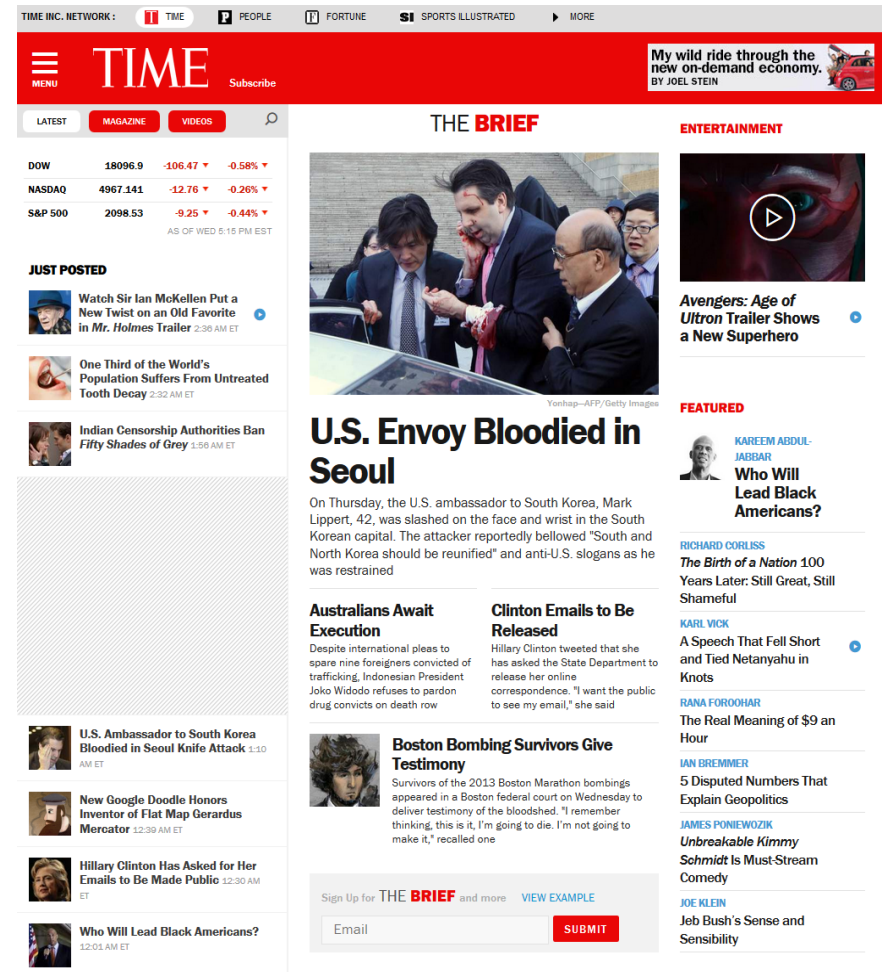


Problembereich 2: Informationsextraktion

- Nur Menschen können Informationen korrekt extrahieren
 - Software (Roboter) fehlt
 - kontextuelles Wissen
 - Allgemeinwissen
- wird benötigt, um Informationen aus ihrer Text- oder Bild-basierten Darstellung zu extrahieren

Anforderung:

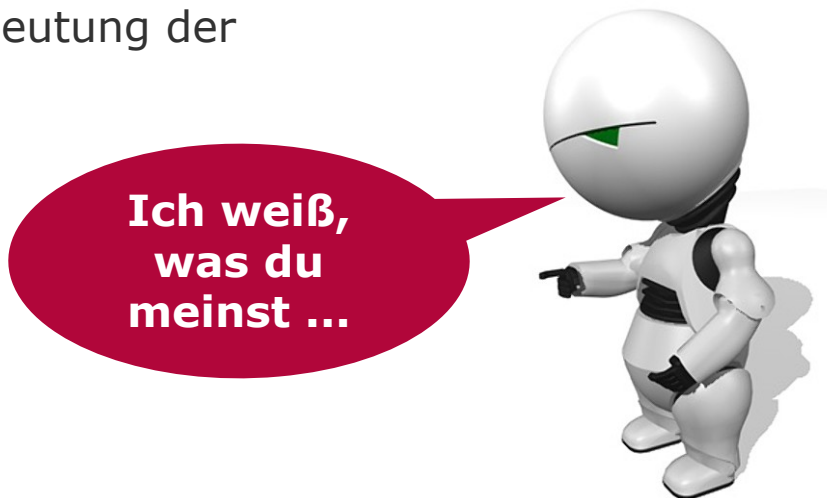
- **Aggregation und Integration von Informationen** aus verschiedenen Quellen



Vision für das Semantic Web

Tim Berners-Lee – Erfinder des WWW – im September 1998:

- „Das Web war als Informationsraum entworfen worden, mit dem Ziel nicht nur für Mensch-zu-Mensch Kommunikation nützlich zu sein, sondern auch, dass es Maschinen möglich ist, sich zu beteiligen und zu helfen.“
- **HTML** als Sprache zur Strukturierung von Informationen im WWW fehlt es an Möglichkeiten, die Bedeutung der Informationen auszudrücken



Ich weiß,
was du
meinst ...

Potenziale des Semantic Web (1/3)

Beispiel 1: Suchmaschinen heute

Suchanfrage: „Golf von 2015“

Nutzer muss Anfrage verfeinern, da Ergebnisse mehrdeutig sind

- Anfrage zu „Golf“:
 - Auto oder Sport oder sonstiges? → Verfeinerung: Auto
- Anfrage zu „2015“:
 - Jahr der Konstruktion oder technischer Parameter?
→ Verfeinerung: Jahr der Konstruktion
- Anfrage zu „Golf von 2015“:
 - Verkaufsangebot, Handbuch oder sonstiges?
→ Verfeinerung: Handbuch
- ...

Suchmaschine muss Nutzer nach Bedeutung der Anfrage fragen

Potenziale des Semantic Web (2/3)

Beispiel 1: Suchmaschinen von morgen

Suchanfrage: „Golf von 2015“

Suchmaschine „kennt“ den Nutzer

- Nutzer überprüft in letzter Zeit regelmäßig Ergebnisse von professionellen Golf-Turnieren
- Nutzer hat online im letzten Jahr Golfschläger gekauft
- Nutzer hat News-Gruppe „en.rec.golf.balls“ abonniert
- Nutzer hat Bildschirmhintergrund „PGA Tour 2014“ gewählt
- Nutzer besitzt die Domain www.thegolffanatic.com
- ...

Suchmaschine kann anhand dieses Wissens Bedeutung der Anfrage erschließen und richtig beantworten ...

Potenziale des Semantic Web (3/3)

Beispiel 2: Planung eines Familienurlaubs

Programm-Eingabe: Urlaub am Meer, ≥ 10 Tage, Familie, ≤ 2000 €

- Programm überprüft Kalender aller Familienmitglieder und sucht nach freien Zeiträumen
- Programm überprüft Urlaubsangebote für Zeiträume anhand der Anforderungen
 - geografische Parameter
 - familienfreundlich?
 - Golfplatz in der Nähe?
 - ...
- Programm überprüft freie Kapazitäten und stellt Auswahl aus passenden Angeboten zusammen
- Programm organisiert Transport anhand des ausgewählten Ortes

Ziele des Semantic Web

- Automatische Organisation von Wissen in der jeweiligen Domäne
- Bereitstellung automatisierter Werkzeuge für Wartung, Entfernung von Inkonsistenzen oder Extraktion von neuem Wissen
- Ersetzung der Schlüsselwort-basierten Suche durch Inhalts-basierte Suche mit „intelligenten“ Frage- und Antwort-Zyklen
- Nutzerfreundliche Suche
- Extraktion und Darstellung von Wissen
- Beantwortung von Fragen zu über verschiedenen Dokumente verteilte Informationen
- Exakte Bestimmung, wer auf welche Teile einer bestimmten Information zugreifen kann

Technische Realisierung des Semantic Web (1/2)

Realisierung des Semantic Web benötigt Anzahl von Technologien für jeweils verschiedene Zwecke:

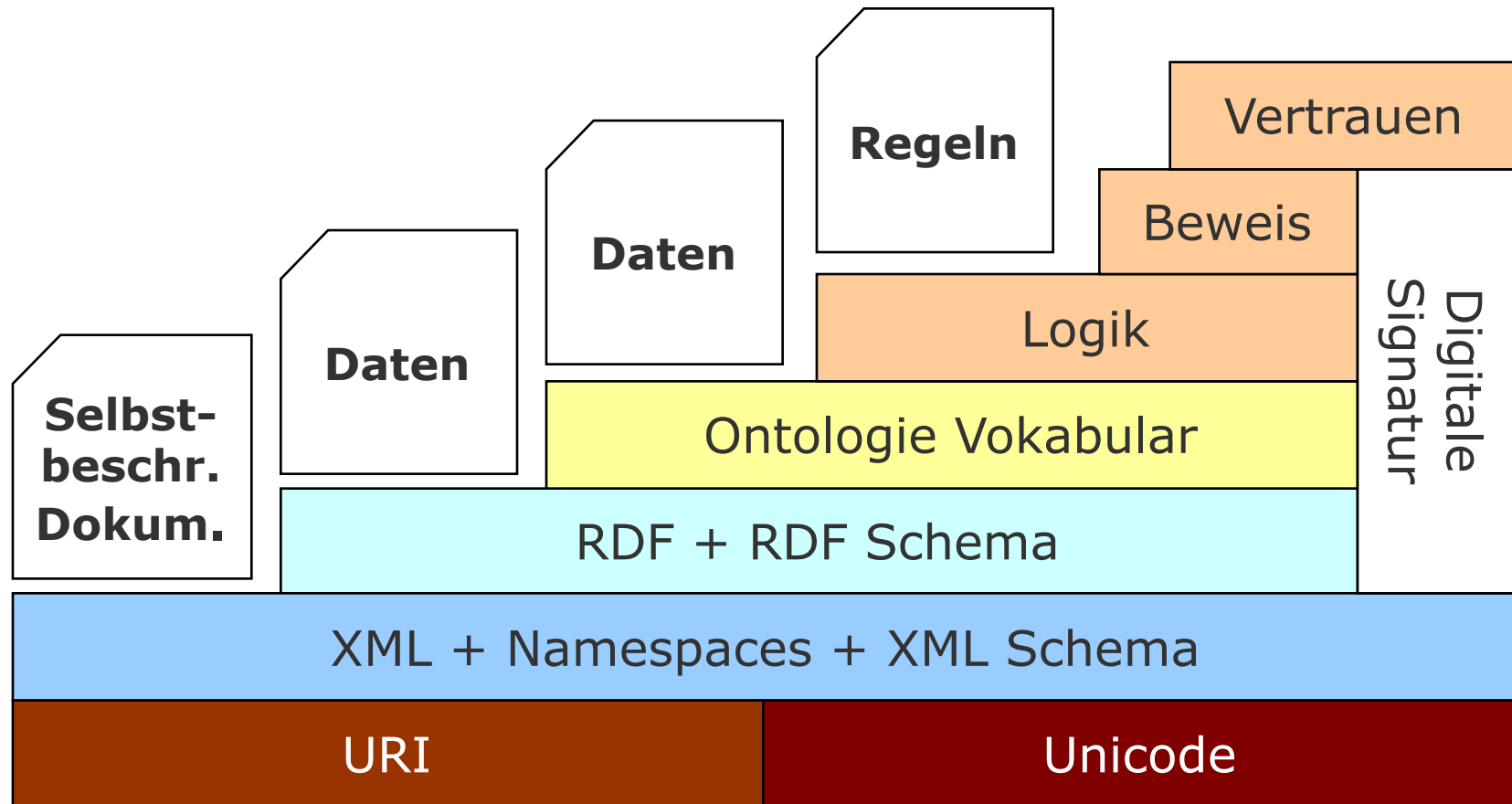
- Jede Informations-Quelle im Semantic Web benötigt eine eindeutige Kennung
→ **URI**
- Informationen im Semantic Web müssen formal ausgedrückt werden, um automatisiert verarbeitet werden zu können; dazu wird einheitliche Syntax gebraucht
→ **XML, XMLSchema**
- Einfache semantische Beziehungen zwischen Informations-Entitäten müssen ausgedrückt werden können
→ **RDF (Resource Description Framework), RDFSchema**
- ...

Technische Realisierung des Semantic Web (2/2)

Realisierung des Semantic Web benötigt eine Anzahl von **Technologien** für verschiedene Zwecke:

- ...
- Verschiedene Semantiken müssen beschreib- und erklärbar sein, um auch scheinbar inkonsistente Informationen kombinieren zu können
 - ➔ **Ontologien**
- Neue Information (Wissen) muss aus vorhandenen Informationen abgeleitet werden können
 - ➔ **Inferenz-Mechanismen**
- Datenschutz (in Bezug auf Vertraulichkeit und Integrität) muss gewährleistet werden können
 - ➔ **XMLEncryption and XMLSignature**

Semantic Web – Technology Stack



Weitergehende Informationen:

➔ openHPI Kurse zu **Semantic Web Technologien** von Prof. Dr. Harald Sack

Linked Open Data (1/2)

- Weltweites Netzwerk im Web frei verfügbarer Daten, die automatisiert von Programmen verarbeitet werden können
- Werden per URI und HTTP abgefragt
- Meist per RDF beschrieben und verlinkt, semantisch abfragbar per SPARQL
- Damit auch Teil des Semantic Web
- Bekannte Datensätze:
 - DBpedia: Extrahierte Informationen aus Wikipedia
 - FOAF: Datensatz über Personen und ihre Beziehungen
 - GeoNames: Informationen über Orte und deren Position
 - ...

Linked Open Data (2/2)

