



Vorlesung XML-Technologien – SoSe 2012

Prof. Dr.-Ing. Robert Tolksdorf
& Markus Luczak-Rösch
Freie Universität Berlin
Institut für Informatik
Netzbasierte Informationssysteme

tolk@ag-nbi.de
markus.luczak-roesch@fu-berlin.de



Einführung

Markus Luczak-Rösch
Freie Universität Berlin
Institut für Informatik
Netzbasierte Informationssysteme

markus.luczak-roesch@fu-berlin.de

Heutiger Termin

- Wie ist diese Vorlesung aufgebaut?
- Warum sollte Sie diese Vorlesung interessieren?
- Was ist XML?
- Ist XML noch aktuell?
- Wie können Sie Selbststudium betreiben?



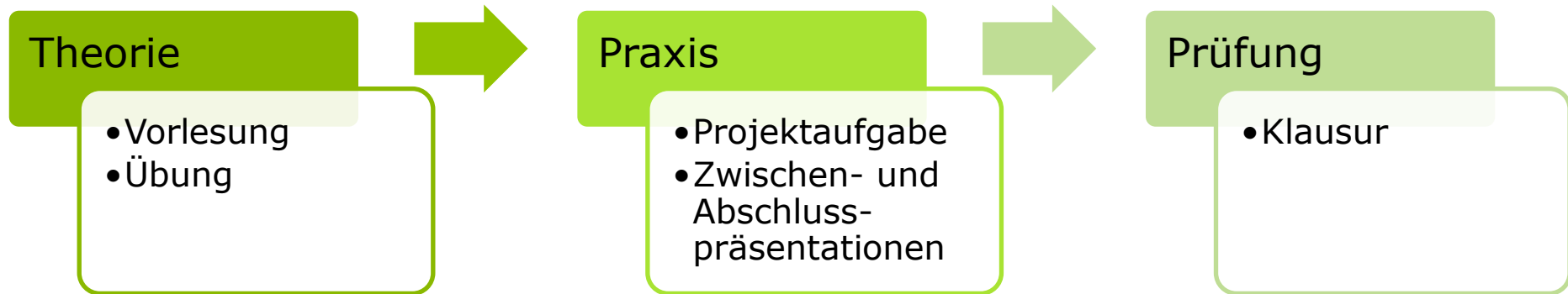
Organisatorisches

Veranstalter

- Vorlesung:
Prof. Dr.-Ing. Robert Tolksdorf, tolk@ag-nbi.de
Markus Luczak-Rösch, markus.luczak-roesch@fu-berlin.de
- AG Netzbasierte Informationssysteme
- Büro: Königin-Luise-Str. 24-26, 1.OG, Raum 118
(NICHT Takustr. 9)
- Sprechstunde:
 - Termine per Mail abstimmen
 - oder bei Herrn Tolksdorf via Form anmelden
<http://flp.cs.tu-berlin.de/%7Etolk/sprechstunde>

- <http://blog.ag-nbi.de/2012/03/19/vorlesung-xml-technologien-2/>
- hier finden sich
 - Folien der Vorlesungen
 - Termine der Vorlesungen, Übungen und Projektarbeit
 - Hinweise/Links auf Literatur

- >60 Teilnehmer haben sich im Online-KVV angemeldet (Stand: 04.04.2012)
- Master- und Bachelor-Studierende:
 - zusätzlich verbindliche Anmeldung mit Unterschrift notwendig
 - Ohne diese Anmeldung dürfen keine Leistungen erbracht werden.
 - verbindliche Anmeldung für Msc-Studierende in der nächsten Woche



- Ab 24.04. bieten wir ein Tutorium an (Di. zur Vorlesungszeit)
 - Behandlung von Vorlesungsstoff in Übungsaufgaben
 - Übungsaufgaben sind **fakultativ**
 - Präsentation einer Musterlösung durch Tutor
 - Beantwortung von Detailfragen

- **verpflichtende Projektarbeit** in Gruppen zu 6 Personen
- Was erwartet Sie?
 - Wir präsentieren zu Beginn der Praxisphase eine oder zwei Projektaufgaben und bilden Projektgruppen, die diese bearbeiten
 - Präsenztermine zur Vorlesungszeit sind Betreuungstermine
 - Di.: indiv. Coaching in Sprechstunde bei Markus Luczak-Rösch (jede Gruppe erhält festen Zeitslot)
 - Mi.: 10 Minuten Stand Up je Gruppe im Hörsaal
 - Meilensteinpräsentation zum Fortschritt in der Mitte der Projektarbeitsphase (alle Gruppen)
 - Abschlusspräsentation am Ende des Semesters (alle Gruppen)

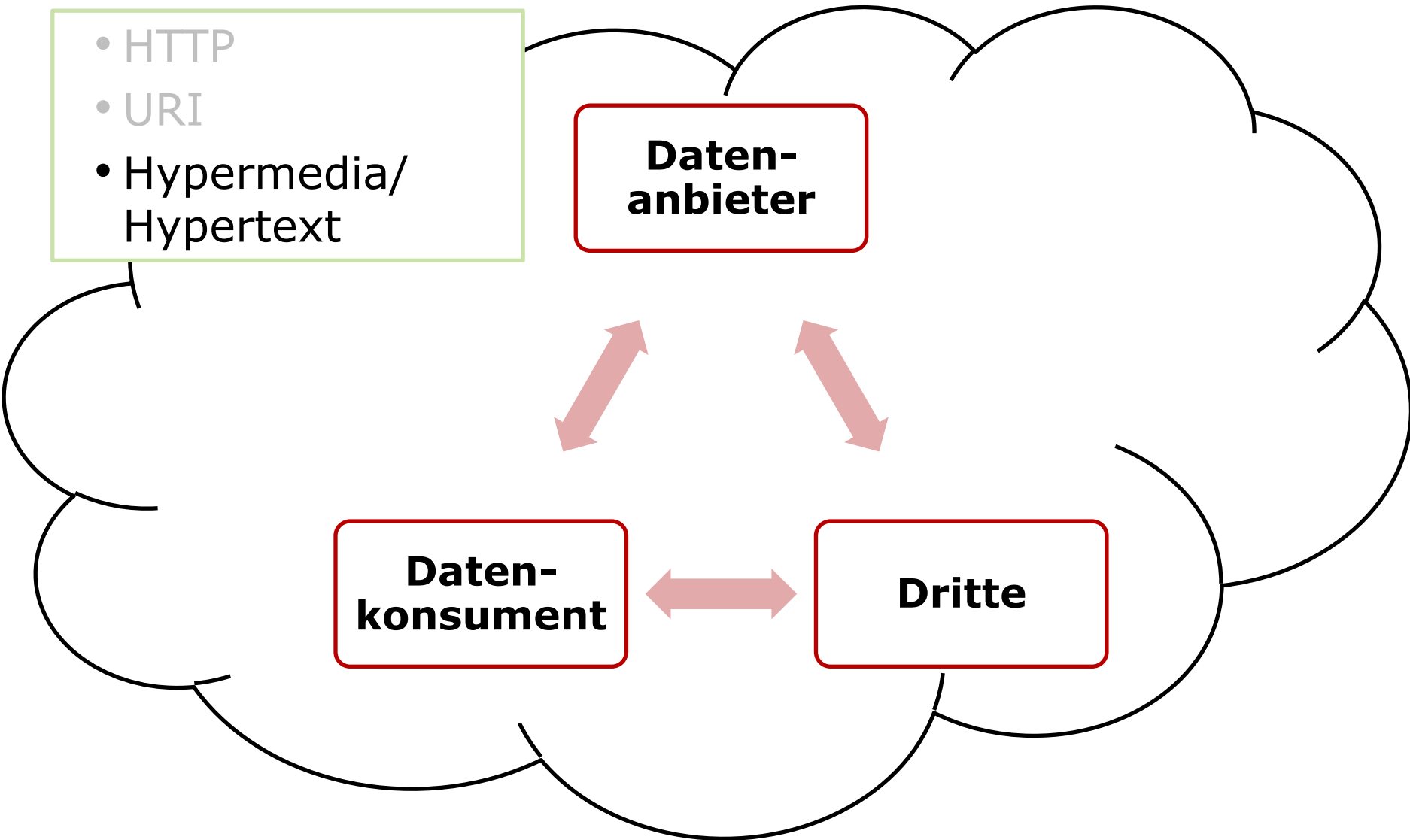
- Klausur bzw. Nachklausur erfolgreich bestanden
 - Teilnahmevoraussetzung: Anmeldung
 - Klausurtermin: 11.07.2012 (letzter Vorlesungstermin)
 - Termin für Nachklausur wird noch bekannt gegeben
- Projektarbeit → aktive Teilnahme
- Note = Klausurnote

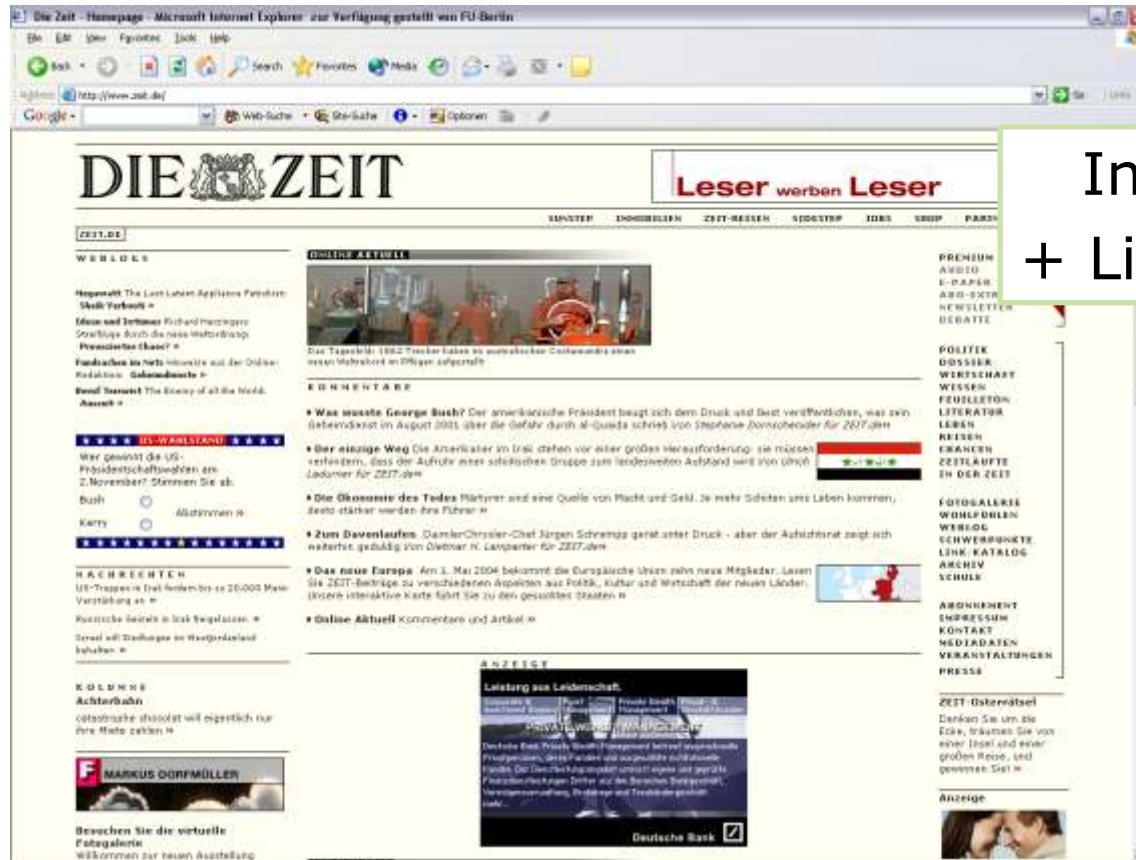
Warum beim Projekt anstrengen?

- Sie lernen dabei vermutlich am meisten!
- Sie müssen im Plenum Ihre Arbeit präsentieren!
- Wir erkennen bei ungenügendem Arbeitsergebnis die aktive Teilnahme nicht an!



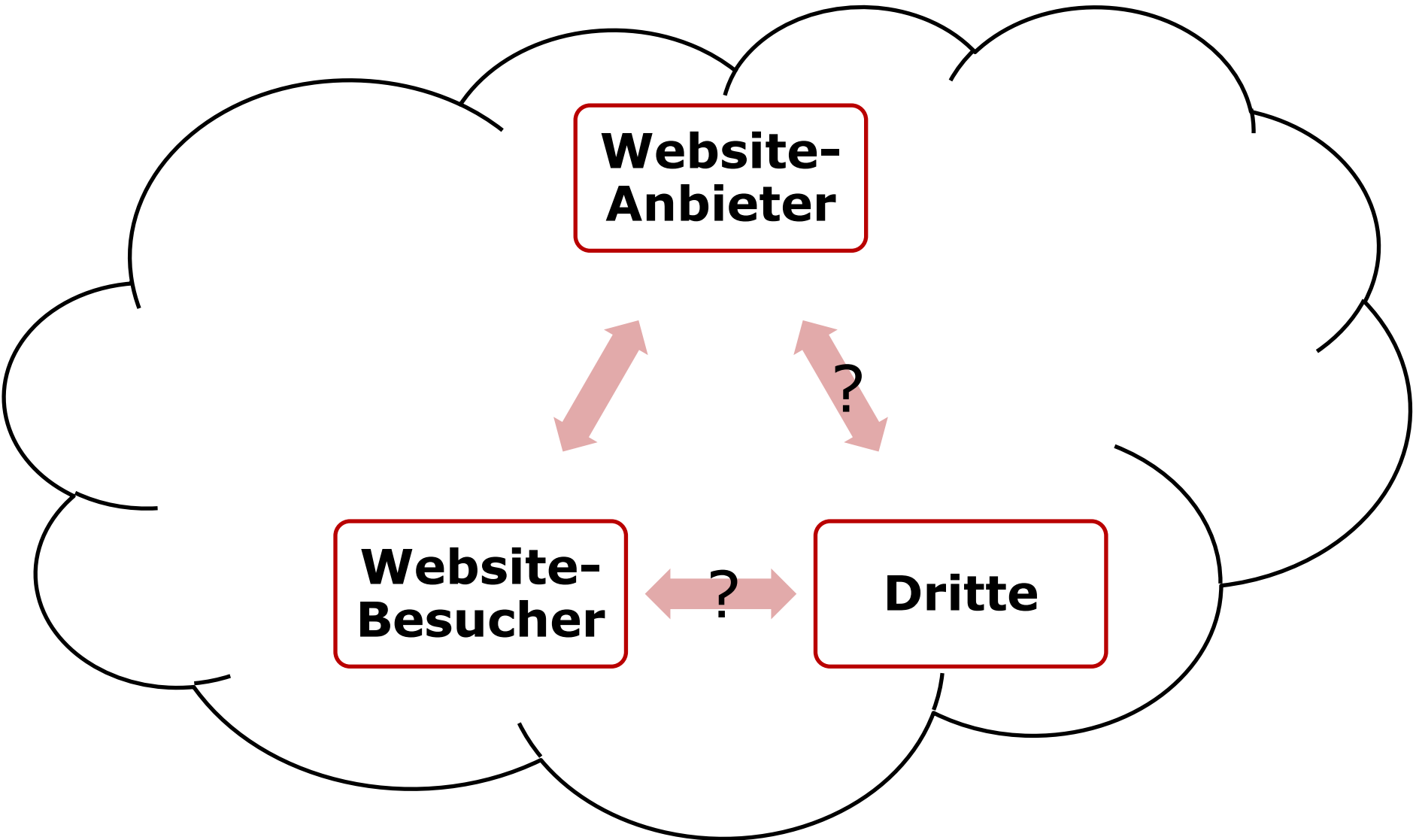
Was bringt eine Vorlesung XML-Technologien?





Inhalt
+ Links

HTML hat sich für die Präsentation von Inhalten bewährt.

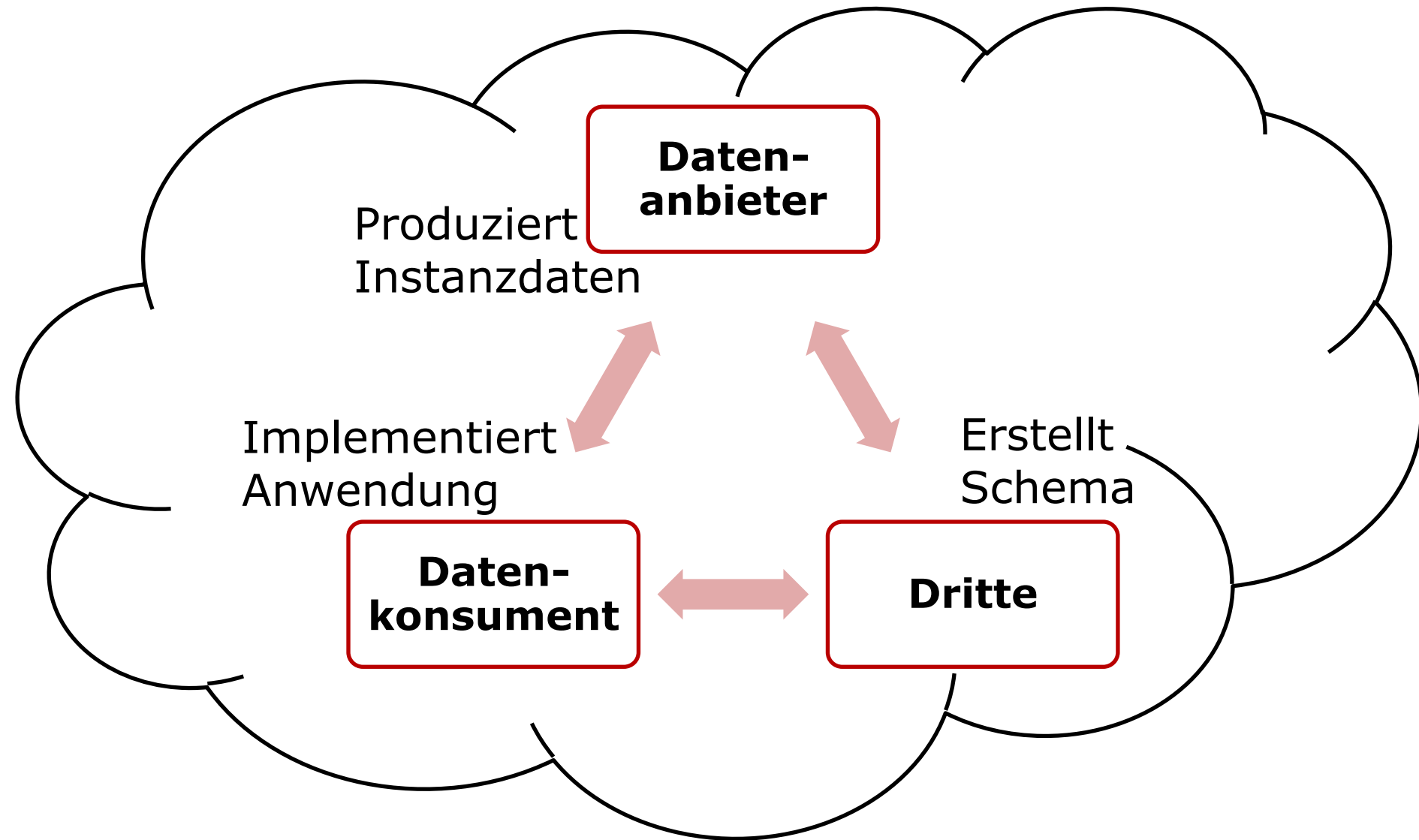


immer häufiger medienneutrale Darstellung nötig:

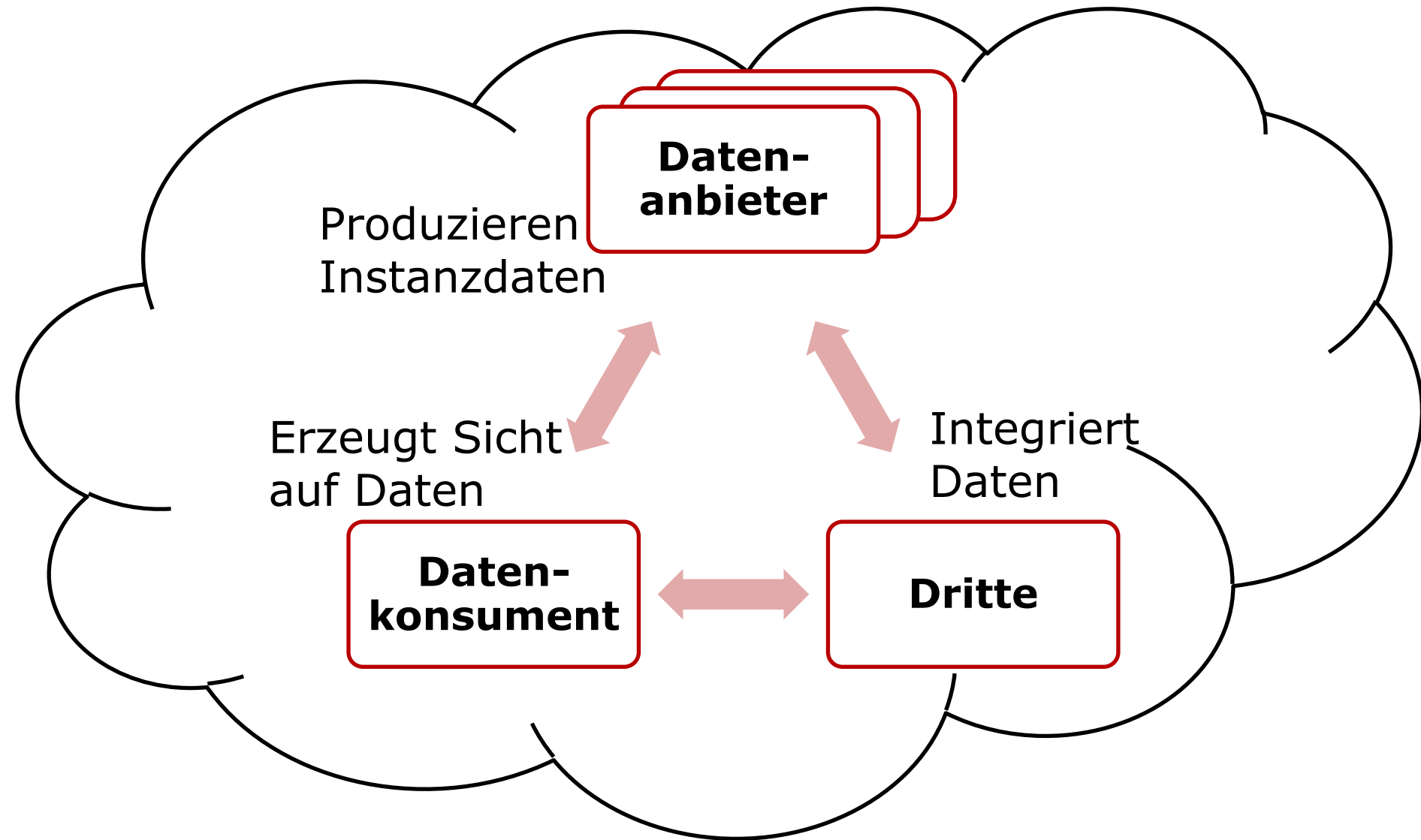
- Vielfalt von Endgeräten (und Bandbreiten) macht Trennung Inhalt von Präsentation nötig
- Austausch von Daten und Dokumenten zwischen Computern
 - ⇒ z.B. Übermittlung eines Bestellformulars
 - ⇒ z.B. Web Services

HTML: keine layoutunabhängige
Darstellung von Inhalten

XML ist die Basis für Web-Interoperabilität



XML ist die Basis für Web-Interoperabilität



- HTTP
- URI
- Hypermedia/
Hypertext
- Metadaten

**Daten-
anbieter**

**Daten-
konsument**

Dritte





Human – Machine

Shopping-Ergebnisse für **galaxy tab**

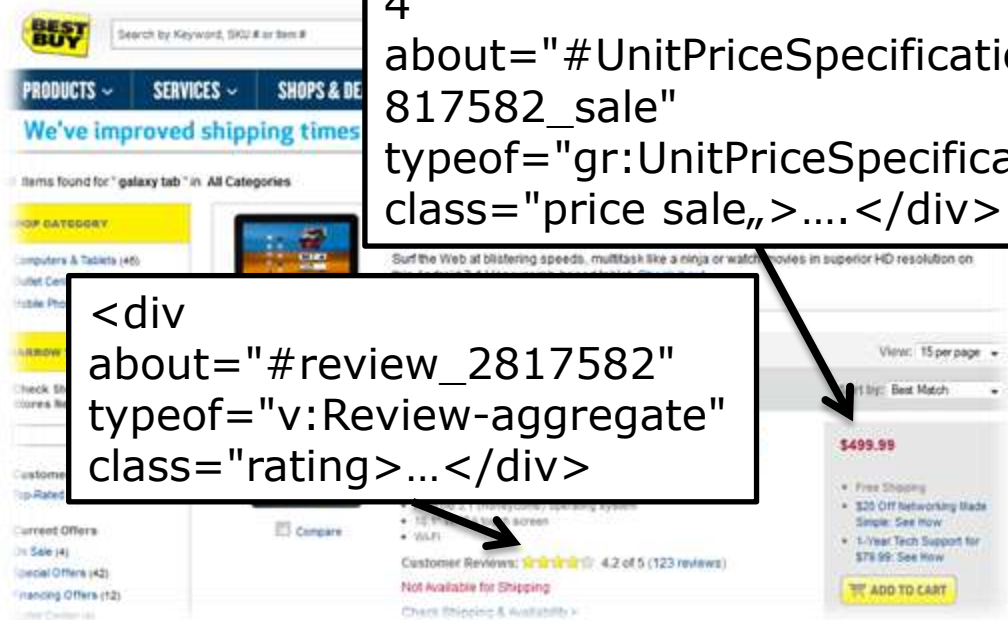


- [Samsung Galaxy Tab 10.1 16 GB - Android 3.0 \(Honeycomb ...](#)
★★★★★ 10 Erfahrungsberichte - 354 € - 59 Anbieter
- [Samsung Galaxy Tab WiFi 16 GB - Android 2.2 1 GHz](#)
★★★★★ 18 Erfahrungsberichte - 279 € - 100 Anbieter
- [Samsung Galaxy Tab 16 GB - Android 2.2 1 GHz](#)
★★★★★ 33 Erfahrungsberichte - 222 € - 94 Anbieter

```
<div
rel="gr:hasPriceSpecification"><h
4
about="#UnitPriceSpecification_2
817582_sale"
typeof="gr:UnitPriceSpecification"
class="price sale">....</div>
```

```
<div
about="#review_2817582"
typeof="v:Review-aggregate"
class="rating">...</div>
```

Machine – Machine

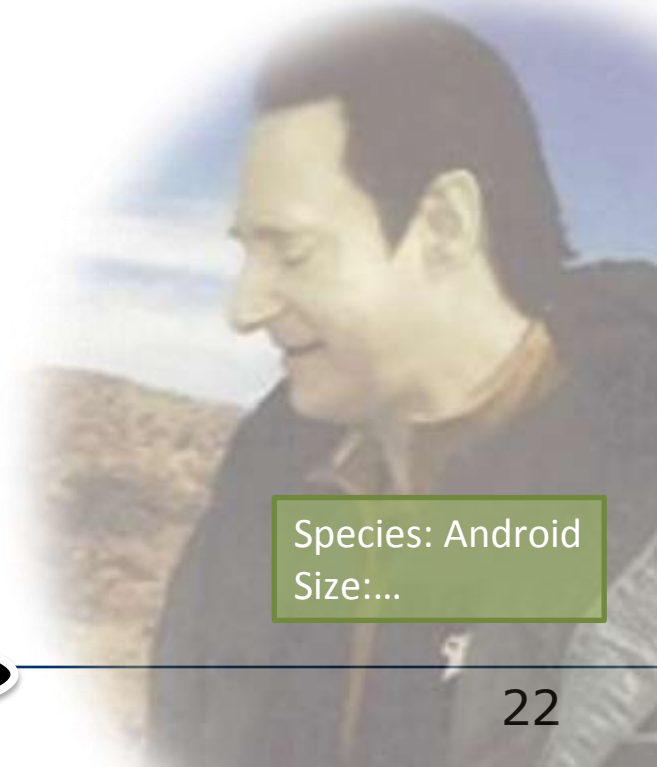
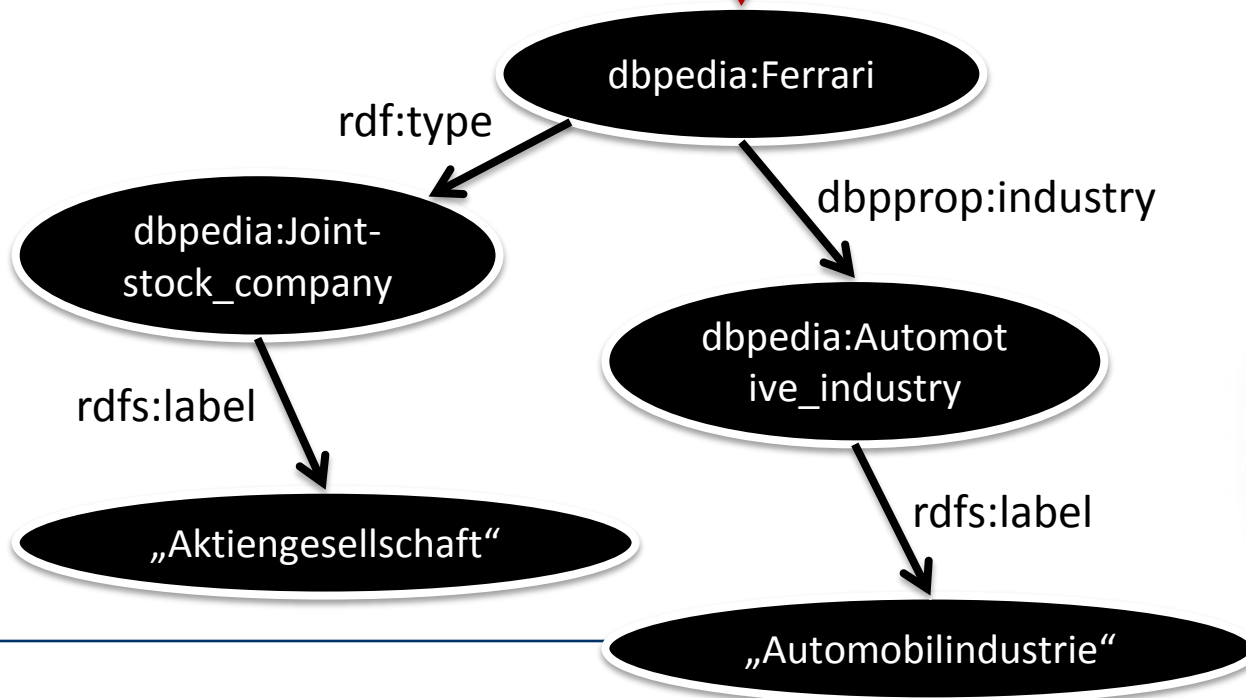


„Machinereadable data about data“



```
<span  
  resource=„dbpedia:Ferrari“  
  property=„rdf:label“>  
  法拉利汽車  
</span>
```

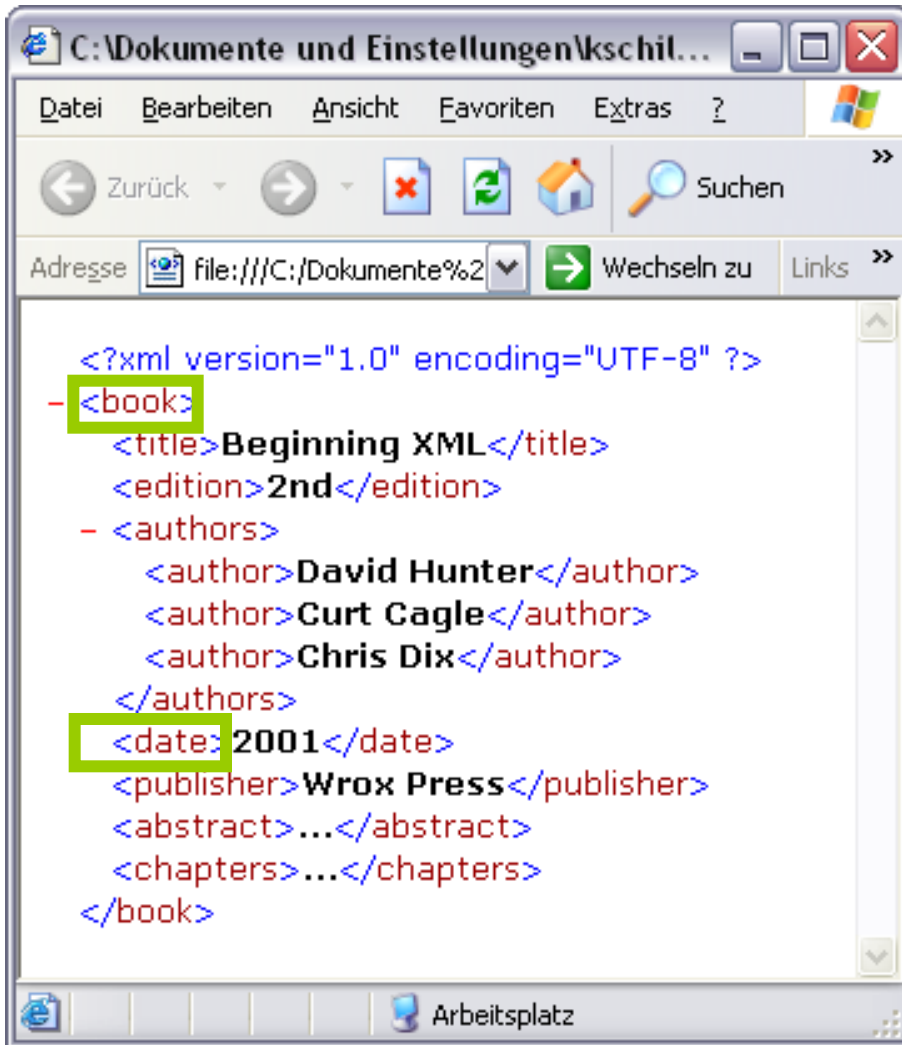
Metadata/Annotations



Species: Android
Size:...



Was ist XML?



```
<?xml version="1.0" encoding="UTF-8" ?>
- <book>
  <title>Beginning XML</title>
  <edition>2nd</edition>
  - <authors>
    <author>David Hunter</author>
    <author>Curt Cagle</author>
    <author>Chris Dix</author>
  </authors>
  <date>2001</date>
  <publisher>Wrox Press</publisher>
  <abstract>...</abstract>
  <chapters>...</chapters>
</book>
```

- Extensible Markup Language
- erlaubt Strukturieren von Inhalten
- Unterschiede zu HTML:
 - Medienneutral
- Tag-Namen
<name>...</name> beliebig
- generische
Auszeichnungssprache

Auszeichnungssprachen

- textbasierte Sprachen, die Dokumente mit zusätzlichen Tags („Markierungen“) versehen:

`<tag-name>ausgezeichneter Text</tag-name>`



- dadurch zusätzliche Information (Metainformationen)
- Beispiel: Hypertext Markup Language (HTML)
- kombinieren Vorteile von Binärdateien mit denjenigen von Textdateien:
 - anwendungsunabhängige Dateiformate, die reichhaltige Metadaten enthalten können

HTML

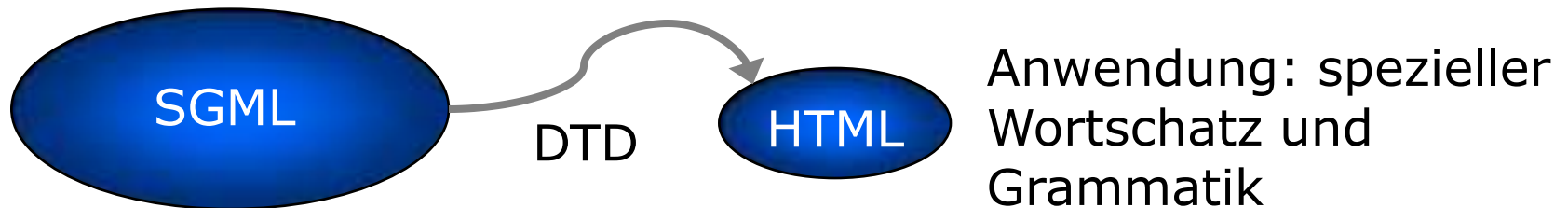
- vorgegebene Auswahl von Tags, keine anderen dürfen verwendet werden

generische Auszeichnungssprache (generalized markup language)

- keine Tags vorgegeben, beliebige Tags erlaubt
- Vorteil: beliebige Metainformationen darstellbar
- Nachteil: Bedeutung der Metainformationen (Tags) offen
- Beispiele: SGML und XML

- Standard Generalized Markup Language
- 1969 von Charles Goldfarb und zwei seiner Kollegen bei IBM für das Dokumentenmanagement entwickelt.
- seit 1986 ein internationaler Standard
- keine vorgegebenen Tags, auch keine für das Layout von Dokumenten
- Vorgänger von XML

- gibt zwar keine konkreten Tags vor
- Mit Document Type Definitions (DTDs) können aber spezielle Auszeichnungssprachen mit konkreten Tags definiert werden:
 - werden Anwendungen von SGML genannt
 - bekannteste Anwendung von SGML: HTML



- Anwendung selbst kann keine Anwendung definieren

Vor- und Nachteile von SGML

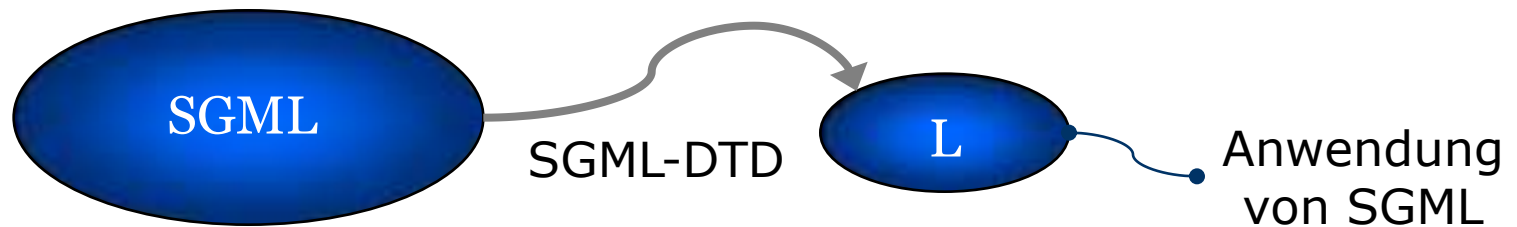
- + kombiniert Vorteile von Binärdateien mit denjenigen von Textdateien
- + beliebig erweiterbar
- + erlaubt die Definition von konkreten Auszeichnungssprachen wie HTML
- sehr komplex: Spezifikation über 600 Seiten lang
- SGML-Parser schwierig zu implementieren

- HTML
 - für Präsentation von Web-Inhalten bewährt
 - keine medienneutrale Darstellung von Inhalten
- medienneutrale Darstellung
 - generische Auszeichnungssprachen (wie SGML) geeignet
- SGML
 - für das Web SGML viel zu komplex

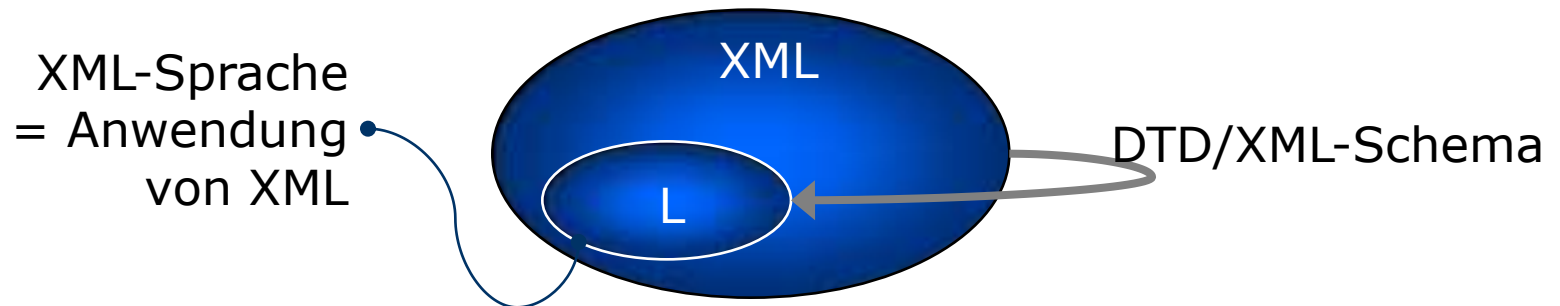
XML: konsequente Vereinfachung von SGML, die für Web-Anwendungen hinreichend allgemein ist.

Was bedeutet Erweiterbarkeit?

- X in XML steht für erweiterbar (extensible).
- Was bedeutet Erweiterbarkeit? → Vergleich HTML vs. XML hilfreich:
- HTML
 - vorgegebene Auswahl an Tags
 - Neues Tag kann nur eingeführt werden, wenn sich das W3C auf eine neue HTML-Version einigt!
- XML
 - beliebige Tags können benutzt werden
 - Anwender des entsprechenden Tags müssen sich auf eine gemeinsame Interpretation des Tags einigen

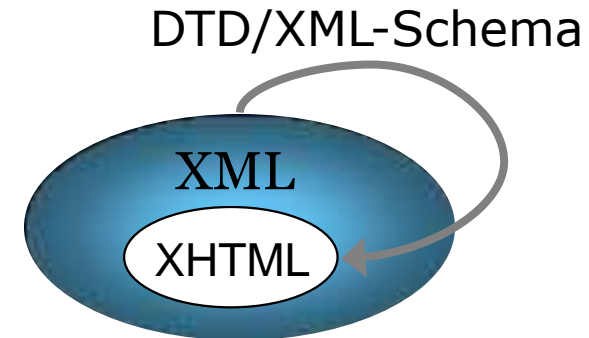
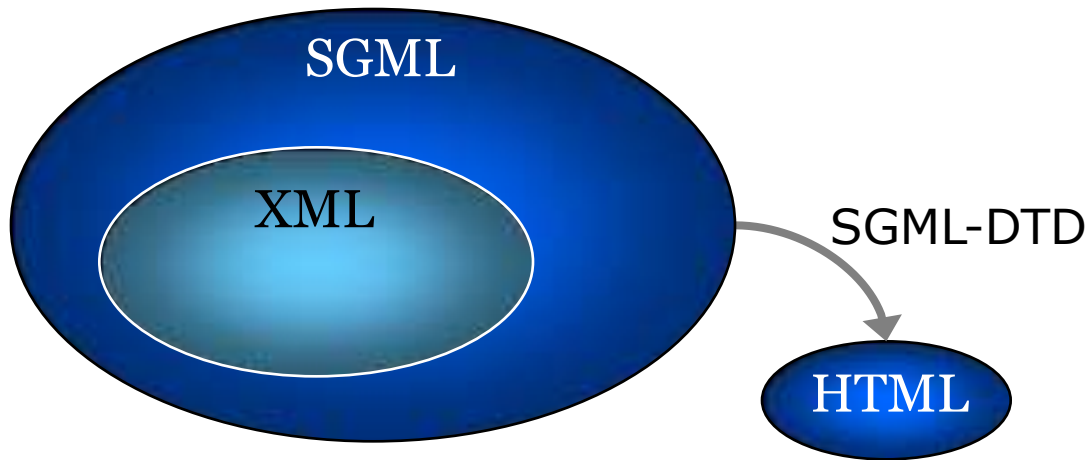


- L muss *nicht* Teilsprache von SGML sein.
- L kann *keine* neue Sprache definieren.
- Beispiel: HTML



- L immer Teilsprache von XML
- L kann *keine* neue Sprache definieren.
- Beispiel: XHTML

SGML, HTML, XML, XHTML?!



HTML

- Anwendung von SGML

XML

- Teilsprache von SGML

XHTML

- XML-Sprache = Anwendung von XML
- alle XHTML-Dokumente immer wohlgeformte XML-Dokumente

Die XML-Familie: Der Kern

- XML 1.0 / 1.1
 - Syntax wohlgeformter XML-Dokumente
 - Definition von Anwendungen (Untermengen) mit DTDs
- Namensräume
 - gleichzeitige Verwendung unterschiedlicher Vokabularien
 - z.B. Unterscheidung Titel einer Person vom Titel eines Buches
 - Festlegung der Bedeutung von Tags
- XML-Schema
 - gleiche Aufgabe wie DTDs
 - jedoch wesentlich mächtiger

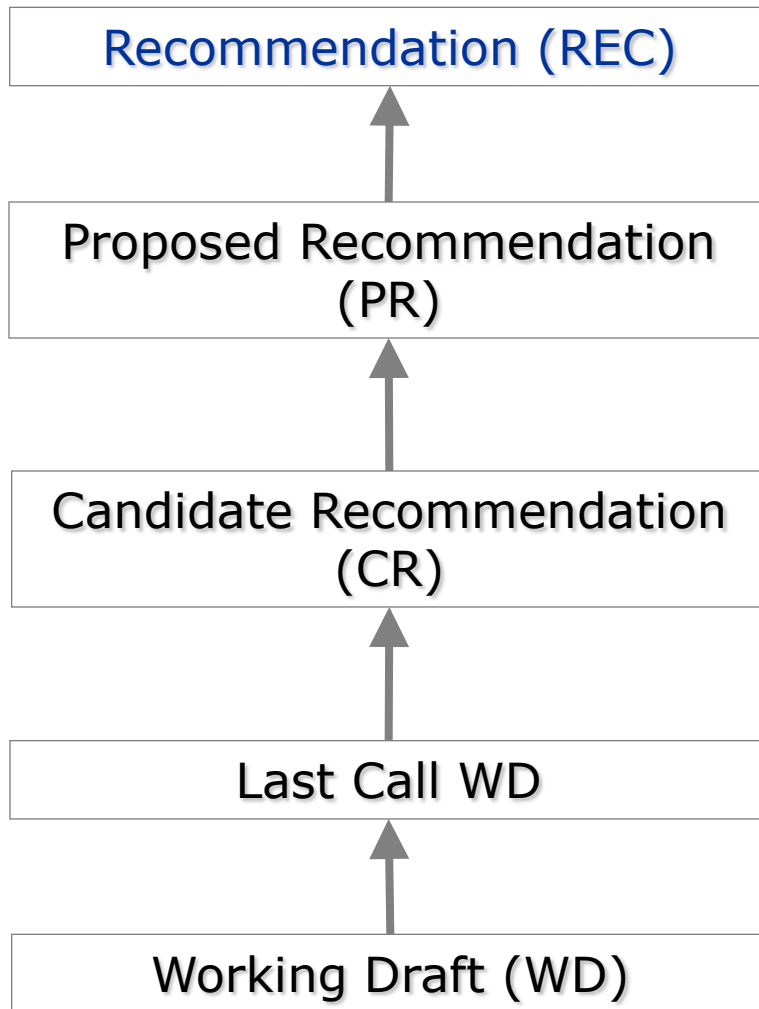
Wichtige XML-Familien-Mitglieder

- Extensible Stylesheet Language (XSLT)
 - Transformation von XML-Dokumenten in beliebige Text-Formate:
XML → HTML / WML / XML / ASCII / ...
- XPath
 - Zugriff auf beliebige Teile eines XML-Dokuments
 - z.B. Zugriff auf alle Buchtitel
- XQuery
 - Abfragesprache
- Document Object Model (DOM)
 - Parsen, Modifizieren und Erstellen von XML-Dokumenten

gesamte XML-Familie besteht aus
lizenzfreien W3C-Standards



- 1994 als Projekt am MIT gegründet
- keine Normierungsorganisation im klassischen Sinn
- kann Einhaltung von Normen nicht auf rechtlichem Wege einklagen
- definiert deshalb lediglich Empfehlungen (recommendations)
- W3C-Recommendations lizenzfrei



offizieller W3C-Standard

offizieller Konsens der betreffenden AG dar, wird dem Advisory Committee übergeben

Direktor: definierte Ziele erreicht, von entsprechender Community begutachtet

letztes WD, definierte Anforderungen erreicht

aktueller Diskussionsstand einer AG



Anwendungen von XML

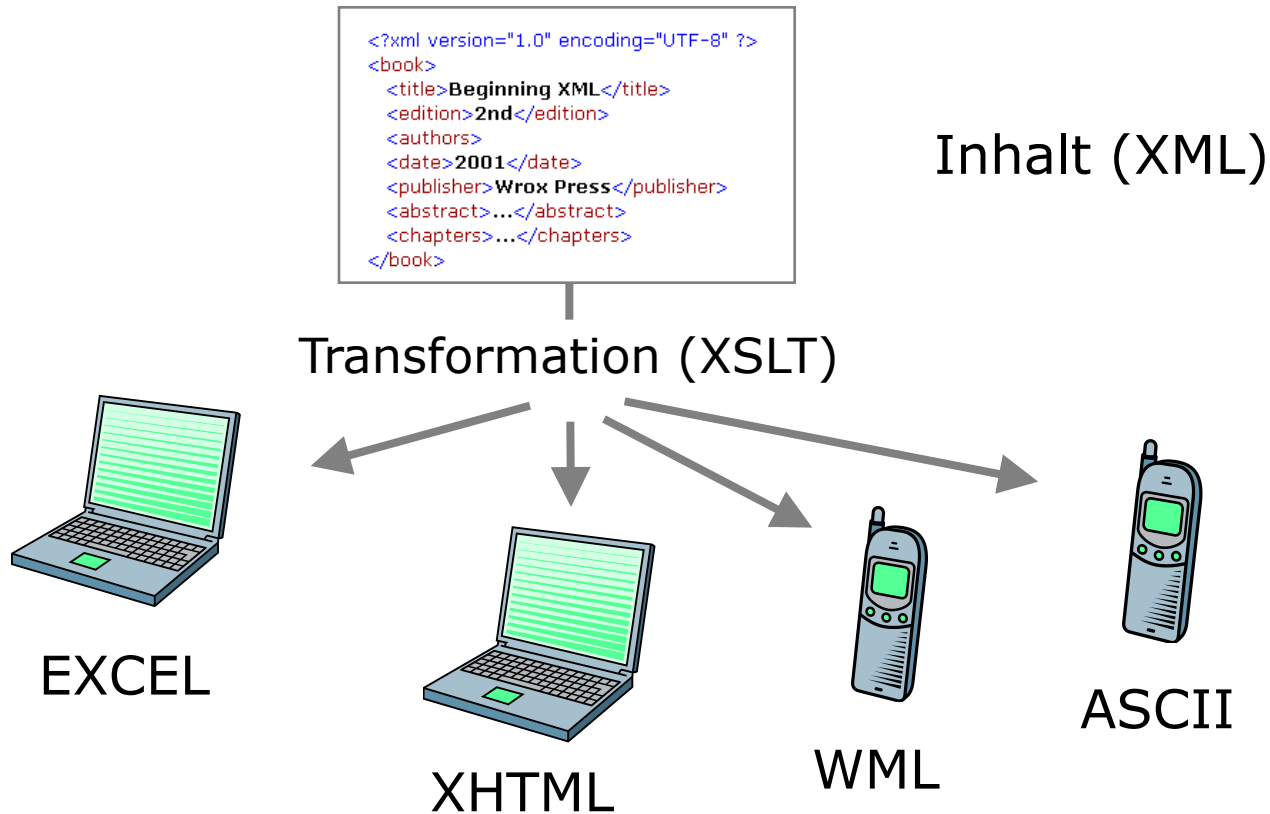
1. Anwendungsspezifische Standards

- XML hat uneingeschränkten Wortschatz:
<xyz>David</xyz>, <aβγ>Hunter</aβγ>
- für spezielle Anwendungen kann jedoch spezifischer Wortschatz und Grammatik festgelegt werden

```
<book>  
  <title> STRING </title>  
  <authors>  
    <author> STRING </author>+  
  </authors>  
  <date> DATE </date>  
  <ISBN> STRING </ISBN> ?  
  <publisher> STRING </publisher>  
</book>
```

- sog. XML-Sprachen (oder Anwendungen von XML)
- mit DTDs und XML-Schemata

2. Trennung Inhalt von Präsentation



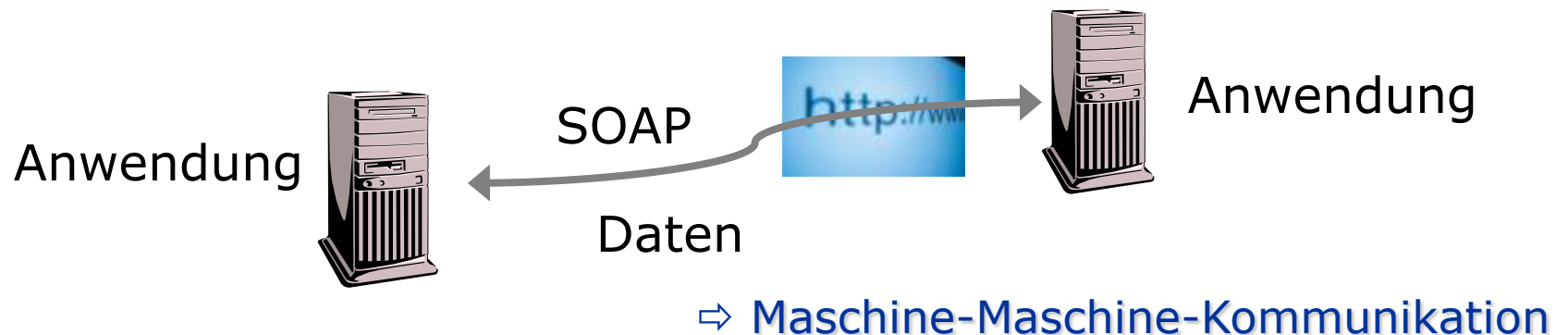
- **Multi-Delivery**: Trennung Inhalt von Präsentation
- weit verbreitet, aber nicht sichtbar!

3. Web-Dienste (Web Services)

traditionelle Web-Anwendung



Web Service



- **Syntax** – die Art und Weise, wie Worte in einem Satz zusammengesetzt wurden.
- **Semantik** – Informationen, die in diesem Sinne kodiert wurden.
- **Pragmatik** – Implikationen aus den Informationen in einem Kontext.

Bildersuche: „Apache“



Maschinen fehlt dieser Kontext aus Begriffen und Zusammenhängen

Kontext muss Maschinen zusätzlich bereitgestellt werden

“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.”

Berners-Lee, Hendler, and Lassila, 2001.



Foto: W3C



Foto: Homepage



Foto: Homepage

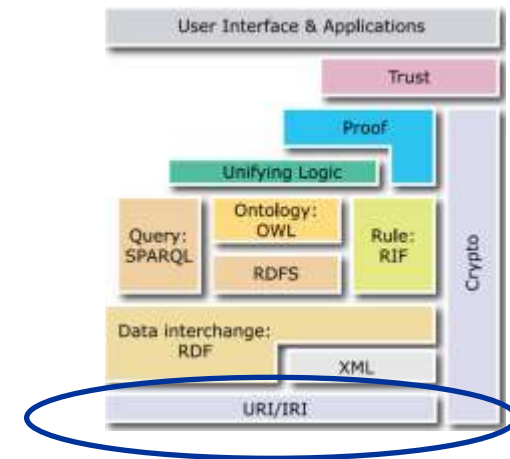
- Damit Metadaten nutzbar sind
 - muss der Informationsanbieter sich so ausdrücken, dass Informationsnutzer ihn verstehen
 - muss der Informationsnachfrager so fragen, dass er etwas finden kann
- Gemeinsame Benutzung von Konzepten
- Gemeinsame Sprache

- Ontologie zur Definition einer gemeinsamen Sprache
 - Es gibt Konzepte, die wir mit „Bank“ und „Sparkasse“ benennen
 - Es gibt ein Konzept, das wir „Geldinstitut“ nennen und das die Konzepte „Bank“ und „Sparkasse“ umfasst

Unicode

jedes Zeichen eigene Nummer (system-, programm- und sprachunabhängig)

Unicode-Codierung – Zeichensätze für fast jede natürliche Sprache



URI – Uniform Resource Identifier

eindeutige Identifikation einer Quelle/Ressource → jedes beliebige Objekt verfügt über einen URI

Mechanismus um Daten verteilt repräsentieren zu können

URLs – Untergruppe von URIs

Syntax vom W3C standardisiert

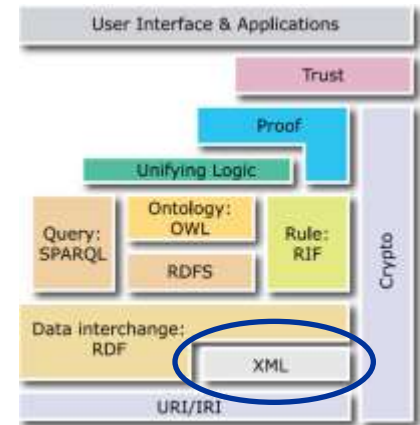
XML + Namensräume + XML-Schema

hierarchisch strukturierte,
medienneutrale Daten

Vokabular kann mit XML-Schema
definiert werden

Bedeutung des Vokabulars kann mit Namensräumen
festgelegt werden

XML-Daten können mit XLink verlinkt werden: Links
können Namen, aber keinen Namensraum haben



⇒ maschinenverarbeitbare verlinkte Daten,
Links jedoch nicht maschinenverarbeitbar

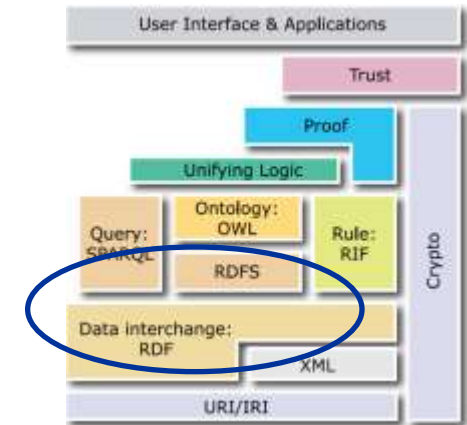
RDF + Namensräume + RDF-Schema

Web als Menge vernetzter Ressourcen

Vokabular für Beziehungen kann
mit RDF-Schema definiert werden

Bedeutung des Vokabulars wird
mit Namensräumen festgelegt

RDF Modell bietet eine syntaxunabhängige Darstellung



⇒ maschinenverarbeitbares
Netzwerk von Beziehungen

Ontologien

Vokabulare

Begriffsbeziehungen (Unterklasse, Untereigenschaft, Wertebereiche, ..., selbstdefinierte)

Sprache für Web-Ontologien:

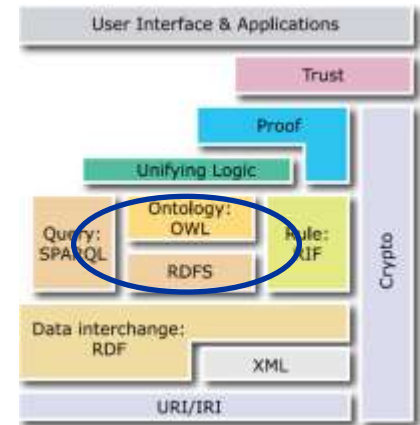
OWL – Web Ontology Language

Erweiterte Beschreibungsmöglichkeiten

In unterschiedlichen Komplexitäten

(OWL-Lite, OWL-DL, OWL-Full)

mittlerweile OWL 2 mit feinerer Unterscheidung der Komplexität

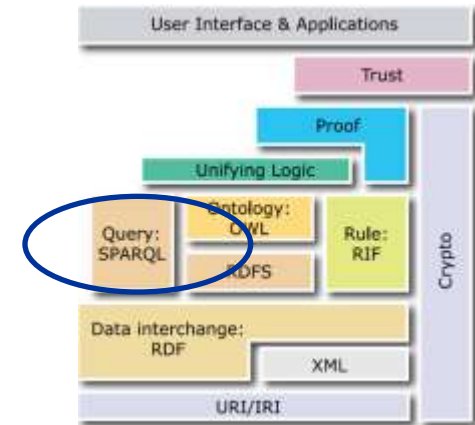


Anfragesprache SPARQL

Dient zur Abfrage von Instanzdaten in einer RDF-Datenbank

„Gib mir alle Menschen, die vor 1900 in Berlin geboren wurden“

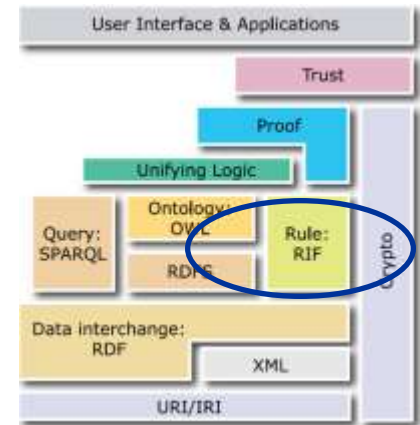
```
SELECT ?name ?birth ?death ?person
WHERE {
  ?person dbpedia2:birthPlace <http://dbpedia.org/resource/Berlin> .
  ?person dbo:birthDate ?birth .
  ?person foaf:name ?name .
  ?person dbo:deathDate ?death
  FILTER (?birth < "1900-01-01"^^xsd:date) .
}
ORDER BY ?name
```



Regelsprachen

bilden die Grundlage für das logische
schließen auf Basis semantischer Daten
früher SWRL (echte Regelsprache für OWL)
als Teil des Layer Cakes

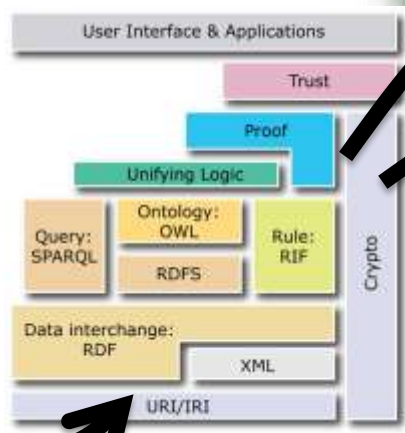
heute RIF als ein Austauschformat
zwischen unterschiedlichen Regelsystemen



Some people say the vision has failed...



... but in reality the „Gauls“ resist and make it again – but now bottom-up.



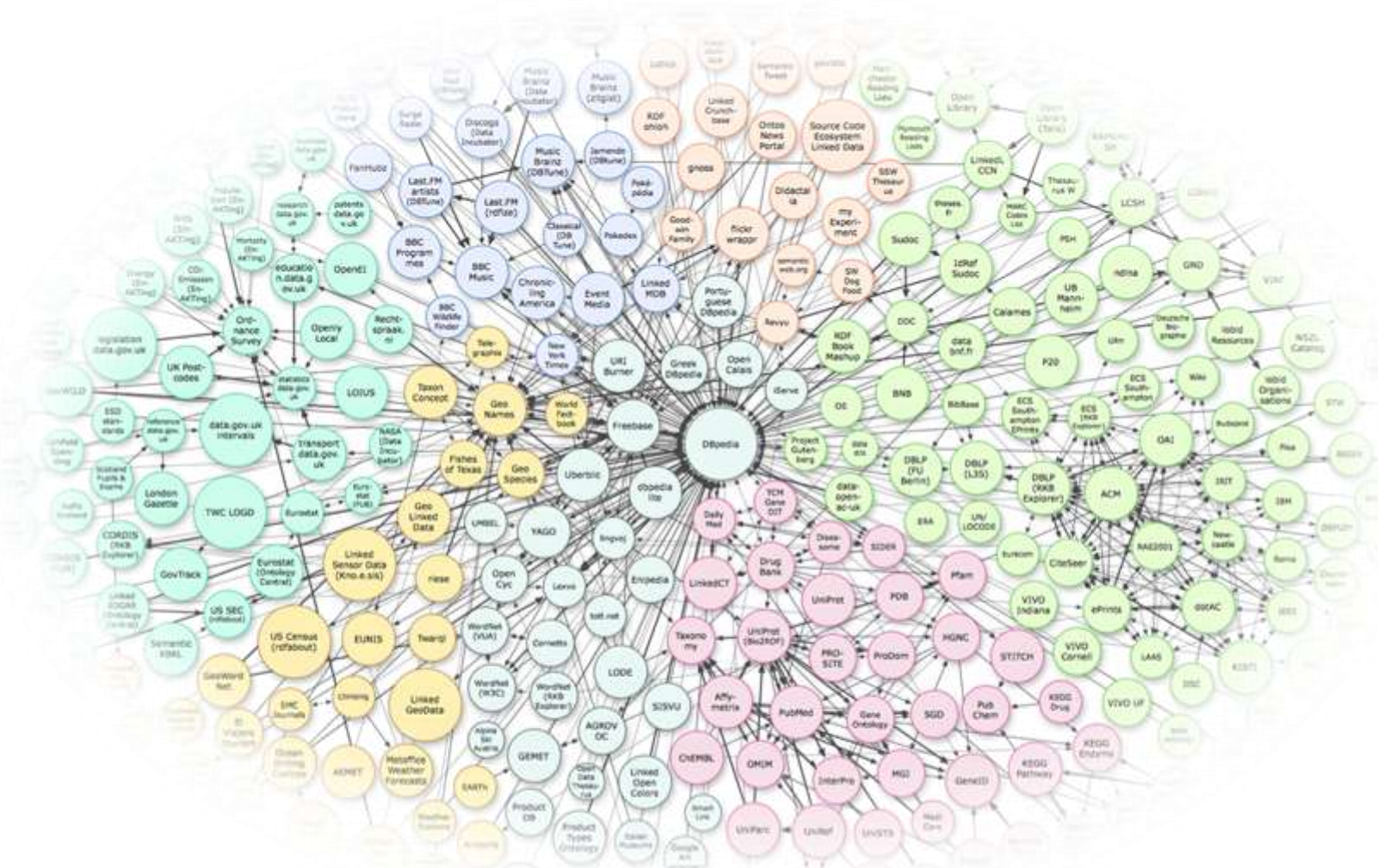
Make it “Bearable”



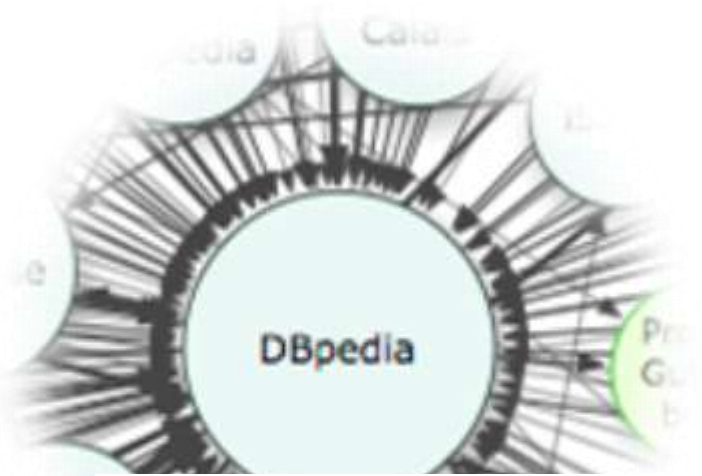
RECYCLE



„A little semantics...“*



*Original quote: „A little semantics goes a long way.“ – Prof. James Hendler in the late 90's



- wraps Wikipedia and transforms information into RDF



Article Discussion

Edsger W. Dijkstra

From Wikipedia, the free encyclopedia

Edsger Wybe Dijkstra (contributions to development)
Shortly before his death in the Dijkstra Prize the falls



Edsger Wybe Dijkstra

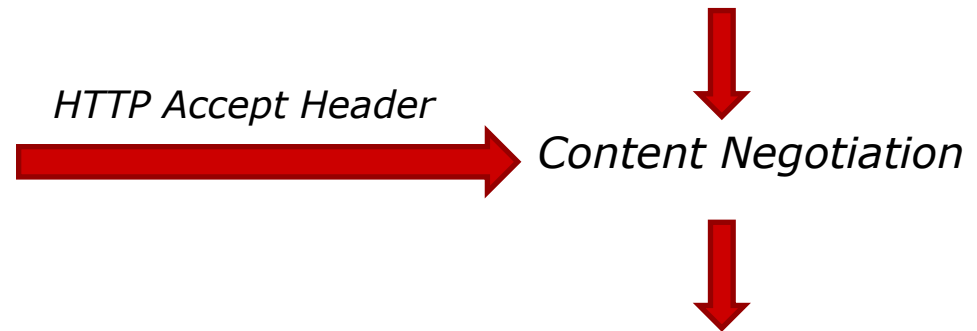
Born	May 11, 1930 Rotterdam, Netherlands
Died	August 6, 2002 (aged 72) Nuenen, Netherlands
Fields	Computer science
Institutions	Mathematisch Centrum Eindhoven University of Technology

```
<rdf:Description
rdf:about="http://dbpedia.org/resource/Edsger_W._Dijkstra">
  <dbpprop:birthDate
    rdf:datatype="http://www.w3.org/2001/XMLSchema#
    date">
    1930-05-11
  </dbpprop:birthDate>
</rdf:Description>
```

1. URIs as names for “things”

`http://dbpedia.org/resource/Berlin`

2. HTTP URIs so that people can look up those names.



3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)

`http://dbpedia.org/page/Berlin`
`http://dbpedia.org/data/Berlin`

4. Include links to other URIs. so that they can discover more things.

`yago-res:Berlin` S
`owl:sameAs` P
`dbpedia:Berlin` O

What is possible?

- „Find all soccer players, who played as goalkeeper for a club that has a stadium with more than 40.000 seats and who were born in a country with more than 10 million inhabitants“

```

SELECT DISTINCT ?player {
  ?s rdfs:type <http://dbpedia.org/ontology/SoccerPlayer> .
  ?s dbpedia:position ?position .
  ?s <http://dbpedia.org/property/club> ?club .
  ?club <http://dbpedia.org/ontology/capacity> ?cap .
  ?s <http://dbpedia.org/ontology/birthPlace> ?place .
  ?place ?population ?pop .
  OPTIONAL { ?s <http://dbpedia.org/ontology/number> ?tricot . }
  Filter (?population in (<http://dbpedia.org/property/populationEstimate>, <http://dbpedia.org/property/
  Results:   
  
```

SPARQL results:

player
http://en.wikipedia.org/wiki/Ferric_Cacho
http://en.wikipedia.org/wiki/Ferris_Krutz
http://en.wikipedia.org/wiki/Dwight_Kahn
http://en.wikipedia.org/wiki/Mat_Taylor
http://en.wikipedia.org/wiki/Lex_Yaeger
http://en.wikipedia.org/wiki/Russell_Ngumahun
http://en.wikipedia.org/wiki/Boo_Vassero
http://en.wikipedia.org/wiki/Jean-Marc_Petit
http://en.wikipedia.org/wiki/Luke_McCormick
http://en.wikipedia.org/wiki/Toby_Wilson
http://en.wikipedia.org/wiki/Shane_Hollop
http://en.wikipedia.org/wiki/Russell_Houtz





Überblick über die Vorlesung

Vorlesungsinhalt

- XML-Basistechnologien
 - 5 Termine
- Interoperabilität im Web
 - 4 Termine
- Einführung Projektarbeit
 - 1 Termin
- Rückblick & Ausblick
 - 1 Termin
- Klausur
 - 1 Termin

- XML-Basistechnologien - 5 Termine
 - XML-Syntax, einschl. Namensräume
 - DTDs und XML-Schemata
 - XML-Parser
 - XSLT, XPath, etc.
- nicht (explizit) behandelt werden:
 - XML-Technologien zur Präsentation von Dokumenten wie XHTML oder WML
 - anwendungsspezifische XML-Standards wie SVG oder VoiceXML



Image: <http://www.morguefile.com/archive/display/211651>

- Interoperabilität im Web
 - Web Services und Web APIs
 - 2 Termine
 - SOAP & WSDL
 - REST
 - JSON
 - Semantic Web Grundlagen und RDF - 1 Termin
 - Linked Data Microformate – 1 Termin
 - Linked Data
 - HTML 5
 - Microformats
 - RDFa



Applications

Image: <http://www.morguefile.com/archive/display/24026>

- Projektarbeit: 1 Termin
 - Einführung in die Praxis des Projektmanagements
 - Aufgabenvorstellung
- Rückblick: 1 Termin
 - kleine Wiederholung
 - Schwerpunkt → Klausurfragen



Image: <http://www.morguefile.com/archive/display/564796>



Literatur



Hunter et al., Beginning XML
(3rd Edition), Wrox Press, 2004.

ca. 41 €

Im Semesterapparat!

- XML
 - 1.0, W3C Recommendation, Sept. 2006, <http://www.w3.org/TR/xml/>
 - 1.1, W3C Recommendation, Sept. 2006, <http://www.w3.org/TR/2006/REC-xml11-20060816/>
- XML-Schema
 - XML Schema Part 0: Primer Second Edition, W3C, 2004
- XSLT
 - XSL Transformations (XSLT) Version 1.0, W3C, Nov. 1999
 - XSL Transformations (XSLT) Version 2.0, W3C, Jan. 2007

- Web Services
 - SOAP Version 1.2 Part 0: Primer (2nd Edition), W3C, April 2007 (<http://www.w3.org/TR/2007/REC-soap12-part0-20070427/>)
 - Web Services Description Language (WSDL)
 - WSDL Version 1.1, W3C, 2001
 - WSDL Version 2.0, W3C Recommendation, Juni 2007
- Semantic Web Grundlagen und RDF
 - W3C RDF Primer: <http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>
 - W3C Semantic Web Standards: <http://www.w3.org/RDF/>
 - Linked Data Design Issues: <http://www.w3.org/DesignIssues/LinkedData.html>
- Moderne Markuptechnologien und Microformate
 - HTML 5 W3C Working Draft: <http://dev.w3.org/html5/spec/Overview.html>
 - W3C RDFa Primer: <http://www.w3.org/TR/xhtml-rdfa-primer/>

Wie geht es weiter?

- ☑ Organisatorisches
 - ☑ Was ist XML?
 - ☑ Überblick über die Vorlesung
 - ☑ Literatur
-
- XML-Syntax
 - Namensräume
 - Semantik von XML-Tags