Grundstudiumspraktikum

# *Mehrsprachigkeit im Semantic Web*

Walther v. Hahn, Cristina Vertan
{vhahn,cri}@nats.informatik.uni-hamburg.de

## Inhaltsübersicht

- Was ist und was bringt das Semantic Web ?
- Wie implementiert man das Semantic Web ?
- Was implementieren wir im Praktikum?
- Organisatorische Details

## Was ist das Semantic Web ?

- Die Idee wurde erstmal von Tim Berners-Lee 1998 vorgeschlagen
- Es gibt unterschiedliche Definitionen entsprechend den unterschiedlichen Aspekten (e-commerce, Netzwerk, KI, Wissensmanagement, Sprachverarbeitung)
- Semantic Web Agreement Group (SWAG) (2000)

SWAG 2001

Das Semantic Web ist ein Web für Dokumente und Dokumentteile, das explizit die Beziehungen zwischen Objekten beschreibt. Es enthält die semantische Information, die für maschinelle Verarbeitung benötigt wird.



## Warum ist das Semantic Web nötig -1-

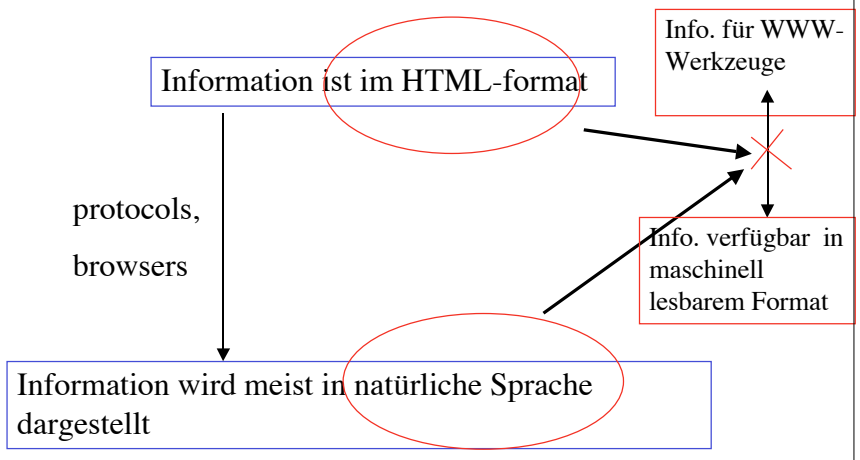
- Ca. 3 Milliarden statische Dokumente in WWW
- ca. 200 Milliarden Benutzer
- diese Ziffern steigen kontinuerlich

WWW heute

- Suchfunktionnen
- Zugänglichkeit
- Darstellungsmöglichkeiten
- Information up-to-date

schwierig

## Grenzen des aktuellen WWW -2-






## Suche in WWW

- meistens stellen die aktuellen Suchmaschinen viel zu viel irrelevante Information zur Verfügung.
- Die Suchmaschinen geben nur Pointer zu Dokumenten. Um die gewünschte Information zu erhalten, muss man die Dokumente lesen
- Wichtige Information für den gesuchte Begriff wird nicht gefunden weil sie mit anderen als den eingegeben Wörtern (oder in anderer Sprache) repräsentiert ist
- Die Situation ist schlechter, wenn man die Suche auch auf multimodale Information erweitert

## Informationsdarstellung im WW

- mit den aktuellen Web-Werkzeugen ist es sehr schwer nicht-redundante und konsistente Information zu pflegen
- Die Webmaster sind sehr oft überfordert
- Viele Webseiten enthalten widersprüchliche Information



## Was bringt das Semantic Web ?

Der Inhalt wird für die Maschinen verfügbar

- verbesserte zielorientierte Suche
- widersprüchliche Information wird automatisch entdeckt
- Informationsextraktion erleichtert und erweitert.
- Die Suche wird sich nicht mehr nur auf die Eingabesprache beschränken

## Inhaltsübersicht

- Was ist und was bringt das Semantic Web ?
- Wie implementiert man das Semantic Web ?
- Was implementieren wir im Praktikum?
- Organisatorische Details



## WWW Veränderung im Semantic Web

Web 2te Generation

Semantic Webseiten:

XML, RDF -Technologien, die Bedeutung mit den Daten verbinden

Web = XML-basierte Datenbank + Softwareagenten

Web 1ste Generation

Dynamische Webseiten:

DHTML, Javascript, Java,

Server-side Technologien

Statische Webseiten:

HTML, Hyperlinks, GIFs, Datenbanken+SQL

## Warum reicht XML-Annotierung nicht aus ?

- Beispiel : wir suchen etwas über *„Java programming language“*
- Google wirft alle Dokumente aus, die eines der eingegebenen Wörter enthalten. Deswegen erhalten wir auch Seiten über die Insel Java...
- mit XML-Annotierung kann man unterscheiden zwischen z.B.:

  <programming> Java </programming> und

  <geographisch>Java </geographisch>
- aber ....



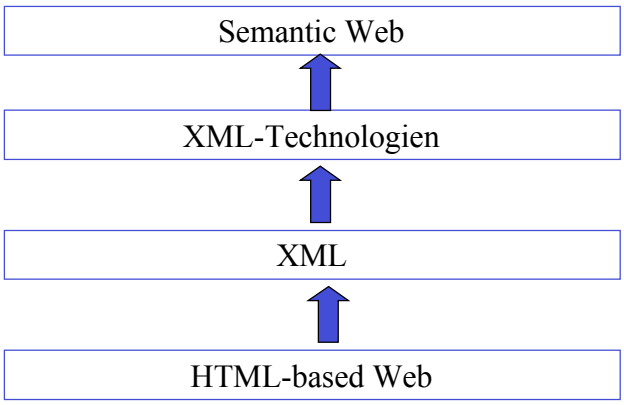
## Was kann XML nicht lösen

- z.B. ein tag:

<programmierung>Java</programmierung>

- wird auch ignoriert weil die Maschine nicht versteht, dass die Bedeutung von programming und programmierung dieselbe ist

## XML und WWW






## XML-Sprachen für das Semantic Web

- Erweitern Web-Daten und Web-Ressourcen durch Bedeutung
- Diese Bedeutung ist hierarchisch spezifiziert
- RDF (Resource Description Framework) definiert ein einfaches Datenmodel als ein Tripel (Subjekt, Prädikat, Objekt)


- Ein RDF Schema (RDFS) beschreibt RDF-Merkmale, und ermöglicht die Darstellung einfacherer Ontologien

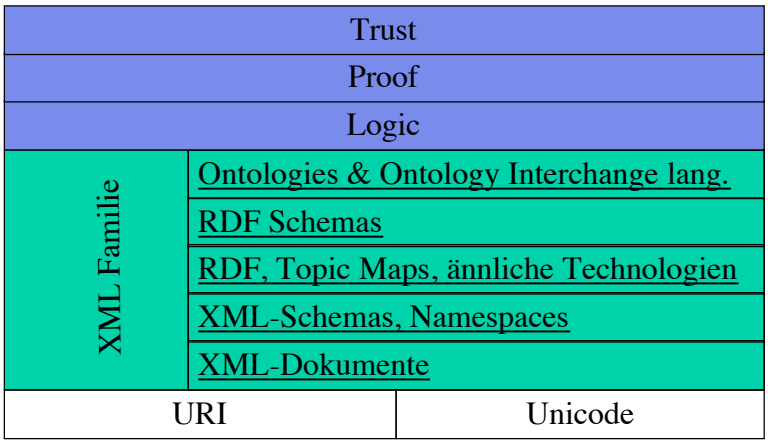
- DAML+OIL (DARPA Agent Markup Language + Ontology Inference Layer), OWL erweitern RDFS für die Darstellung von komplexeren Ontologien

## Ontologien

- Sind hierarchische Konzeptstrukturen mit semantischen Relationen (Ober/Unterbegriff, Teil-von, etc.).
- Sie definieren Beziehungen zwischen Tags und erlauben damit semantische Annotation und ihre Abbildung auf natürlichsprachliche Wörter.



## „Layer-cake“ Architektur (nach Tim Berners-Lee)



## Inhaltsübersicht

- Was ist und was bringt Semantic Web ?
- Wie implementiert man das Semantic Web ?
- Was implementieren wir im Praktikum?
- Organisatorische Details



## Funktionalität des Systems - Grobe Beschreibung

- Ein Demosystem zur Durchsuchung und zum Information-Retrieval von zweisprachigen Webseiten
- Die Webseiten werden durch eine Sammlung von deutschen und englischen Texten (Korpora) simuliert, wobei für jeden Text die reale Internet-Addresse angegeben ist.
- Die Eingaben des Systems sind in natürliche Sprache. Schlüsselwörter werden daraus extrahiert.
- Mit Hilfe von RDF-Annotationen im Text und eine Konzeptontologie werden relevante Paragraphen durchgesucht und als Antwort präsentiert.


## Worüber suchen wir Information ? -1-

http:// www.siebenburgen.de

http:// www.siebenburger.de



## Worüber sucht man Informationen ? -2-

## Vorhandenes Material

- Eine Sammlung von Texten auf Deutsch und Englisch über Siebenbürgen und wichtige Orte
  - Jede Datei enthält ein Hyperlink zu der WebSeite, aus der sie extrahiert wurde.
  - Für einen Ort sind die englischen und deutschen Texte (d.h. der Inhalt) nicht unbedingt identisch
- Eine Datei (Orte.txt), gibt die Namenkorrespondenz an
- Eine Datei mit Eingabebeispielen (muß erweitert werden)



## Deutsche /Englische Korpora - Beispiel-

Das Rothenburg Siebenbürgens

Das historische Zentrum von Sighisoara (Schässburg) wurde von der UNESCO auf die Liste des Weltkulturerbes gesetzt. Und das zu Recht. In der Mitte steht wie einst die mächtige Burg , die besterhaltene Siebenbürgens, die im 12. und 13. Jahrhundert auf den Ruinen eines römischen Kastells errichtet wurde. Das auffälligste und sehenswerteste Gebäude ist der 64 m hohe Stundturm .

.....

http://reisen.transylvaniatravel.net/staedteurlaub/reise311.html

Sighisoara

Strategically overlooking a valley full of picturesque villages boasting fantastic fortress churches, Sighisoara, with its small but beautifully preserved medieval core seems to be set for a prosperous future in tourism. But where are the tourists? While an attractive town like this would certainly be swamped daily by coachloads of visitors if it were in Hungary or Austria, Sighisoara can often be eerily quiet and deserted, with the notable exception of the yearly Medieval Folk Festival.

.....

http://www.inyourpocket.com/romania/sighisoara/en/

## Fragekorpus

- Zur Zeit enthält das Fragekorpus nur etwa 15 Beispiele
- Nach Anschauen der Texte sollte dieses Korpus zu etwa 50 Fragen (Eingaben) erweitert werden.
- Die Fragen sollen nicht unbedingt Wörter sein, die in den Texten erscheinen, sondern ähnliche Wörter (für verwandte Konzepte).



## Aufgaben mit den Texten

- Man muss die Texte erstmal analysieren und relevante Konzepte extrahieren.
- Dieses konzepte werden danach als tags für Paragraphe/Ausdrucke/Wörter benutzt werden.
- Beispiel:

*<beschreibung>Das historische Zentrum von <stadt_burg>Sighisoara (Schässburg)<stadt_burg> wurde von der UNESCO auf die Liste des Weltkulturerbes gesetzt. Und das zurecht. In der Mitte steht wie einst die mächtige Burg , die besterhaltene Siebenbürgens, die .....*

*</beschreibung>*

- Die Tags sollen auf deutsch für die deutsche Texte und auf englisch für die englische Texte

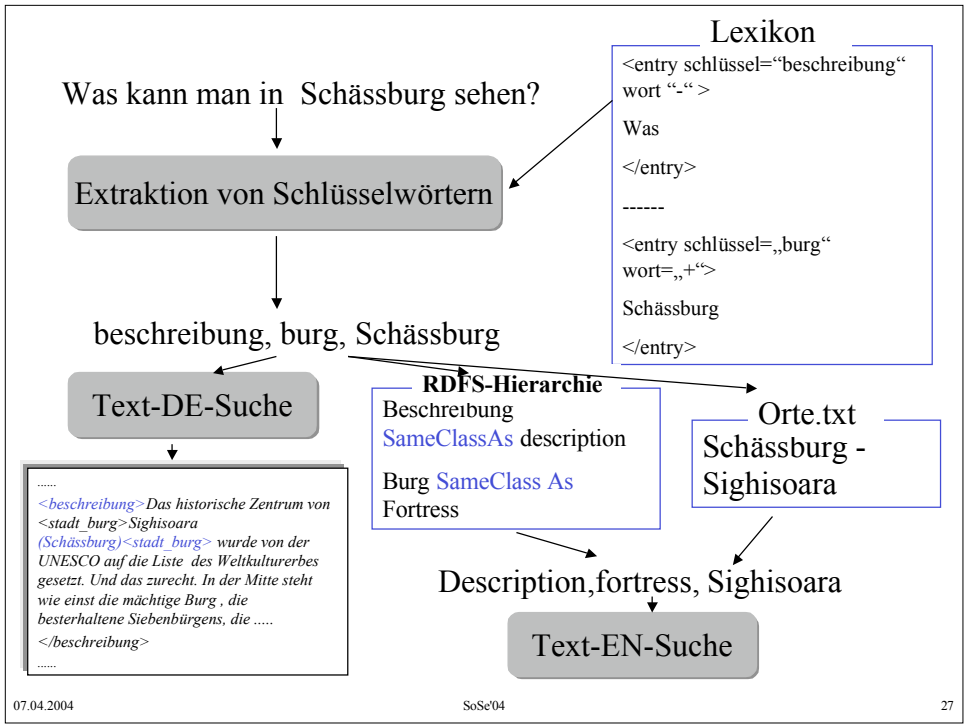
## Aufgaben mit den Tags

- Für jedes Tag muss beschrieben werden:
  - Welches andere Tag kann es enthalten (= Properties in RDFS)
  - In welcher Beziehung steht es mit anderen Tags (= Class/Subclass in RDFS)
  - Welche ist (wenn vorhanden) die Korrespondenz in anderen Sprachen (= sameClassAs in RDF)



## Aufgaben mit dem Fragekorpus

- Für jede Frage muss man Schlüsselwörter identifizieren
- Mit diesen Schlüsselwörtern erzeugt man ein Lexikon:
- Beispieleintrag im Lexikon :
  - <entry suche=„beschreibung“> was </entry>
- Von einer beliebigen Frage werden zunächst nur die Schlüsselwörter extrahiert.





## Inhaltsübersicht

- Was ist und was bringt Semantic Web ?
- Wie implementiert man das Semantic Web ?
- Was implementieren wir im Praktikum?
- Organisatorische Details

## Termine und Themenliste

| Nr. | Datum | Themen | Zwischentermine |
|---|---|---|---|
| 1 | 07.04 | T: Einführung im Semantic Web, Praktikum-Ablauf<br>P: Accounts; Analyse der vorhandenen Daten, Erweiterung des gegeben Korpus | Test-Korpus fertig |
| 2 | 14.04 | T: Annotation von texten im Semantic Web (RDF)<br>P: Annotation eines Beispeiltexts mit RDF | |
| 3 | 21.04 | T: Grundprinzipien der Ontologien. Ontologie-Darstellung mit RDFS<br>Lexikon-Darstellung. Mapping von Lexika auf Ontologien<br>P: Erstellung einer Beispiel-Ontologie | |
| 4 | 28.04 | T: System-Architektur; Verteilung von individueller Aufgaben | |
| 5 | 05.05 | P: Implementierung | Erste menge von RDF-tags sovie erste Version deer Lexika fertig |
| 6 | 12.05 | P: Implementierung | Erste Ontologie-Version fertig |
| 7 | 19.05 | P: Implementierung | |
| 8 | 26.05 | T: Kriterien für eigene Evaluation des Systems<br>P: Implementierung | |
| 9 | 09.06 | **P: Zwischenevaluation** | Eine erste version des Systems (für ein paar texte, ein Teil der Wörter und Konzepte) soll fertig sein |
| 10 | 16.06 | P: Implementierung | |
| 11 | 23.06 | P: Implementierung | |
| 12 | 30.06 | P: Implementierung | |
| 13 | 07.07 | **P: End-Evaluation** | Demosystem fast fertig |
| 14 | 14.07 | **P: Zusammenfassung; Öffentliche Präsentation des Systems** | Demosystem fertig |



## Scheinkriterien

- Individueller Nachweis von
  - Programmieranteil (obligatorisch)
  - Präsentation
  - Teil im Endbericht
  - Teil der Evaluation (nach Kriterien)

## Web-Seite des Praktikums

- http://nats-www.informatik.uni-hamburg.de/view/MSW/

- Wenn sie selbst Materialien und Zwischenergebnisse einfügen, melden Sie sich bitte an unter
- http://nats-www.informatik.uni-hamburg.de/view/TWiki/TWikiRegistration
  als "Group" bitte auswählen "MSWGroup"