

Kreativität und Technik

**Vorlesung im Modul 10-201-2334
im Wahlbereich Bachelor GSW
sowie im Modul 10-201-2333
im Bachelor Informatik**

Sommersemester 2018

Prof. Dr. Hans-Gert Gräbe

<http://bis.informatik.uni-leipzig.de/HansGertGraebe>

Namensräume und Begriffswelten

Kommunikation wird durch die Einführung von **Namensräumen** als URI-Präfix unterstützt.

Namensräume erlauben es, überlappungsfrei URIs zu generieren.

- Damit können Beschreibungen erzeugt werden, welche die *Fiktionen* MEINE Welt, MEINE Begriffe, ICH-Kern, Welt und Wirklichkeiten, Wirklichkeitskonstruktion bedienen, ohne diese gedanklich transzendieren zu müssen.
- Aber wir wollen ja mehr: Kooperation mit konkreten Anderen.
- Semantik = pragmatisch kontextualisierte Bildung von *Modellen* als Basis für gemeinsame *Verfahrensweisen*.
- Wir brauchen Sprache, um über die Modelle selbst sprechen zu können, und damit Wege der Formalisierung von Semantik.

Konzeptualisierungen und Begriffswelten

Aber: **Das Turmbau von Babel Phänomen**

- Was bedeutet es, dass jeder kommunikative Zusammenhang hierbei zusammen mit den eigenen Verfahrensweisen auch seine eigenen Modelle und damit seine eigene Sprache entwickelt und spricht?
- Welche Konzepte können Übersetzungsleistungen unterstützen?
- **Ontologien** (oder Vokabulare): WIR einigen uns auf die Verwendung gemeinsamer Namensräume (foaf, skos, org, sioc usw.) für *spezielle* Zwecke und damit auf gemeinsame *Teilmodelle* von Welt.
 - Phänomen der Kohärenz privater und kooperativer Sprachpraxen.
- Wie geht das genau?
 - **Inhalt:** Pragmatisch kontextualisierte Bildung von *Modellen* als Basis für gemeinsame *Verfahrensweisen*.
 - **Form:** *Ontologisierung* als pragmatisch kontextualisierte Semantifizierung von Syntax.
 - Beispiele auf der nächsten Folie

Konzeptualisierungen und Begriffswelten

- **Beispiel foaf: Friend of a Friend**
 - foaf: <<http://xmlns.com/foaf/0.1/>>
 - Weiterleitung auf <http://xmlns.com/foaf/spec/>
 - Wir studieren das dort entwickelte Modell und die Beschreibungsformen von Semantik und Syntax
- **Beispiel skos: Simple Knowledge Organization System**
 - skos: <<http://www.w3.org/2004/02/skos/core#>>
 - Weiterleitung auf tabellarische Übersicht
<https://www.w3.org/2009/08/skos-reference/skos.html>
 - Ganz am Ende der Seite drei Referenzen
- **Beispiel org: The Organization Ontology**
 - org: <<http://www.w3.org/ns/org#>>
 - Weiterleitung auf eine Turtle-Datei. Herunterladen und anschauen
 - rdfs:seeAlso <<https://www.w3.org/TR/vocab-org/>>

Konzeptualisierungen und Begriffswelten

- Sozial ein extrem schwieriger Prozess, aber das ist **der Kern semantischer Technologien**: Die *Institutionalisierung* maschinenlesbarer gemeinsamer Begriffswelten *als sozialer Prozess*.
- Damit verbunden sind *Modellbildungen*, Bedingtheiten (Kontextualität verschiedener Wirklichkeiten) und der Prozess des Transzendierens von Kontexten, wenn Ontologien nicht wie ursprünglich vorgesehen angewendet werden.
 - Miteinander sprechen – Vereinbarung von Ontologien
 - Weiterentwicklung von Ontologien
 - Große Datenbanken von Ontologien: <http://prefix.cc> oder <http://lov.okfn.org> (Linked Open Vocabularies)
 - Kreativität im kooperativen Kontext. Formalisierungserfordernis, um Informationen als Daten auszutauschen. Noch einmal das Konzertbeispiel.

Zwei Beispiele

- DBPedia – Extrahiert strukturierte Information aus Wikipedia
 - DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. DBpedia allows you to ask sophisticated queries against Wikipedia, and to link the different data sets on the Web to Wikipedia data. We hope that this work will make it easier for the huge amount of information in Wikipedia to be used in some new interesting ways. Furthermore, it might inspire new mechanisms for navigating, linking, and improving the encyclopedia itself.
 - Beispiel: <http://dbpedia.org/page/Leipzig>
- Linked-Data-Service der Deutschen Nationalbibliothek
 - Die Deutsche Nationalbibliothek erstellt einen Linked-Data-Service, der langfristig eine Nutzung der gesamten nationalbibliografischen Daten inklusive aller Normdaten durch die Semantic-Web-Community erlaubt. Sie ist bestrebt, durch diesen Datenservice einen Beitrag zur Informationsinfrastruktur weltweit zu leisten und damit eine Voraussetzung für moderne kommerzielle und nicht-kommerzielle Webdienstleistungen anzubieten.
 - <http://www.dnb.de/lds>

Schema.org

- Anderer Zugang: <http://schema.org> - Googles Ontologisierung der Welt und Einbau in Webseiten statt Aufbau einer verteilten Datenbank wie in der Linked Open Data Cloud.
- Schema.org und Microdata: <https://schema.org/docs/gs.html>
 - itemscope, itemtype und itemprop und die Verbindung zu RDF.
- Auszeichnung von Webseiten mit diesem Markup erhöht deren Sichtbarkeit bei Google.

Googles Knowledge Graph

- **Googles Knowledge Vault:** Extrahiert durch supervised learning aus den untersuchten Webseiten entsprechende Fakten als Googles Wissensbasis.
 - Enthielt 2014 über 1.6 Milliarden Fakten, die mit einem probabilistischen Konfidenzwert bewertet sind.
- **Google Knowledge Graph:** Konsolidierung und Anreicherung mit strukturierten Fakten aus Freebase (2007 gegründet, 2010 von Google aufgekauft), Wikipedia und Wikidata.
 - Enthielt 2016 über 70 Mrd. Fakten.
 - Ende 2015 wurde die Google Knowledge Graph API veröffentlicht, über die Webentwickler auf den Bestand zugreifen können.

Wolfram Alpha

- Ebenfalls Suchmaschine, die auf Fakten aufbaut, die aus eigener Recherche gewonnen wurden. Zusammen mit *Mathematica* als Compute Engine lassen sich komplexere Präsentationen und Visualisierungen erstellen. Ziel ist die Vernetzung von mathematischem Wissen und Allgemeinwissen.
- <https://www.wolframalpha.com>
 - Beispiel „Leipzig“.

XML - Extensible Markup Language

Quelle: http://de.wikipedia.org/wiki/Extensible_Markup_Language

- XML ist eine Auszeichnungssprache zur Darstellung hierarchisch strukturierter Daten in Form von Textdateien. XML wird für den plattform- und implementationsunabhängigen Austausch von Daten zwischen Computersystemen eingesetzt.
- Die vom W3C herausgegebene XML-Spezifikation (Recommendation, erste Ausgabe vom 10.02.1998, aktuell ist die fünfte Ausgabe vom 26.11.2008) definiert eine Metasprache, auf deren Basis durch strukturelle und inhaltliche Einschränkungen anwendungsspezifische Sprachen definiert werden.
- Diese Einschränkungen werden durch Schemasprachen, insbesondere XML Schema, ausgedrückt. Beispiele für XML-Sprachen sind: RSS, MathML, GraphML, XHTML, XAML, Scalable Vector Graphics (SVG), GPX, aber auch XML-Schema selbst.
- Ein XML-Dokument besteht aus Textzeichen, im einfachsten Fall in ASCII- bzw. UTF-Kodierung, und ist damit von Menschen lesbar.

XML und Text Encoding

- XML = **E**Xtended **M**arkup **L**anguage
- Markup wird verwendet, um Textteile auszuzeichnen
- `<tag a1="a1wert" a2="a2wert"> Text </tag>`
 - a1, a2 - Attribute
- Der Text kann selbst wieder Tags enthalten
- Darstellung als Baum → XML-DOM = Document Object Model
 - Das Dokument besitzt genau ein Wurzelement
- Die Reihenfolge der Zweige im Baum ist bedeutsam (Listensemantik), die Reihenfolge der Attribute eines Elements nicht (Mengensemantik).
- Die Struktur eines Dokuments sollte in einem *Schema* fixiert sein (XML Schema, DTD, RELAX NG als verbreitete Schemasprachen), das mit dem Dokument verbunden ist.

XML und Text Encoding

- Schemabeschreibungen enthalten oft auch Annotationen, um die Semantik der ausgezeichneten Textteile näher zu beschreiben.
- Begriffe Wohlgeformtheit und Validität.
- XML ist im Wesentlichen ein deklaratives Markup, das auf verschiedene Weise interpretiert (prozessiert) werden kann.
- XML wird verwendet, um annotierte Texte zu erfassen. Grundlage für den TEI-Standard der Digital Humanities zur editorischen Erfassung von Texten.
- Mehr: A Gentle Introduction to XML,
<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html>
- Beispiel aus dem Deutschen Textarchiv anschauen (Text-Bild-Ansicht) <http://www.deutsches-textarchiv.de>
- Beschreibung der einzelnen Elemente
<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-lb.html>

Das deutsche Textarchiv

- RDF = *Vielfalt* von Begriffswelten (und damit Ontologien) wird nach einheitlichen Metagrundsätzen strukturiert.
- XML/TEI = Aufbau *einer* Begriffswelt und eines XML-Bindings speziell für die Zwecke der editorischen Erfassung von Texten.
- Große Texterfassungsprojekte:
 - Deutsches Textarchiv – unter Leitung der BBAW in den Jahren 2007-2015 gefördertes DFG-Projekt.
 - <http://www.deutsches-textarchiv.de/doku/ueberblick>
 - TextGrid – Übernahme und Aufbereitung als XML/TEI von Texten aus der digitalen Bibliothek von editura (zeno.org)
 - <https://textgrid.de/digitale-bibliothek>
 - TextGrid ist allerdings mehr, eine komplette virtuelle Forschungsumgebung und Kooperationsplattform.

DTA, TextGrid und DARIAH-DE

- Das *Deutsche Textarchiv* wird von der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) seit 2013 als Langzeitprojekt betrieben.
 - CLARIN-Servicezentrum des Zentrums Sprache an der BBAW
<http://clarin.bbaw.de/de/>
- Die Projektförderung für *TextGrid* endete 2015 und wurde in die ebenfalls vom BMBF geförderte digitale Forschungsinfrastruktur DARIAH-DE – Digital Research Infrastructure for the Arts and Humanities übernommen. Damit wird die dauerhafte und langfristige Nutzung der Angebote von TextGrid auf neuestem technologischen Stand gewährleistet. (Quelle: <https://textgrid.de/>)
- DARIAH-DE ist Teil einer europaweiten Forschungsinfrastruktur, siehe <https://de.dariah.eu/>.

Die Deutsche Digitale Bibliothek

- Das Ganze bettet sich ein in die öffentliche digitale Verfügbarmachung von Kulturgütern
- Die Deutsche Digitale Bibliothek -
<https://www.deutsche-digitale-bibliothek.de>
 - Gemeinschaftsprojekt von Bund, Ländern und Kommunen
 - Ein freier und zentraler Zugang zu Informationen gehört zu einer modernen, funktionierenden Demokratie. Diese Erkenntnis gab den Anstoß zur Einrichtung des Internetportals der europäischen digitalen Bibliothek Europeana, die Kulturgüter aller Mitgliedsstaaten der Europäischen Union weltweit zugänglich macht. Mit der DDB kommt Deutschland der Aufgabe nach, seinerseits ein nationales Portal für digitalisierte Kunst- und Kulturgüter zu schaffen und so seinen Teil des europäischen kulturellen Gedächtnisses einzubringen.