Gesellschaftliche Strukturen im digitalen Wandel

Vorlesung im Modul 10-201-2333 im Wahlbereich Bachelor GSW sowie im Modul 10-202-2330 im Master und Lehramt Informatik

Wintersemester 2018/19

Prof. Dr. Hans-Gert Gräbe

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

UNIVERSITÄT LEIPZIG
Institut für Informatik
Betriebliche Informationssysteme

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.

UNIVERSITÄT LEIPZIG
Institut für Informatik
Betriebliche Informationssysteme

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.

UNIVERSITÄT LEIPZIG
Institut für Informatik
Betriebliche Informationssysteme

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“.

UNIVERSITÄT LEIPZIG
Institut für Informatik
Betriebliche Informationssysteme

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.

UNIVERSITÄT LEIPZIG
Institut für Informatik
Betriebliche Informationssysteme

XML und Text Encoding

- XML = E**X**tended **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 Wurzelelement
- 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

UNIVERSITÄT LEIPZIG
Institut für Informatik
Betriebliche Informationssysteme

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.

TEI – Das DTA Basisformat

Grundstruktur jedes TEI-Dokuments

Jedes TEI-Dokument im DTA besteht aus einem **Header**, der Metadaten zum publizierten Text umfasst, und einem **Textbereich**, der alle Transkriptionen mit den zugehörigen Annotationen enthält. Dieser Volltext umfasst dabei nicht allein den eigentlichen Buchtext, sondern alle Textbestandteile, so auch Titelseite und Vorwort sowie in der Regel die Register, Beigaben und Anhänge.

Die **Metadaten im TEI-Header** umfassen:

- die Titel- und Quellenangaben zur vorliegenden Textausgabe (innerhalb von <fileDesc>; siehe dazu Kap. Bibliographische Angaben),
- Angaben zu den editorischen Richtlinien, welche der Ausgabe zugrunde liegen (innerhalb von <encodingDesc>; siehe dazu Kap. Editorische Richtlinien) sowie
- erste inhaltliche Angaben zum Text (innerhalb von <profileDesc>; siehe dazu Kap. Dokumentklassifikationen).

Siehe dazu http://www.deutschestextarchiv.de/doku/basisformat_header

UNIVERSITÄT LEIPZIG
Institut für Informatik
Betriebliche Informationssysteme

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/.

UNIVERSITÄT LEIPZIG
Institut für Informatik
Betriebliche Informationssysteme

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 und Ländern.
  - Der Sitz der Geschäftsstelle befindet sich bei der Stiftung Preußischer Kulturbesitz in Berlin.
  - Ziel der Deutschen Digitalen Bibliothek (DDB) ist es, jedem über das Internet freien Zugang zum kulturellen und wissenschaftlichen Erbe Deutschlands zu eröffnen, also zu Millionen von Büchern, Archivalien, Bildern, Skulpturen, Musikstücken und anderen Tondokumenten, Filmen und Noten. Als zentrales nationales Portal soll die DDB perspektivisch die digitalen Angebote aller deutschen Kultur- und Wissenschaftseinrichtungen miteinander vernetzen.