

# Semantische Datenintegration

## Daten und ihre Modellierung

Jakob Voß

Hochschule Hannover

2017-04-01

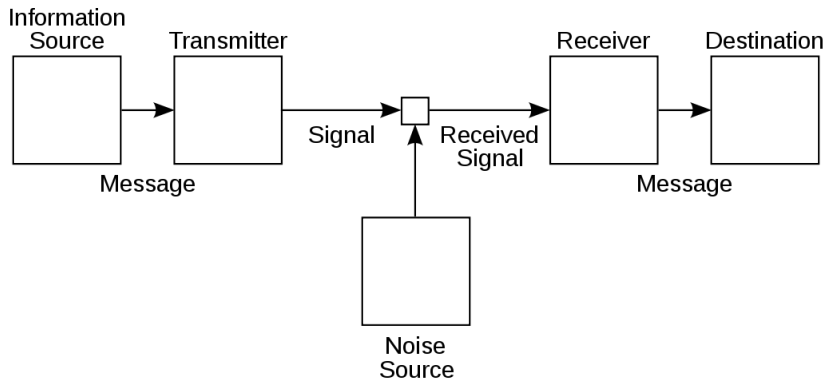
# Daten

# Frage

“Was sind Daten?”

# Daten und Informationen

- ▶ Mindestens bis in die 1960er keine fachliche Unterscheidung
- ▶ Informationstheorie (Shannon 1948) = Datentheorie



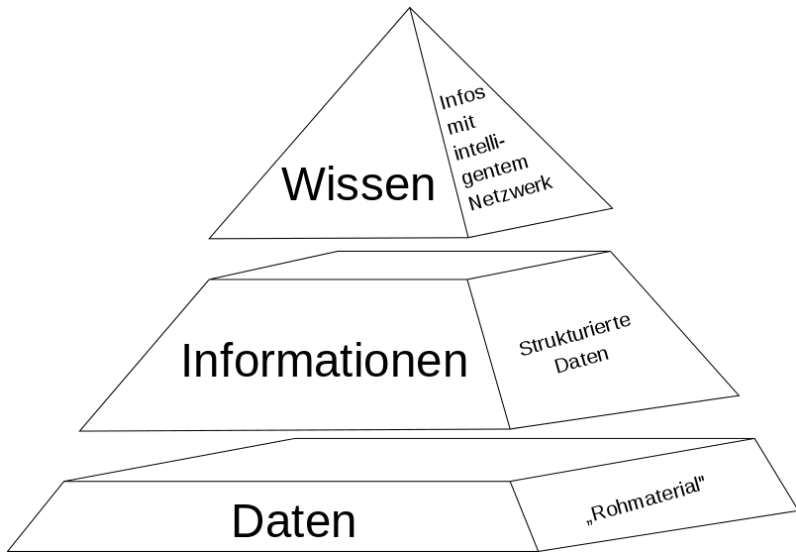
A mathematical theory of communication

# Daten im Rahmen der Informationstheorie

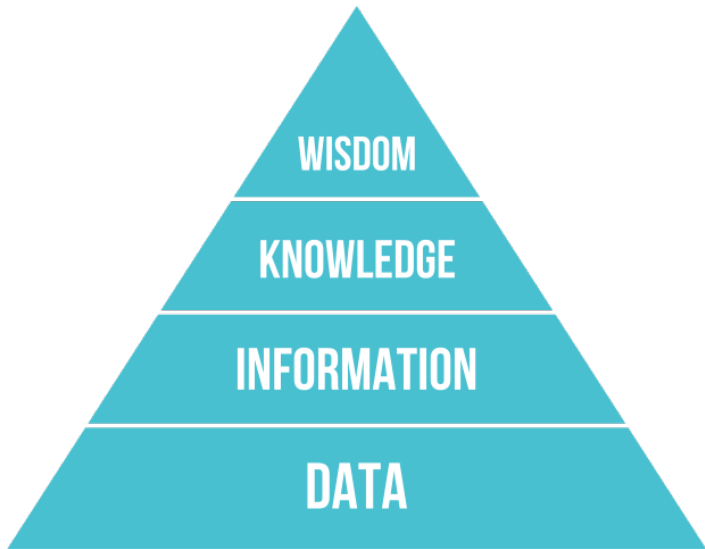
- ▶ Daten bestehen aus **bits** (0 oder 1)
- ▶ Eindeutige Information (Syntax)
- ▶ ohne Bedeutung (Semantik)

# Informationen und Daten

- ▶ Werden oft als praktisch gleich behandelt
- ▶ Hängen irgendwie zusammen
- ▶ Bauen aufeinander auf?



Wissenspyramide (Wikipedia)



Wissenspyramide (CC-BY-SA Longlivetheux)



# Computerunterstützte Integration heterogenen Wissens

- ▶ Daten zusammenführen
- ▶  $\Rightarrow$  Informationen zusammenführen
- ▶  $\Rightarrow$  Wissen zusammenführen

- ▶ Gibt es nicht in dieser Form
  - ▶ Vorschlag “Datalogy” (Naur 1966)
- ▶ Stattdessen Trends
  - ▶ EDV (1970/80er)
  - ▶ Linked Data (2006)
  - ▶ Big Data (2012)
  - ▶ Data Science (2013)

*Siehe Google Trends und ngram viewer*

# Daten-Trends

## EDV (1970/80er)

- ▶ Daten können automatisch verarbeitet werden

## Linked (Open) Data (ab 2006)

- ▶ Publikation von Daten in RDF

# Daten-Trends

## Big Data (2012)

- ▶ Immer mehr Daten werden automatisch erzeugt
- ▶ Viele Daten können statistisch ausgewertet werden

## Data (driven) Science (2013)

- ▶ (statistische) Datenanalyse
  - ▶ Data Mining
  - ▶ Künstliches Intelligenz
  - ▶ Visualisierung
  - ▶ ...

# Gemeinsamkeiten

- ▶ Daten können automatisch verarbeitet werden
  - ▶ weil eindeutig
- ▶ Immer mehr Daten werden publiziert
  - ▶ Computersysteme erzeugen mehr Daten
- ▶ Viele Daten können statistisch ausgewertet werden
  - ▶ Ist das relevant?

# Unterschiede

- ▶ Drei übliche Vorstellungen von Daten (Ballsun-Stanton 2012)
  - ▶ Daten als Fakten
  - ▶ Daten als Beobachtungen
  - ▶ Daten als binäre Nachrichten
- ▶ Welche können/wollen wir integrieren?

# Daten als Fakten

- ▶ Reproduzierbare Ergebnisse von Messungen
- ▶ Beispiele
  - ▶ Masse der Erde
  - ▶ Einwohnerzahl einer Stadt
- ▶ Paradigma
  - ▶ Semantic Web / Linked Data
  - ▶ Metadaten?
- ▶ Problem
  - ▶ Post-Faktisches Zeitalter
  - ▶ Kontext

# Daten als Beobachtungen

- ▶ Aufgezeichnete Wahrnehmungen
- ▶ Beispiele
  - ▶ Audio- und Videoaufzeichnungen
  - ▶ Historische Aufzeichnungen
- ▶ Paradigma
  - ▶ Big Data / Statistik
- ▶ Problem
  - ▶ Fokus auf quantitative Analyse



# Daten als binäre Nachrichten

- ▶ Zeichen, die zur Kommunikation dienen
- ▶ Eindeutig, aber ohne direkten Bezug zur Realität
- ▶ Letztendlich eine Folge von Bits
- ▶ Paradigma
  - ▶ Forschungsdaten
  - ▶ digitale Dokumente
- ▶ Problem
  - ▶ Es kommt auf den Einzelfall an

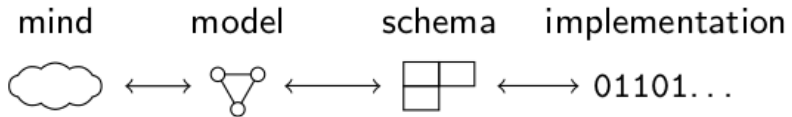
# Daten als binäre Nachrichten

- ▶ Kommunikativer Akt steht im Vordergrund  
(*Was will uns . . . mit diesen Daten sagen?*)
- ▶ Daten sind digitale **Dokumente**
  - ▶ Haben Ursprung, Form und Zwecke
- ▶ Kernthema der Bibliotheks- und Informationswissenschaft

# Zusammenfassung

- ▶ Daten als Fakten  
⇒ Einzelwissenschaften
- ▶ Daten als Beobachtungen  
⇒ Statistik & Maschinelles Lernen
- ▶ Daten als Dokumente  
⇒ Informationsmanagement

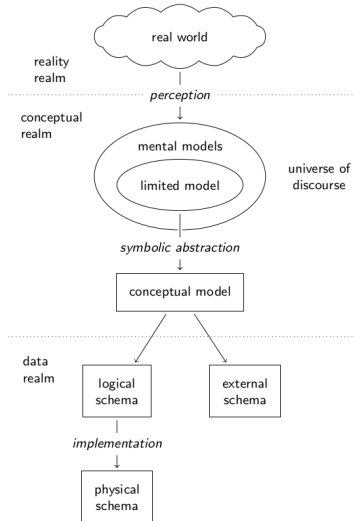
# Datenmodellierung



Datenmodellierungsprozess

# Ebenen der Datenmodellierung

- ▶ Vorstellungen
  - ▶ von der Realität
  - ▶ von dem was in Daten enthalten ist/sein soll
- ▶ Modelle
  - ▶ mentale Modelle (z.B. Mind-Maps)
  - ▶ konkrete Modelle (Modellierungssprachen)
- ▶ Schemas
  - ▶ Schemasprachen (SQL, XML Schema...)
  - ▶ Datenstrukturierungssprachen (XML, JSON, CSV...)
- ▶ Umsetzung in Daten



*Welche Studiengänge und ProfessorInnen gibt es im Deutschen Bibliotheks- und Dokumentationswesen?*



# Beispiel

- ▶ Objekte, Eigenschaften, Beziehungen. . .
- ▶ Möglichkeiten und Beschränkungen
- ▶ Schreibweisen/Formate

# Mögliche Datenquellen

- ▶ `http://www.kleinefaecher.de/bibliothekswissenschaft/`
- ▶ `http://www.kleinefaecher.de/informationwissenschaft/`
- ▶ `https://studieren.de/bibliotheks-und-dokumentationswesen.fachbereiche.t-0.f-67.html`
- ▶ Hochschuleseiten
- ▶ Hochschullehrerverzeichnis
- ▶ ...

# Mögliche Umsetzungen

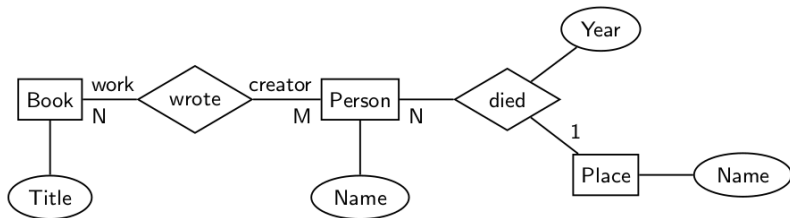
- ▶ Tabelle
- ▶ Strukturiertes Dokument
- ▶ Eigene Datenbank
- ▶ Vorhandene Datenbank (Wikidata)
- ▶ Abtippen
- ▶ Wrapper
- ▶ APIs

# Datensprachen

# Arten von Datensprachen

- ▶ Modellierungs-Sprachen (UML, ERM...)
- ▶ Schema-Sprachen (RDF Schema, XML Schema, RegExp...)
  - ▶ Abfragesprachen (SQL, XPath...)
- ▶ Datenstrukturierungssprachen (CSV, XML, JSON...)
  - ▶ Auszeichnungssprachen (HTML, TEI, Markdown...)
- ▶ Kodierungen (Unicode, ASCII, Binärcode...)

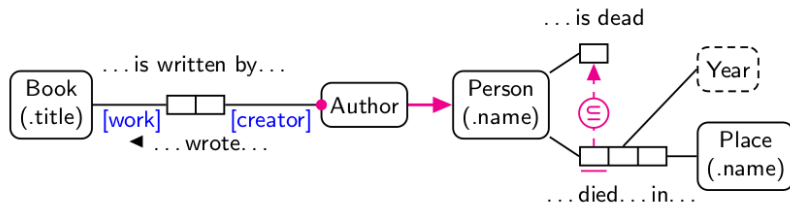
# Modellierungssprachen: ERM



# Modellierungssprachen

- ▶ Entity-Relationship Model (ERM)
- ▶ Unified Modeling Language (UML)
- ▶ Object Role Modeling (ORM2)
- ▶ ...

# Modellierungssprachen: ORM2





*the impact of the very substantial amount of work on modeling languages appears to be minimal, with modelers apparently preferring to work with the DBMS language*

(Simsion 2007, 345)

# Schemasprachen

- ▶ Auch bekannt als
  - ▶ Datendefinition
  - ▶ Datenbeschreibung
  - ▶ Formatbeschreibung
  - ▶ ...

# Beispiele für Schemasprachen

- ▶ Backus-Naur-Form und Reguläre Ausdrücke
- ▶ XML Schema
- ▶ RDF Schema
- ▶ SQL
- ▶ JSON Schema
- ▶ ...

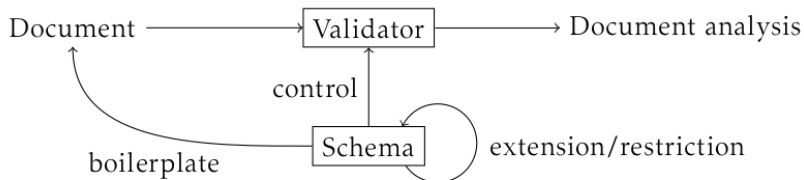
## Beispiel: SQL

```
CREATE TABLE Authorships (  
    workID int NOT NULL,  
    authorID int NOT NULL,  
    FOREIGN KEY (workID) REFERENCES Works(id),  
    FOREIGN KEY (authorID) REFERENCES Authors(id),  
    UNIQUE (workID, authorID)  
)
```



# Schemasprachen

- ▶ Eigene Syntax (mit Varianten!)
- ▶ Automatisierbar
- ▶ Anwendung für konkrete Datenstruktur



# Abfragesprachen

- ▶ XPath
- ▶ XQuery
- ▶ SQL
- ▶ ...
- ▶ Programmiersprachen
- ▶ APIs
- ▶ ...
- ▶ Feldnamen

# Abfragesprachen

Auswahl von Teilen aus bestehenden Daten.  
*Wichtig für jede Nutzung und Integration*



# Strukturierungssprachen

“Data structuring languages (DSL)” oder “data serialization languages” bilden einen sehr groben Rahmen zur Formulierung von Daten.

- ▶ CSV
- ▶ XML
- ▶ INI
- ▶ JSON
- ▶ YAML
- ▶ RDF ohne Semantik
- ▶ ...

# Eigenschaften von Strukturierungssprachen

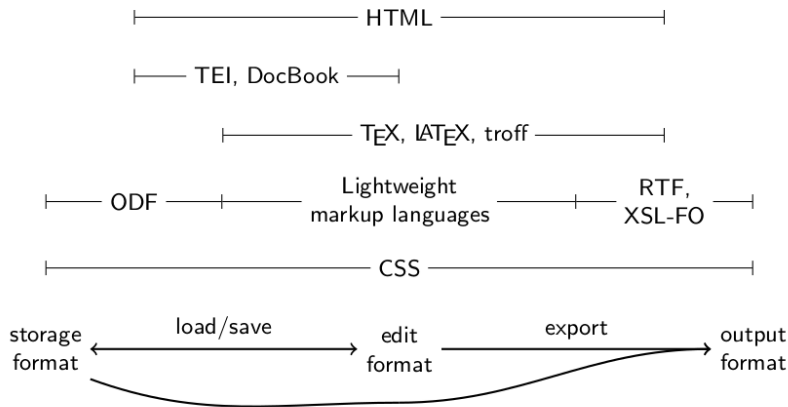
- ▶ Allgemeines Datenmodell
- ▶ Datentypen
  - ▶ Zahlen, Zeichenketten, Boolean...
  - ▶ Listen
  - ▶ Komplexere Typen (= eigene Formate)
- ▶ Syntax (mit Varianten)

# Allgemeine Datenmodelle

- ▶ Hierarchie (XML)
- ▶ Tabelle (CSV)
- ▶ Netzwerk (RDF)

Mischformen möglich durch Datentypen

# Auszeichnungssprachen





# Kodierungen

- ▶ Zeichen (ASCII, Unicode)
- ▶ Zahlen
- ▶ Identifier-Systeme

# Kodierungen

encoding	hexadecimal	binary
US-ASCII	—	—
ISO 646 DK/NO/SE	5D	1011101
EBCDIC CP37 etc.	67	01100111
Mac OS Roman	81	10000001
Allegro-DOS/IBM437	8F	10001111
NeXTSTEP	86	10000110
ISO 8859-1	C5	11000101
ANSEL (MARC-8) combining ° + A	EA 41	11101010 01000001

# Kodierungen

encoding	symbols
named HTML entity	&Aring;
XML character entity	&#xc5;; &#xC5;; &#197;; A&#x030A; ...
Swedish 6 dot Braille	 pattern P16 (Unicode U+2821)
Eurobraille 8 dot	 pattern P34567 (Unicode U+287C)
Transcription	Aa
Morse code (å = à, no case)	. - - - -

# Zusammenfassung



# Zusammenfassung Datensprachen

- ▶ Modellierungs-Sprachen (UML, ERM...)
- ▶ Schema-Sprachen (RDF Schema, XML Schema, RegExp...)
  - ▶ Abfragesprachen (SQL, XPath...)
- ▶ Datenstrukturierungssprachen (CSV, XML, JSON...)
  - ▶ Auszeichnungssprachen (HTML, TEI, Markdown...)
- ▶ Kodierungen (Unicode, ASCII, Binärcode...)

*Frage: Wo sind die meisten Probleme bei der Integration?*

# Zusammenfassung Daten

- ▶ Meist eher implizit behandelt
- ▶ Verschiedene Auffassungen
  - ▶ Daten als Fakten
  - ▶ Daten als Beobachtungen
  - ▶ Daten als Dokumente

# Weiterer Art der Gruppierung

Daten / Metadaten / Content

- ▶ Hängt mit Datensprachen und Auffassungen zusammen!

- Ballsun-Stanton, Brian. 2012. „Asking About Data: Exploring Different Realities of Data via the Social Data Flow Network Methodology“. Dissertation, University of New South Wales.
- Floridi, Luciano. 2005. „Is Information Meaningful Data?‘ *Philosophy and Phenomenological Research* 70 (2): 351–70.  
<http://philsci-archive.pitt.edu/archive/00002536/>.
- Naur, Peter. 1966. „The science of datalogy“. *Communications of the ACM* 9 (7): 485.
- Shannon, Claude Elwood. 1948. „A mathematical theory of communication“. *Bell Systems Technical Journal* 27: 379–423, 623–56.
- Simsion, Graeme. 2007. *Data Modeling Theory and Practise*. Technics Publications.
- Voß, Jakob. 2013. „Was sind eigentlich Daten?‘ *LIBREAS. Library Ideas*, Nr. 23. <http://libreas.eu/ausgabe23/02voss/>.