

Semantische Datenintegration

Identität und Identifizierung

Jakob Voß

Hochschule Hannover

2017-04-01

Identität und Entitäten

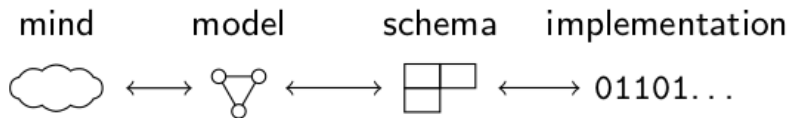
Grundannahme

- ▶ Daten allein haben keine Semantik
- ▶ Semantische Daten beziehen sich auf etwas
 - ▶ Werte (Zahlen, Namen...)
 - ▶ Dinge (Konzepte, Entitäten...)

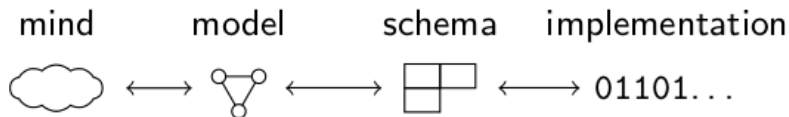
Daten-Werte

- ▶ Zahlen oder Zeichenketten
- ▶ Anzahl, Größe, Datum, Namen. . .
- ▶ Datentyp mit Kodierung
 - ▶ Integer, Unicode. . .
 - ▶ Mathematisch definierte Mengen
 - ▶ Wertebereiche und Einheiten!
- ▶ Vorstellung von Daten als Fakten oder Beobachtungen

Daten-Dinge



Daten-Dinge



- ▶ Im Kopf: Dinge
- ▶ Im Modell: Entitäten
- ▶ Im Schema: Identifier

Grundlage: Kent (1978), Chapter 1

Was ist ein Ding?

- ▶ Einheit
 - ▶ Was ist “ein Ding”?
- ▶ Gleichheit
 - ▶ Wann sind zwei Dinge das Gleiche?
 - ▶ Wann ist ein Ding nicht mehr das Gleiche?
- ▶ Kategorien
 - ▶ Von welcher Art ist ein Ding?

Entitäten-Einheit

- ▶ Beispiel: ein “Buch”
 - ▶ Ein Werk
 - ▶ Eine Auflage
 - ▶ Ein Exemplar
 - ▶ “nichtperiodische Publikationen mit einem Umfang von 49 Seiten oder mehr” (UNESCO)
- ▶ Worte und Konzepte sind vage!

Eintitäten-Gleichheit: Identitätsprinzip

1. Wenn Dinge nicht unterscheidbar sind, sind sie identisch
2. Identische Dinge sind nicht unterscheidbar

Entitäten-Gleichheit: Probleme

- ▶ Dinge erfüllen mehrere Rollen
 - ▶ z.B. Bibliothek als Gebäude, Sammlung, Organisation
- ▶ Dinge ändern sich

Welche Eigenschaften sind für ein Informationssystem relevant?

Beispiel: Schiff des Theseus

Music [\[edit \]](#)

[Sugababes](#), a [British](#) band,^[16] "were formed in 1998 [..] but one by one they left, till by September 2009 none of the founders remained in the band; each had been replaced by another member, just like the planks of Theseus's boat."^[17] The three original members reunited in 2011 under the name [Mutya Keisha Siobhan](#), with the "original" Sugababes still in existence.^{[12][18]}

https://en.wikipedia.org/wiki/List_of_Ship_of_Theseus_examples

Entitäten-Kategorien

- ▶ Legen fest was ein Ding ist
- ▶ Einteilung in (meist disjunkte) Mengen
- ▶ Klassifikation

Beispiel: Emporio celestial de conocimientos benévolos

... los animales se dividen en (a) pertenecientes al Emperador, (b) embalsamados, (c) amaestrados, (d) lechones, (e) sirenas, (f) fabulosos, (g) perros sueltos, (h) incluidos en esta clasificación, (i) que se agitan como locos, (j) innumerables, (k) dibujados con un pincel finísimo de pelo de camello, (l) etcétera, (m) que acaban de romper el jarrón, (n) que de lejos parecen moscas.

(Borges 1952)

Beispiel: Classification and its Consequences

- ▶ International Classification of Diseases (ICD)
- ▶ Südafrikanische Rassenklassifikation

(Bowker und Star 1999)

Umsetzung von Kategorien

- ▶ Disjunkte Datentypen
 - ▶ “Person”
 - ▶ “Organisation”
 - ▶ “Tätigkeit”
 - ▶ ...
- ▶ Implementierung
 - ▶ SQL-Tabellen
 - ▶ RDF-Klassen
 - ▶ Formate
 - ▶ ...

Zusammenfassung: Fallstricke bei der Datenintegration

- ▶ Daten beziehen sich auf Werte oder Dinge
- ▶ Dinge sind nicht eindeutig (Einheit)
- ▶ Dinge ändern sich (Gleichheit)
- ▶ Dinge ordnen die Welt (Kategorien)

... Notoriamente no hay clasificación del universo que no sea arbitraria y conjetural. La razón es muy simple: no sabemos qué cosa es el universo.

(Borges 1952)

Beispiel: Data and Reality

Zusammenstellung in einem integrierten Informationssystem:

<https://www.wikidata.org/wiki/Q25625532>

Kent (1978)

Identifizier

Motivation

- ▶ Umsetzung von Entitäten in Daten
- ▶ Hilfreiche Strategien
 - ▶ Identifier-Systeme
 - ▶ Normdateien

Identifizier-Systeme

Beispiel: ISBN

| | |
|---------------------|------------------------|
| ISBN-10 with hyphen | 1-4909-3186-4 |
| ISBN-10 with space | 1 4909 3186 4 |
| plain ISBN-10 | 1490931864 |
| EAN | 9781490931869 |
| EAN barcode aligned | 9 78149 093186 9 |
| ISBN-13 with hyphen | 978-1-4909-3186-9 |
| plain ISBN-13 | 9781490931869 |
| URN-ISBN | URN:ISBN:1-4909-3186-4 |

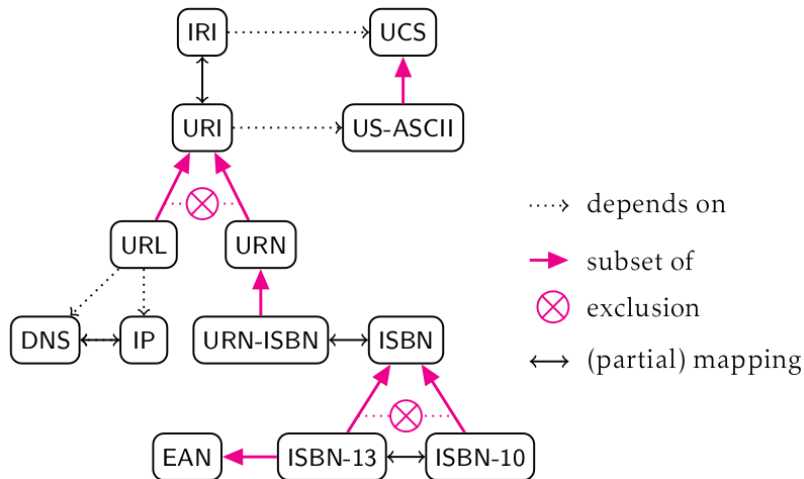
Identifizier-Heterogenität auf Syntaxebene

Bestandteile von Identifier(-Systemen)

Daten (Zeichenkette, Zahl... = *Identifier*) verweisen auf ein *Ding*

- ▶ In welchem Kontext (*Wo?*)
- ▶ Durch welche Vereinbarung oder Autorität (*Wer?*)
- ▶ Wo festgelegt und abrufbar (*Wie?*)

Beziehungen zwischen Identifier-Systemen



Anforderungen an Identifier-Systeme

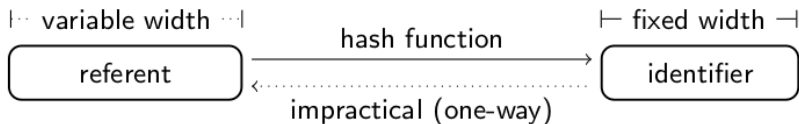
- ▶ Eindeutige Identifier
- ▶ Dauerhafte Identifier
- ▶ Überschaubare (kurze) Identifier
- ▶ Lesbare/Geordnete/Verteilte/Strukturierte Identifier
- ▶ Ausführbare Identifier (z.B. URL)
- ▶ *Verlässliche Autoritäten und Infrastruktur!*

Beispiel: <http://purl.org/>

Identifizier-Namensräume

| identifier | local | qualifier | syntax | description |
|----------------|-----------|-----------|-------------|------------------------|
| Frankfurt/Main | Frankfurt | Main | <i>L/Q</i> | city name |
| Dublin, Ohio | Dublin | Ohio | <i>L, Q</i> | city name |
| US-OH | OH | US | <i>Q-L</i> | ISO 3166-2 area code |
| std::set | set | std | <i>Q::L</i> | C++ identifier |
| rdf:type | type | rdf | <i>Q:L</i> | URI reference in RDF |
| 10.1000/182 | 1000/182 | 10 | <i>Q.L</i> | DOI as specific Handle |
| sgn-US | US | sgn | <i>Q-L</i> | IANA language & subtag |

Sonderfall: Hashcodes



Kryptologische Hashfunktion

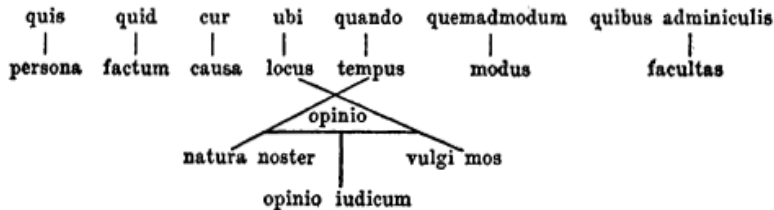
Beispiel: SHA1 8624bcdae55baeef00cd11d5dfcfa60f68710a02

Normdateien

- ▶ aka Stammdaten
- ▶ *Beispiele?*

Möglichkeiten und Grenzen von Identifizier-Systemen

Worauf beziehen wir uns?



Victorinus über Cicero und Hermagoras nach D. W. Robertson (1946)

Wer? Was? Warum? Wo? Wann? Wie? Womit?

Identität für lösbare Fälle

Eher lösbare Identitäten

- ▶ Zeitangaben (*Wann?*)
- ▶ Orte (*Wo?*)
- ▶ Personen (*Wer?*)
- ▶ Mittel und Formen (*Womit?*) → teilweise

Identifizier für Zeitangaben

- ▶ Genaue Zeitangaben: Daten als Fakten
 - ▶ ISO 8601 (YYYY-MM-DDThh:mm:ss)
- ▶ Erweiterte Zeitangaben: EDTF, 2012/2014
<https://www.loc.gov/standards/datetime/>
 - ▶ Intervalle (2017-04-01/open)
 - ▶ Ungenauigkeit (1984?, 1984~, 198x)
 - ▶ Fehlende Angaben (199u)
 - ▶ Jahreszeiten und große Jahreszahlen
 - ▶ Listen von Zeitangaben

Syntax prüfbar durch erweiterte reguläre Ausdrücke


PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS *THE* CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013.II.27. $27\frac{1}{2}$ -13 2013.158904109
MMXIII-II-XXVII MMXIII $\frac{\text{LVII}}{\text{CCCLXV}}$ 1330300800
 $((3+3) \times (11+1) - 1) \times 3 / 3 - 1 / 3^3$ 2013
10/1101/1101 02/27/20/13 $\begin{matrix} 2 & 3 & 1 & 4 \\ 0 & 1 & 2 & 3 & 7 \\ & 5 & 6 & 7 & 8 \end{matrix}$ 

Normdateien für Zeitangaben

Zeitangaben in RDF mit **PeriodO** <http://perio.do/>

- ▶ Name(n) und Beschreibung
- ▶ Quellenangaben(n)
- ▶ Zeitliche Eingrenzung
- ▶ Räumliche Eingrenzung

Beispiel: <http://n2t.net/ark:/99152/p0qp9rs3drk>

Identifizier für Orte

- ▶ Geographische Koordinaten
- ▶ Verweise auf Ortsdatenbanken (mit Koordinaten)
- ▶ Gebäude u.A. Strukturen

Kontinentalplatten bewegen sich einige cm pro Jahr

Normdateien für Orte


- ▶ ISO 3166
- ▶ Offizielle Kennziffern/Codes für Gemeinden u.Ä.
- ▶ GeoNames
- ▶ OpenStreetMap
- ▶ Gazetteers
- ▶ ...

Wikidata kennt fast 200 Properties für Orts-IDs


Beispiel: Pleiades

A major Etruscan city of North Etruria.

Canonical URI for this page:

<https://pleiades.stoa.org/places/403292> 

Representative Point (Latitude, Longitude):

43.401604, 10.8600625 



Locations:

- DARE Location (750 BC - AD 640)
- DARMC location 18992 (750 BC - AD 640)

Names:

- Οὐολατέρρα (Ovolatérra; 30 BC - AD 300)
- ΦΕΙΛΑΘΡΑ, velaθri (Velathri; 750 BC - AD 640)
- Volaterrae (750 BC - AD 640)
- Volterra (modern)



Makes a connection with:

- Etruria/Tuscia (Attested dates needed)



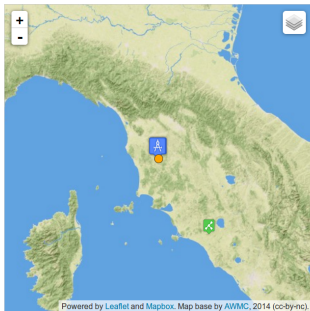
Has a connection with:

- Roman amphitheater at Volaterrae (Attested dates needed)
- Etruscan acropolis at Volaterrae (Attested dates needed)
- Roman bath at Volaterrae (Attested dates needed)
- Roman theater at Volaterrae (Attested dates needed)

Place type:

urban area, settlement

References:



Powered by Leaflet and Mapbox. Map base by AWMC, 2014 (cc-by-nc).

Show place in AWMC's Antiquity À-la-carte, Google Earth, or Pelagios' Peripleo.

Show area in GeoNames, Google Maps, or OpenStreetMap.



Theater at Volterra (I)
by Isawnyu.

[2 other related photos...](#)

Use this tag in Flickr to
mark depictions of this
place's site(s):

[pleiades:depicts=403292](#)

or this one to mark
objects found here:

[pleiades:findspot=403292](#)

Related Content from Pelagios

Velathri/Volaterrae

Epigraphic Database
Heidelberg (29);
American Numismatic
Society (26); University
of Graz (2); Inventory
of Greek Coin Hoards
(1); Fasti Online (1)

[Pelagios Datasets](#)

Zeiten und Orte

- ▶ Orte sind eine 4-dimensionale Kartoffeln
 - ▶ Datenmodell CRM_{geo} (Niccolucci und Hermon 2016)
- ▶ Nicht trivial, aber Referenzrahmen in der physischen Welt

Personen

Wikidata kennt über 500 Properties für Personen-Identifizierung

[https://www.wikidata.org/wiki/Wikidata:
List_of_properties/Person/Authority_control](https://www.wikidata.org/wiki/Wikidata:List_of_properties/Person/Authority_control)

```
SELECT ?p ?pLabel ?pDescription {  
  ?p wdt:P31 wd:Q19595382 .  
  SERVICE wikibase:label {  
    bd:serviceParam wikibase:language "de,en"  
  }  
}
```

Aufgabe: Finde Nicht-Personen die eine Personen-ID in Wikidata haben

Identität für unlösbare Fälle

Eher nicht lösbare Identitäten

- ▶ Allgemeine Konzepte (*Was?*)
- ▶ Prozesse (*Wie?*)
- ▶ Ursachen (*Warum?*)

Allgemeine Konzepte

- ▶ Wissensorganisationssysteme
 - ▶ Glossare
 - ▶ Klassifikationen
 - ▶ Thesauri
 - ▶ Ontologien
 - ▶ ...

<http://bartoc.org/>

Prozesse und Ursachen

- ▶ Netzwerke von weiteren Entitäten
- ▶ Dynamisch
- ▶ Nicht vollständig beschreibbar (Halteproblem)
- ▶ Behandlung wie
 - ▶ Allgemeine Konzepte
 - ▶ Dokumente/Werke

Identifizierung bei der Datenintegration

- ▶ Gleiche Standards
- ▶ Konkordanzen und Mappings
- ▶ Entity Recognition/Reconciliation

Konkordanzen und Mappings

- ▶ Wenn möglich 1-zu-1
 - ▶ Verschiedene Rollen?
 - ▶ Teil-Ganzes?
 - ▶ Grundproblem: Dinge \leftrightarrow Entitäten
- ▶ Automatische Verfahren
 - ▶ Labelbasiert
 - ▶ Instanzbasiert
 - ▶ ...

—→ *mehr siehe nächste Einheit*

Beispiel: Mix'n'match

- ▶ Tool zur Bearbeitung von Wikidata
 - ▶ <https://tools.wmflabs.org/mix-n-match/>
- ▶ Beispiele für Properties
 - ▶ GCD series ID
<https://www.wikidata.org/wiki/Property:P3589>
 - ▶ ISIL ID
<https://www.wikidata.org/wiki/Property:P791>

Entity Recognition und Mapping

- ▶ Objektidentifikation
- ▶ Heuristiken
 - ▶ Sprachtechnologie
 - ▶ Hash-Werte
 - ▶ Edit-Distance
 - ▶ Vektorraum-Retrieval
 - ▶ ...

Beispiel: OpenRefine

► Features

- Datenbereinigung
- Gruppierung und Clustering
- Reconciling: Abgleich von Datensätzen mittels externer API

► Tutorial

- Demo unter <http://vimeo.com/142641953>
- Kurs (30 Minuten) unter <https://data-lessons.github.io/library-openrefine/>
- <http://www.mnyslc.org/fellows/2017/03/17/using-openrefine-to-reconcile-name-entities/>

► Reconciliation APIs

- <https://tools.wmflabs.org/openrefine-wikidata/>
- <http://lobid.org/organisations/reconcile>
- <http://refine.codefork.com/reconcile/viaf>
- <https://vivo.tib.eu/fis/reconcile>
- ...

Zusammenfassung

- ▶ Was ist ein Ding?
 - ▶ Einheit, Gleichheit, Kategorien
 - ▶ Festlegung durch Modellierung
- ▶ Identifier-Systeme
 - ▶ Notwendige Festlegung von Beziehungen
 - ▶ Verschiedene Anforderungen und Komplexitäten
 - ▶ Wann, Wo, Wer...
 - ▶ Was, Wie, Warum...
 - ▶ Einfache vs. schwierige Fälle
 - ▶ Sonderfall Hashcodes
 - ▶ Normdateien helfen
- ▶ Integration durch Standards, Mappings und Objektidentifikation

Quellen dieser Folien: <https://github.com/hshdb/MWM-317-02/>

Borges, Jorge Luis. 1952. „El Idioma Analítico de John Wilkins“. In *Otras inquisiciones (1937-1952)*, 139–44. Buenos Aires: Sur.

Bowker, Geoffrey C., und Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. The MIT Press.

D. W. Robertson, Jr. 1946. „A Note on the Classical Origin of ‚Circumstances‘ in the Medieval Confessional“. *Studies in Philology* 43 (1): 6–14.

Kent, William. 1978. *Data and Reality. Basic assumptions in data processing reconsidered*. North-Holland.

Niccolucci, Franco, und Sorin Hermon. 2016. „Representing gazetteers and period thesauri in four-dimensional space-time.“ *International Journal on Digital Libraries* 17 (1): 63–69. doi:10.1007/s00799-015-0159-x.