Grundlagen von Datenformaten

DINI AG KIM Workshop 2021

Jakob Voß

28. April 2021

Daten

Was ist das?

Was ist das?

Bitte raten unter menti.com
Code 5234 8088
oder
https://www.menti.com/k5zd3cbuzh

Was ist das?

- Eine Antilope
- Eine Addax-Antilope
- Eine Abbildung einer Addax-Antilope
- Ein Digitalisat einer (älteren) Abbildung einer Addax-Antilope
- Eine Bilddatei
- ...

Interpretation erfordert Hintergrundwissen

Verschiedene Sichten auf Daten

- Daten als Fakten (“So sieht eine Addax aus”)
- Daten als Beobachtungen (“Abbildung einer Addax”)
- Daten als **Dokumente** (“Bilddatei”)

Daten-Dokumente entschlüsseln

- Was steht in den Daten (Inhalt: Semantik)
  - Antilope, Abbildung...
- Wo kommt das Dokument her? (Kontext: Pragmatik)
  - Identifier, Schnittstellen, Verarbeitung...
- Wie sind Daten strukturiert? (Struktur: Syntax)
  - JPEG-Datei, 790x878 Pixel...

Entschlüsselung erfordert Wissen über Datenformate

Welche Datenformate kennen Sie?

Nächste Frage unter menti.com

Code 5234 8088

oder https://www.menti.com/k5zd3cbuzh

Einordnung von Datenformaten

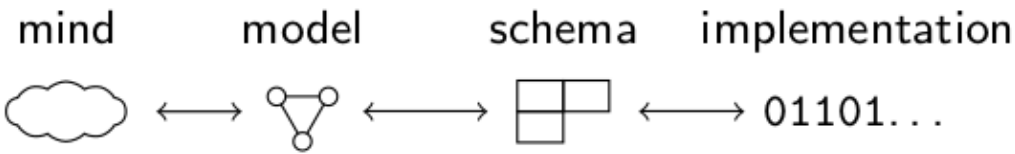
Grobe Übersicht nach Anwendung (1/2)

- Dokumentformate:
  HTML, TEI, Office, Markdown, Bild- & Audio-Formate...
- (bibliographische) **Metadatenformate**:
  MARC, MODS, LIDO, EAD, PICA...

Grobe Übersicht nach Anwendung (2/2)

- **Strukturierungssprachen**:
  CSV, JSON, RDF...
- **Schemasprachen**:
  Reguläre Ausdrücke, XSD, JSON Schema...
- **Abfragesprachen**:
  SQL, XPath, XQuery, CSS Selector...
- Datenmodelle:
  BIBFRAME, CIDOC-CRM, Dublin Core...

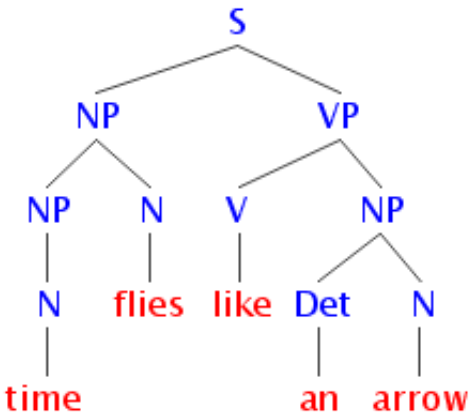
Datenmodelle



- Hinter Daten steht immer (implizit oder explizit) mindestens ein Modell
- Datenformate bewegen sich zwischen Modell und Implementierung

Eine “Grammatik” der Datenformate

- Lehre vom Bau einer Sprache, ihren Formen, Gesetzmäßigkeiten und Funktionen im Satz
- Elemente wie Wortarten, Satzbau, Regeln...
- Verschiedene Sprachen und Dialekte
- Unabhängig von der Bedeutung



Datenstrukturierungssprachen

Ein Datensatz (ohne Datenformat)

Name	Lebensdaten
Douglas Noël Adams	1952-2001

Erkennen Sie Modell und Schema?

Ein Datensatz (CSV)

```
name,dates
Douglas Noël Adams,1952-2001
```

Ein Datensatz (YAML)

```yaml
name: Douglas Noël Adams
dates: 1952-2001
```

Ein Datensatz (JSON)

```
{
  "name": "Douglas Noël Adams",
  "dates": "1952-2001"
}
```

Ein Datensatz (XML)

```
<name>Douglas Noël Adams</name>
<dates>1952-2001</dates>
```

Datenstrukturierungssprachen

- JSON, YAML, CSV, XML, RDF...
- Allgemeine Struktur in der sich fast alles ausdrücken lässt
- Je nach Sprache einiges besser als anderes
- Vergleich: Deutsch, Englisch, Spanisch, Japanisch...
- Konkrete Datenformate sind eher “Fachsprachen”

Praxisbeispiel

Ein JSON-Datensatz

```
{
  "name": "Douglas Noël Adams",
  "dates": "1952-2001"
}
```

Ein JSON-Datensatz

```
{
  "xxxx": "xxxxxxxxxxxxxxxxx",
  "xxxxx": "xxxxxxxxx"
}
```

Anatomie eines JSON-Datensatz

Syntax	Struktur
`"..."`	Unicode-Zeichenketten
`{...}`	Key-Value-Paare (“Object”)

JSON hat noch mehr, ist aber überschaubar!

Praxis mit dem JSON Editor

https://jsoneditoronline.org/beta/

```
{
  "name": "Douglas Noël Adams",
  "dates": "1952-2001"
}
```

Aufgabe in Kleingruppen á 5 Personen

1. Video an Ton an, bitte laut gemeinsam denken!
2. Datensatz eingeben
3. Weitere Daten hinzufügen (z.B. Alter, Identifier…)
4. Feedback: Menti Code 4936 6360 oder https://www.menti.com/5r2gnu61md

Feedback

...

Lessons Learned

- Passende Daten-Editoren helfen ungemein
- Ohne Syntax-Highlighting ist das Leben trist und grau
  - Weitere Werkzeuge: Code-Formatierung, Linter
- Manche Syntax-Elemente sind irrelevant
- Syntax-Elemente sind nicht der Inhalt...
- ...sondern Gerüst für eine Datenstruktur (z.B. JSON)

Wo ist der Inhalt?

```
{
  "xxxx": "xxxxxxxxxxxxxxxxxx",
  "xxxxx": "xxxxxxxxx"
}
```

Wo ist der Inhalt?

```
{
  "xxxx": "xxxxxxxxxxxxxxxxx",
  "xxxxx": "xxxxxxxxx"
}
```

- `xxx..`: Unicode-Zeichenketten
- Unicode ist auch nur ein Datenformat mit Modell, Schema und Implementierungen
- Daten haben keinen Inhalt

Es kommt auf die Betrachtungsebene an

Zeichenkette →JSON-Struktur →konkretes JSON-Format →... →Bedeutung

```
{
  "xxxx": "xxxxxxxxxxxxxxxxx",
  "xxxxx": "xxxxxxxxx"
}
```

```
{
  "name": "Xxxxxxx Xxxx Xxxxx",
  "dates": "0000-0000"
}
```

Datenformat: Menge von Daten mit einigen Beschränkungen und Freiheitsgraden

Daten-Schemas

Was ist ein Schema?

- Beschränkungen und Freiheitsgrade über einer Datenstrukturierungssprache
- Beispiel: JSON-Dokument mit notwendigen und optionalen Feldern
- Oft gibt es keine *expliziten* Schemas sondern
  - (Best-?)Practice
  - Standards
  - Anwendungen
  - ...
- Besser: formale Schemas in einer Schemasprache

Beispiel: JSON-Schema

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "type": "object",
  "properties": {
    "name":  { "type": "string" },
    "dates": { "type": "string" }
  }
}
```

Beispiel: JSON-Schema

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "type": "object",
  "properties": {
    "name":  { "type": "string" },
    "dates": {
      "type": "string",
      "pattern": "^[0-9]{4}(-[0-9]{4})?$"
    }
  }
}
```

Reguläre Ausdrücke: Schemas für Zeichenketten

Übung im JSON-Editor (ohne beta!)

https://jsoneditoronline.org/

Options > JSON Schema > Url:

https://raw.githubusercontent.com/jakobib/diniagkim2021/main/json-schema.json

Aufgabe:

1. Validieren Sie das Beispiel-Dokument
2. Ändern Sie das Dokument so, dass es keine Syntaxfehler hat aber dem Schema nicht mehr entspricht

Feedback

...

Übung: Finde die Fehler

Menti-Code 4936 6360 oder https://www.menti.com/5r2gnu61md

```
{
  "name": "Douglas NoÃ«l Adams"
  "dates" "1952-2001",
}
```

Wo sind Fehler und auf welcher Datenebene?

Übung: Finde die Fehler

```
{
  "name": "Kari Nordmann",
  "date": "1942-2019 ",
  "age": "76"
}
```

Wo sind Fehler und auf welcher Datenebene?

Wann ist ein Datensatz korrekt?

- Syntax (kaputt oder nicht kaputt)
- Struktur (valide oder fehlerhaft)
- Inhalt (tja...)

Was kann ein Datenformat festlegen?

- Syntax: Datenstrukturierungssprachen
- Struktur: Mittels Schemas
- Inhalt: Standards und Erfassungsregeln (nicht vollautomatisierbar)

Abfragesprachen

Was sind Abfragesprachen?

- Kleine Datenformate (“Sprachen”) um Abfragen zu formulieren
- Ergebnis: Teilmenge von vorhandenen Daten
- Abfragen über eine Datenstrukturierungssprache
  - XPath für XML-Daten
  - SPARQL für RDF-Daten
  - SQL für Relationale Datenbanken
  - ...
- APIs sind oder beinhalten ebenfalls Abfragesprachen

Abfrage von JSON-Daten

Verschiedene Abfragesprachen und Möglichkeiten (leider keine vollständig etabliert)

- JSON Path
- JSON Pointer
- jq
- ...

Übung: https://jqplay.org/

Zusammenfassung

Daten entschlüsseln

- Verschiedene Sichten auf Daten sind möglich und relevant
  - Fakten vs. Beobachtungen vs. Dokumente
  - Inhalt vs. Struktur vs. Kontext
  - Kodierungsebenen (z.B. Bytes, Struktur, Format…)
- Erfahrung im Umgang mit Datenformaten (Data Literacy)

Datenstrukturierungssprachen

- JSON, XML, CSV, RDF...
- Stellen Elemente zur Strukturierung bereit
  z.B. Zeichenketten, Verschachtelung, Key-Value-Paare...
- Folgen eigenen Paradigmen
  z.B. Tabelle, Baustruktur, Graph...
- Eigenes Ökysystem von Werkzeugen z.B. Editoren
- Sagen nichts über konkretes Format und Inhalt aus

Daten-Schemas

- Formale Standards was in einem Format erlaubt und erfordert ist
- Schemasprachen für jeweilige Strukturierungssprachen
  - XML Schema (XSD) für XML
  - JSON Schema für JSON
  - Avram für MARC/PICA
  - ...
- Ermöglichen automatische Validierung
- Ohne Validierung sind Fehler vorprogrammiert!

Abfragesprachen

- Teilmenge von Date in einer Datenstrukturierungssprache
- Geht mit jeder Programmiersprache, besser zielgerichtete Sprachen
  - XPath für XML
  - jq oder JSON Pointer für JSON
  - ...

Allgemeine Datenwerkzeuge

- **Erstellung:** Editoren
- **Validierung:** Validatoren (Syntax- & Schema-Ebene)
- **Abfrage:** Abfragesprachen, Datenbanken...

Mehr zu Datenformaten...

Übung macht die Meister:in!

https://format.gbv.de

Feedback

Bitte Feedback zur Veranstaltung!

http://etherpad.lobid.org/p/kimws21-Feedback-Hands-On-Tutorial_Datenformate

Anhang

Wo kommt die Antilope her?

Bertuch & Bertuch (1824): Bilderbuch für Kinder, Band 11, Seite 195
https://doi.org/10.11588/diglit.3218#0197

Wo kommt die Antilope her?

International Image Interoperability Framework (IIIF)

Datenformat für Metadaten (Seitenaufteilung etc.):
https://digi.ub.uni-heidelberg.de/diglit/iiif/bertuch1824bd11/manifest.json

Abfrageformat IIIF Image API https://digi.ub.uni-heidelberg.de/iiif/2/bertuch1824bd11%3A195.jpg/1150,1450,790,850/790,/0/default.jpg

`{scheme}://{server}{/prefix}/{identifier}/{region}/{size}/{rotation}/{quali`