Metadaten für multimodale Corpora

Bernhard Jackl
jackl@phonetik.uni-muenchen.de

09. April 2015

We kill people based on metadata.

– General Michael Hayden,
former director of the NSA and the CIA

http://www.nybooks.com/blogs/nyrblog/2014/may/10/we-kill-people-based-metadata/

Arten von Daten

- Primärdaten (Rohdaten)
  - Audiodaten
  - Video
  - Bilder
  - Time-Series (z. B. Motion Capture Daten)
- Sekundärdaten
  - $\Rightarrow$ Annotationen
- Metadaten

Metadaten

Wozu Metadaten?

- Corporaerzeugung zeitaufwendig & teuer
- Corpora ohne Metadaten schwer auffindbar
- „Passt dieses Corpus auf meine Fragestellung?“
- verhindert unnötige „Kopien“ von Corpora
- verhindert Wegwerfcorpora
- Aber: Metadaten allein helfen nicht, wenn nicht öffentlich

Metadaten (Cont.)

Beispiele

- Aufnahmesessions
- Teilnehmer (Actors)
- Primärdaten
- Sekundärdaten
- Corpus
- ...

Metadaten (Cont.)

Beispiele für Sessions

- Session ID (!)
- Aufnahmedatum & -zeit (!)
- Aufnahmeort
- Aufnahmebedingung
- Vorkommende Sprachen
- Genre
- ...

Metadaten (Cont.)

Beispiele für Actors

- Actor ID (!)
- Geburtsdatum (!)
- Geschlecht (!)
- Sprachen (Mutter-, Fremdsprachen, Dialekte)
- Informationen über Eltern
- Sprachstörungen
- Spezielle Attribute
  - Raucher (y/n)
  - Händigkeit (l/r/b)
  - Brille, Bart, Piercing (y/n)
  - Blut-/Atemalkoholkonzentration
  - ...

Metadaten (Cont.)

Beispiele zu Mediendateien

- Formate (MIME-Type) (!)
- Technische Daten (!)
- Länge (Bei Audio & Video)
- Frame Elemente (bei Time Series)
- Auflösung (Bei Video & Bildern)
- Aufnahmehardware/-software (Rechner, Mikrophon, ...)
- ...

Metadaten (Cont.)

Beispiele zu Annotationen

- Formate (plain text, binär)
- Encoding (UTF-8, ASCII, ...)
- Art der Annotation (orthogr., phonet., phonol., ...)
- automatisch/manuell erstellt
- ...

Metadaten (Cont.)

Beispiele für Corpus

- Corpus Titel
- Besitzer
- zugehöriges Projekt
- Enstehungszeitraum
- Ansprechpartner
- Gesamtumfang (GB, Stunden)
- Zugangsbeschränkung (public, academic, restricted, ...)
- Validierung?
- ...

Metadatenformate

Eigene

- Tabellenformat
  - \+ auch von Menschen gut les- u. editierbar
- XML
  - \+ erlaubt Verschachtelung
  - \+ gut maschinenlesbar
  - – schlecht menschenlesbar
  - – sehr verbose
- rel. Datenbank (PostgreSQL, MySQL, etc.)
  - \+ leichte Umformung/Export

Metadatenformate (Cont.)

Standards (Auswahl)

- Dublin Core
- OLAC (Open Language Archives Community)
- IMDI (ISLE MetaData Initiative, *veraltet*)
- CMDI (Component MetaData Initiative)

Metadatenformate (Cont.)

Dublin Core

- Dublin Core Metadata Initiative (DCMI)
- ursprünglich 15 *core elements*
  - identifier
  - format
  - type
  - ...
- sehr allgemeine Elemente (für jegliche Medien geeignet)

Metadatenformate (Cont.)

CMDI

- Einführung von Profilen
  - bestehen aus Components (die selbst aus weiteren C. und atomaren Elements bestehen)
  - registriert in der ComponentRegistry
  - Begriffe definiert in ISOcat
- wenn abwärtskompatibel können Profile erweitert werden
- Profile des BAS:
  - media-corpus-profile (v1.1)
  - media-session-profile (v1.2)
- Problem: Vielzahl an Profilen, Gefahr der Zersplitterung

CMDI-Erstellung

- Einfaches Corpus
  - wenige Sessions oder Actors
  - wenige Mediendateien
    - ⇒ CMDI mithilfe eines Editors erstellen
- Großes Corpus
  - große Menge an Sessions, Actors, Mediendatein ...
  - Metadaten bereits in einem strukturierten Format
    - ⇒ Batch-CMDI-Erstellung

CMDI-Erstellung (Cont.)

Editoren

- COMEDI (http://clarino.uib.no/comedi)
  - WebApp
  - benötigt Login über z. B. CLARIN IdP
  - noch in der Testphase
  - zum betrachten und editieren einzelner Dateien
- Arbil (https://tla.mpi.nl/tools/tla-tools/arbil/)
  - Java webstart oder Install für Windows/Mac/Linux
  - Viewer & Editor für Hierarchien
  - lokal & remote Daten

CMDI-Erstellung (Cont.)

COALA

- Konvertierung aus Zwischenformat nach CMDI
- für media-session-profile & media-corpus-profile
- nur für Corpora mit Multimediadaten, nicht Zeitungsartikel, o. ä.

CMDI-Erstellung (Cont.)

COALA – Art des Zwischenformates

- 5 Tabellen
  1. Actors
  2. Sessions
  3. Bundles (einzelne Aufnahme, bei uns: ein Prompt)
  4. MediaFiles (Signaldateien, Teil eines Bundles)
  5. WrittenResources (Annotationen, Teil eines Bundles oder einer ganzen Session)
- Können erstellt werden:
  - per Hand (Excel, OpenOffice.org, ...)
  - mit einem (Shell-) Script
  - als Export aus einer Datenbank

CMDI-Erstellung (Cont.)

COALA – WebApp

- https://clarin.phonetik.uni-muenchen.de/BASWebServices/#/services/Coala
- frei verfügbar
- Detaillierte Anleitung
- Beispiele dort verfügbar
- Videotutorial (https://youtu.be/EaIHujLkOdc)

Kurze Demo

Take Home Message

- Metadaten sind wichtig
- besser zuviel, als zuwenig speichern
- Verschiedene Formate
- CLARIN verwendet vorwiegend CMDI
- Erstellung mit Editoren möglich

Vielen Dank
für Ihre Aufmerksamkeit!