Federated Content Search

Making (legacy) lexicographical data interoperable

Axel Herold

Berlin-Brandenburgische Akademie der Wissenschaften
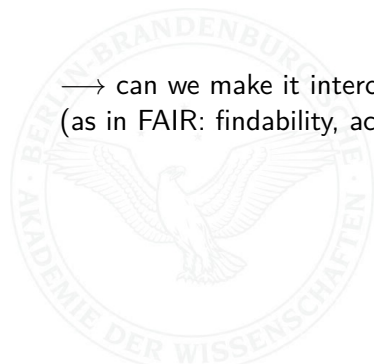
February 17th, 2023

Data models

Heterogeneity of lexical models

- ▶ legacy data from times before standardization
- ▶ tailor-made lexicographical models
- ▶ community specific (sub)standards
- ▶ old versions of standards

$\longrightarrow$ can we make it interoperable?
(as in FAIR: findability, accessibility, interoperability, reuse)

Data models

Example: DWDS

- historically (approx. 12 years ago): (pure) TEI representation of the „Wörterbuch der deutschen Gegenwartssprache" (WDG, 1964–1977, see `https://www.dwds.de/d/wdg`)
- edition of the WDG by senior lexicographers (from Grimm's dictionary)
- slowly emerging target entry model (*ad hoc*, not *a priori*)
- switch to DWDS specific XML dialect:
  - swifter and unrestricted model changes
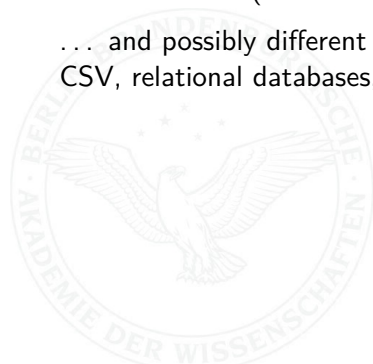 - readability for senior staff

Data models

Major design decisions

- ▶ purely lexicographical view (see talk on TEI modeling)
- ▶ elements contain information to be presented directly
- ▶ attributes carry metadata
- ▶ restricted datatypes and extensional enumerations wherever possible
- ▶ liberal re-use of common structures (such as usage labels, forms, comments)

Data models

More examples of idiosyncratic models

- ▶ Wahrig
- ▶ Neologismenwörterbuch (IDS)
- ▶ Wortgeschichten Digital (ZDL)
- ▶ GermaNet (Univ. Tübingen)

... and possibly different serializations (XML, JSON, RDF triples, CSV, relational databases, graph databases, ...)

Interoperability

How can we achieve it?

Interoperability

How can we achieve it?

- ▶ mapping data categories

Interoperability

How can we achieve it?

- ▶ mapping data categories
- ▶ using common agreed standard

Interoperability

How can we achieve it?

- ▶ mapping data categories
- ▶ using common agreed standard
- ▶ focus in subset of data categories

How can we achieve it?

- ▶ mapping data categories
- ▶ using common agreed standard
- ▶ focus in subset of data categories
- ▶ n different parsers $\longrightarrow$ pivot representation

Interoperability

Federated Content Search

- ▶ has existed for text corpora for quite some time
  (e. g. C4 corpora, CLARIN FCS)
- ▶ more challenging for lexical data
  (idiosyncratic tree structures as opposed to text annotation
  tiers or syntactic trees/graphs)
- ▶ needs to consider query and presentation
- ▶ Text+ project currently working on lexical FCS
  (https://www.text-plus.org/en/home/)
- ▶ prototype implementation at
  http://lexfcs-demo.wortschatz-leipzig.de

Interoperability

Text+: Lexical FCS

- ▶ data centers provide API endpoints for their resources
- ▶ common rest API for querying using CQL
  (Contextual Query Language,
  https://www.loc.gov/standards/sru/cql/)
- ▶ CQL can be extended for lexicological/lexicographical needs
  (e. g. specific fields)
- ▶ interfacing with recources (mapping, query translation) needs
  to be implemented by data providers
- ▶ aggregator allows to query several endpoints simultaneously
- ▶ common representation format (will be TEI Lex-0)