

# Data Modelling for Lexicographical Data

## A Short Introduction and Overview

Axel Herold

Berlin-Brandenburgische Akademie der Wissenschaften

February 23<sup>th</sup>, 2024



Data models – Why?

Data Structures and Representation Formats

Community Standard: TEI

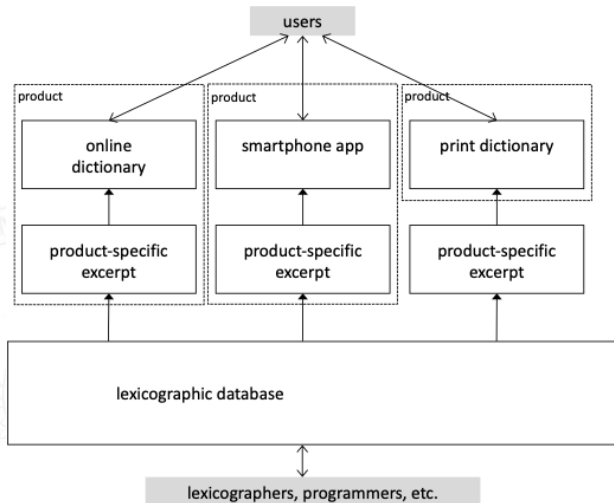
Traditional entry model(s) in TEI

TEI Lex-0

Hands-on

Literature

# Data models – Why?



## Structured data models foster

- ▶ precise querying and data retrieval  
(not just fulltext search)
- ▶ alignment across resources  
(linking to entries, senses, citations, ...)
- ▶ consistency checks, quality assurance
- ▶ transformations (restructuring, consistent rendering)
- ▶ fine-grained metadata (even within entries)
- ▶ informed training of classifiers, annotation programmes,  
machine learning algorithms

## Visual and Lexicographic Structures

Task: Describe the visual structuring of the entry and try to assign each part of the entry to a specific lexicographic structure.



**Fanclub**, der; -s; -s; auch: **Fanklub**; zu engl. club ‚Verein‘ (zu engl. clubbe ‚Keule, Knüppel‘); kein exklusives Etablissement, sondern: *Zusammenschluss von →Fans (1), dessen Hauptaufgabe in der Unterstützung des von ihnen vergötterten →Vereins u. seiner →Mannschaften besteht (z. B. durch →Fan-Gesänge, →Sprechchöre u. →Choreographien):* Der BVB hat 987 Fanclubs mit insgesamt 65.000 Mitgliedern. Auf den Sondertrikots sind die Namen aller offiziellen Fanclubs zu sehen, sie bilden in einer Grafik das Wort Danke (31.03.22; sport.sky.de).

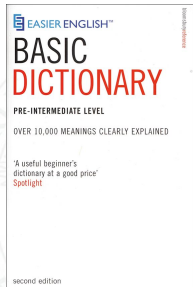
## Visual and Lexicographic Structures

- ▶ headword
- ▶ grammatical props
- ▶ „label“
- ▶ variant headword form
- ▶ etymology
- ▶ definition
- ▶ cross-references
- ▶ citation
- ▶ bibliographic info
- ▶ (... in general, there's much more)

**Fanclub**, der; -s; -s; auch: **Fanklub**; zu engl. club ‚Verein‘ (zu mengl. clubbe ‚Keule, Knüppel‘); kein exklusives Etablissement, sondern: *Zusammenschluss von →Fans (1), dessen Hauptaufgabe in der Unterstützung des von ihnen vergötterten →Vereins u. seiner →Mannschaften besteht* (z. B. durch →Fan-Gesänge, →Sprechchöre u. →Choreographien): Der BVB hat 987 Fanclubs mit insgesamt 65.000 Mitgliedern. Auf den Sondertrikots sind die Namen aller offiziellen Fanclubs zu sehen, sie bilden in einer Grafik das Wort Danke (31.03.22; sport.sky.de).

## Visual and Lexicographic Structures

Task: Describe the visual structuring of the entry and try to assign each part of the entry to a specific lexicographic structure.



**aid** /eid/ *noun* **1.** help, especially money, food or other gifts given to people living in difficult conditions ○ *aid to the earthquake zone* ○ *an aid worker* (NOTE: This meaning of **aid** has no plural.) □ **in aid of** in order to help ○ *We give money in aid of the Red Cross.* ○ *They are collecting money in aid of refugees.* **2.** something which helps you to do something ○ *kitchen aids* ■ **verb** **1.** to help something to happen **2.** to help someone

## Visual and Lexicographic Structures

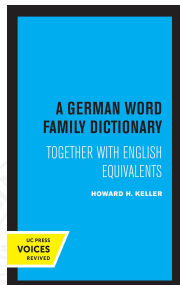
- ▶ phonological transcription
- ▶ „armchair“ examples
- ▶ multi word entities
- ▶ metamarks
- ▶ notes

**aid** /eid/ *noun* **1.** help, especially money, food or other gifts given to people living in difficult conditions ○ *aid to the earthquake zone* ○ *an aid worker* (NOTE: This meaning of **aid** has no plural.) □ **in aid of** in order to help ○ *We give money in aid of the Red Cross.* ○ *They are collecting money in aid of refugees.* **2.** something which helps you to do something ○ *kitchen aids* ■ **verb** **1.** to help something to happen **2.** to help someone



## Visual and Lexicographic Structures

Task: Describe the visual structuring of the entry and try to assign each part of the entry to a specific lexicographic structure.



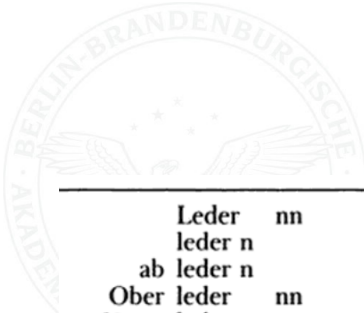
---

Leder	nn	leather
leder n		of leather; leathern, leathery, tough
ab leder n		wipe with chamois skin
Ober leder	nn	upper leather of shoe
Unter leder	nn	sole leather

---

## Visual and Lexicographic Structures

- ▶ multiple languages (de, en)
- ▶ morphological information
- ▶ nested entries



Leder	nn	leather
leder n		of leather; leathern, leathery, tough
ab leder n		wipe with chamois skin
Ober leder	nn	upper leather of shoe
Unter leder	nn	sole leather

## Commonly used Data Models/Formats

- ▶ XML
- ▶ Json
- ▶ wiki-based formats
- ▶ relational databases
- ▶ graph databases
- ▶ RDF triples
- ▶ (legacy) custom solutions

often accompanied by schemas and formal constraints

here, we'll focus on XML

## Commonly used Representation Formats

- ▶ printed pages
- ▶ PDF, EPUB, other (e-book) formats
- ▶ custom application specific formats
- ▶ common web based technology (HTML, JavaScript, ...)



## Brief overview

- ▶ operating since the (late) 1980s
- ▶ common standard for text representation in the DH for all sorts of printed or written documents, currently based on open X\* standards
  - ▶ prose, verse, drama
  - ▶ dictionaries
  - ▶ list, accounting data
  - ▶ scholarly editions
  - ▶ born-digital communication
  - ▶ ...
- ▶ *lots* of active projects rely on the TEI
- ▶ open collaboration and development through special interest groups
- ▶ <http://www.tei-c.org>, <https://github.com/TEIC>

## Traditional entry model(s) in TEI

- ▶ modeling  $\approx$  mapping of objects and their properties (and relations) onto symbolic representations (generally also abstraction)
- ▶ modeling lexical data (esp. in digitization of printed resources) is multi-layered modeling:
  - ▶ printed characters  $\longrightarrow$  codepoints (e.g. Unicode)
  - ▶ spacial relation of characters  $\longrightarrow$  words (tokens)
  - ▶ typographical properties  $\longrightarrow$  (hints as to) functions of words (tokens)
  - ▶ ...
- ▶ every level relies on interpretation and may introduce uncertainty
- ▶ alternative and even incompatible interpretations (and therefore conflicting models) are possible

## Traditional entry model(s) in TEI

different „views“ on lexical data:

**typographical** „the two-dimensional printed page, including information about line and page breaks and other features of layout“

**editorial** „the one-dimensional sequence of tokens which can be seen as the input to the typesetting process . . . “

**lexicographic** „... the underlying information represented in a dictionary, without concern for its exact textual form“

(TEI Guidelines, chapter 9)

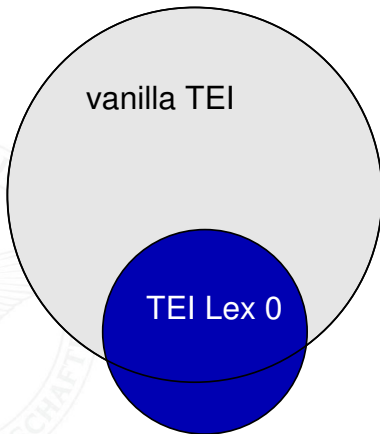
## TEI Lex-0

- ▶ international working group with strong ties to the TEI
- ▶ work started in 2016, supported by ENeL, Dariah, independent research institutes (currently cooperating with Elexis)
- ▶ general use-case: mapping typographic structures onto lexical structures
- ▶ not a chapter 9 replacement, rather a proposed baseline by way of a TEI customization
- ▶ aims at interoperability by
  - ▶ restricting some alternatives
  - ▶ streamlining some content models
  - ▶ closing and fixing vocabulary
  - ▶ ...
- ▶ quite some proposals went upstream already



## Relation between vanilla TEI and TEI Lex-0

„not a chapter 9 replacement, rather a proposed baseline“  
for lexicographic information



## TEI Lex-0

- ▶ frequent group meetings
- ▶ open collaboration on GitHub:  
<https://github.com/DARIAH-ERIC/lexicalresources>  
(take a look!)
- ▶ close collaboration with the TEI consortium
- ▶ TEI Lex-0 is still work in progress (version 0.9.3)

## Let's get our hands dirty!

- ▶ task: plain text → TEI Lex-0 mark-up
- ▶ XML editor (actually, any text editor will do)
- ▶ XML validation
- ▶ we will ignore metadata for now
- ▶ data can be found in the course's GitHub project
  - ▶ plain text entries
  - ▶ TEI Lex-0 schema (0.9.3, Relax NG)
- ▶ correct TEI data made available after the course

## This introduction is based on

- ▶ Herold/Meyer/Müller-Spitzer (2016): Datenmodellierung. In: Klosa/Müller-Spitzer (ed.): Internet-Lexikografie. Ein Kompendium. Berlin/Boston 2016
- ▶ Herold/Meyer/Wiegand (to appear): Data Modelling. In: Klosa (ed.): Internet Lexicography. to appear

## Dictionaries

- ▶ Burkhardt, Armin: Wörterbuch der Fußballsprache. Hildesheim: Arete <sup>2</sup>2022
- ▶ Easier English Basic Dictionary. Bloomsbury Publishing <sup>3</sup>2009
- ▶ Keller, Howard H.: A German Word Family Dictionary. University of California Press 1978

# Thank you for participating!

