

# TEI: Vertiefung am Beispiel von "Dictionaries"



---

Prof. Dr. Christof Schöch

---

Modul Auszeichnungssprachen  
MSc. Digital Humanities, Universität Trier

---



# Überblick

1. Die Module der TEI-Guidelines
2. Das Modul 9: Dictionaries
3. Weitere Perspektiven: TEI Lex-0

# (1) Die Module der TEI-Guidelines

# Die 23 Module der TEI

## Front Matter

### [Title](#)

- i. [Releases of the TEI Guidelines](#)
- ii. [Dedication](#)
- iii. [Preface and Acknowledgments](#)
- ⊕ iv. [About These Guidelines](#)
- ⊕ v. [A Gentle Introduction to XML](#)
- ⊕ vi. [Languages and Character Sets](#)

## Back Matter

- ⊕ Appendix A [Model Classes](#)
- ⊕ Appendix B [Attribute Classes](#)
- ⊕ Appendix C [Elements](#)
- ⊕ Appendix D [Attributes](#)
- ⊕ Appendix E [Datatypes and Other](#)

### [Macros](#)

- ⊕ Appendix F [Bibliography](#)
- ⊕ Appendix G [Deprecations](#)
- ⊕ Appendix H [Prefatory Notes](#)
- Appendix I [Colophon](#)

## Text Body

- ⊕ 1 [The TEI Infrastructure](#)
- ⊕ 2 [The TEI Header](#)
- ⊕ 3 [Elements Available in All TEI Documents](#)
- ⊕ 4 [Default Text Structure](#)
- ⊕ 5 [Characters, Glyphs, and Writing Modes](#)
- ⊕ 6 [Verse](#)
- ⊕ 7 [Performance Texts](#)
- ⊕ 8 [Transcriptions of Speech](#)
- ⊕ 9 [Dictionaries](#)
- ⊕ 10 [Manuscript Description](#)
- ⊕ 11 [Representation of Primary Sources](#)
- ⊕ 12 [Critical Apparatus](#)
- ⊕ 13 [Names, Dates, People, and Places](#)
- ⊕ 14 [Tables, Formulæ, Graphics and Notated Music](#)
- ⊕ 15 [Language Corpora](#)
- ⊕ 16 [Linking, Segmentation, and Alignment](#)
- ⊕ 17 [Simple Analytic Mechanisms](#)
- ⊕ 18 [Feature Structures](#)
- ⊕ 19 [Graphs, Networks, and Trees](#)
- ⊕ 20 [Non-hierarchical Structures](#)
- ⊕ 21 [Certainty, Precision, and Responsibility](#)
- ⊕ 22 [Documentation Elements](#)
- ⊕ 23 [Using the TEI](#)




## TEI Guidelines: Inhaltsverzeichnis

# Projektabhängiges TEI Schema



- Motto: "So präzise wie möglich, so flexibel wie nötig"
- Funktionen des Schemas
  - Konsistente Kodierung erlauben
  - Datenmodell formulieren und kommunizieren
  - Verarbeitungsroutinen entwickeln
- Werkzeug "Roma" für ODD / Schema
  - <https://roma.tei-c.org/>
  - Beta: <https://romabeta.tei-c.org/>











# Roma: TEI-ODD-Tool

Roma - ODD Customization (jTEI input customization)  
Version 0.3.0

 SETTINGS  START OVER  DOWNLOAD

☒ Elements ☐ Attribute Classes ☐ Model Classes ☐ Datatypes

 alphabetically  

<input type="checkbox"/> <b>entry</b> contains a single structured entry in any kind of lexical resource, such as a dictionary or lexicon.	 (dictionaries)
<input type="checkbox"/> <b>entryFree</b> (unstructured entry) contains a single unstructured entry in any kind of lexical resource, such as a dictionary or lexicon.	 (dictionaries)
<input type="checkbox"/> <b>etym</b> (etymology) encloses the etymological information in a dictionary entry.	 (dictionaries)
<input type="checkbox"/> <b>form</b> (form information group) groups all the information on the written and spoken forms of one headword.	 (dictionaries)
<input type="checkbox"/> <b>gen</b> (gender) identifies the morphological gender of a lexical item, as given in the dictionary.	 (dictionaries)
<input type="checkbox"/> <b>gram</b> (grammatical information) within an entry in a dictionary or a terminological data file, contains grammatical information relating to a term, word, or form.	 (dictionaries)
<input type="checkbox"/> <b>gramGrp</b> (grammatical information group) groups morpho-syntactic information about a lexical item, e.g. pos, gen, number, case, or iType (inflectional class).	 (dictionaries)
<input type="checkbox"/> <b>hom</b> (homograph) groups information relating to one homograph within an entry.	 (dictionaries)
<input type="checkbox"/> <b>hyph</b> (hyphenation) contains a hyphenated form of a dictionary headword, or hyphenation information in some other form.	 (dictionaries) 

Neue Version von Roma

## (2) Das Modul 9: Dictionaries

# Wörterbucheintrag (print, annotiert)

**<p>**

**Lemma**

**Wortartangabe**

**kommentar zu grammatischem Form**

**Artikel**

**EINKUPPELMANÖVER** n. zuss. mit einkuppeln vb.  
 vorgang der verbindung zweier raumschiffe: [1979] laut  
 Tass wird das kosmosforschungsprogramm durch ein-  
 kuppelmanöver mit der raumstation Salut 6 fortgesetzt  
 gött. tagebl. 48. titels.

**EINKUPPELN** vb. maschinen  
 oder maschinenteile mittels einer kupplung verbinden: **Wortartangabe**  
 [1918] größere schiffe haben verschiedene steuerstationen,  
 die entsprechend eingekuppelt werden können **ARVAY**  
 seemannswesen 55. j. [1973] verzögernd auf die eingabe  
 wirkt endlich noch der eigene arbeitsrhythmus einzelner  
 eingabegeräte. ein kartenleser z. b. kann nicht zu jedem  
 beliebigen zeitpunkt seines eigenen arbeitsrhythmus  
 eingekuppelt werden **DÜRR** datenerfassung 57. j. jünger  
 meist absolut. bei kraftwagen durch loslassen des kupp-  
 lungspedals eine verbindung zwischen motor und getriebe  
 herstellen und das fahrzeug in bewegung setzen: [1928]  
 Roth machte eine vielsagende gebärde und ließ seinen  
 chauffeur einkuppeln **BRONNEN** film 205. j. [1966] wenn ich  
 jetzt einkupple und gas gebe **ZWERENZ** Casanova  
 461. j.

**Artikel**

**Wortartangabe**

**= Belegdatum**

**= Belegautor**

**[...] = Beleg**

**— = bibliograph. Nachweis; abgekürzt**

**<p>**

**definiertionsartige Bedeutungsangabe (mit  
 kommentar zu historischen gebrauch-  
 sänderungen)**

TCDH: 2DWB



# Wörterbucheintrag (online)


**D W D S** Der deutsche Wortschatz von 1600 bis heute. [Anmelden](#)

[Startseite](#) / [Wörterbuch](#) / [Rebstock](#) – Schreibung, Definition, Bedeutung, Synonyme, Beispiele

Rebstock

**Rebstock**, der


**Grammatik** Substantiv (Maskulinum)

**Aussprache** 

**Worttrennung** Reb-stock

**Wortzerlegung** ↗[Rebe](#) ↗[Stock](#)<sup>1</sup>

**Bedeutung** eWDG, 1974

 **landschaftlich** [Weinstock](#)

**Thesaurus** www.openthesaurus.de (08/2020)


**Botanik**

➤ **Synonymgruppe**

↗[Rebe](#) · Rebstock · ↗[Wein](#) · ↗[Weinrebe](#) · ↗[Weinstock](#)

**Worthäufigkeit**

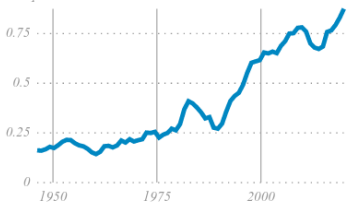
selten häufig



**Wortverlaufskurve**

[ab 1600](#) [ab 1946](#)

Frequenz / Mio Tokens



**Weitere Wörterbücher**

| [Deutsches Wörterbuch \(¹DWB\)](#) (1)

| [Deutsches Wörterbuch, Neubearbeitung \(²DWB\)](#) (0)

| [Wörterbuch der deutschen Gegenwartssprache \(WDG\)](#) (2)

DWDS, <https://www.dwds.de/>

# Herausforderungen

- Struktur(en)
  - Wörterbucheinträge sind meist stark strukturiert
  - Jedes Wörterbuch aber meist auf unterschiedliche Weise
- Perspektive(n)
  - (gedruckte) Wörterbücher haben starke typographische Konventionen
  - (digitale) Wörterbücher benötigen eine starke (semantische) Struktur

# Viele "Standards"!

## 3. Data Formats

- 11 In choosing a uniform encoding system for all ICLTT data, the department's staff surveyed data formats in use. Although most of the relevant dictionary productions of the recent past have relied on digital data and methods, there is little consensus on standards. A great number of divergent formats have coexisted: MULTILEX and GENELEX (GENERIC LEXicon) are systems that are associated with the Expert Advisory Group on Language Engineering Standards (EAGLES).<sup>4</sup> Other formats used in digital dictionary projects are OLIF (Open Lexicon Interchange Format),<sup>5</sup> MILE (Multilingual ISLE Lexical Entry),<sup>6</sup> LIFT (Lexicon Interchange Format),<sup>7</sup> OWL (Web Ontology Language)<sup>8</sup> and DICT (Dictionary Server Protocol),<sup>9</sup> the latter being an important dictionary delivery format (Faith 1997).
- 12 Another standard considered was ISO 1951 ("Presentation/representation of entries in dictionaries – requirements, recommendations and information"). Although this standard focuses on encoding the presentation of lexicographical data in dictionaries for human use in what is called LEXml (Lexicographical Markup Language), it seems that after a few years of existence only few publishing houses have been using this format (such as Langenscheidt, Munich) for their dictionary production line.
- 13 Last but not least, when looking for an encoding standard for machine readable dictionaries, ISO 24613:2008 ("Language resource management – Lexical markup framework (LMF)"), the ISO standard for natural language processing (NLP) and machine-readable dictionaries (MRD), must be considered. Recently, there have been discussions about the possibility of creating a TEI serialization of LMF (Romary 2010).

<https://doi.org/10.4000/jtei.522>

# TEI: Makrostruktur

```
<tei>
  <teiheader>
    ...
  </teiheader>
  <text>
    <front>...</front>

    <entry>...</entry>
    <entry>...</entry>
    <entry>...</entry>
    ...

    <back>...</back>
  </text>
</tei>
```

# Kernstück: <entry>

## 9.2 The Structure of Dictionary Entries

« 9.1 Dictionary Body  
and Overall Structure

» 9.3 Top-level  
Constituents of Entries

[Home](#)

A simple dictionary entry may contain information about the form of the word treated, its grammatical characterization, its definition, synonyms, or translation equivalents, its etymology, cross-references to other entries, usage information, and examples. These we refer to as the *constituent parts* or *constituents* of the entry; some dictionary constituents possess no internal structure, while others are most naturally viewed as groups of smaller elements, which may be marked in their own right. In some styles of markup, tags will be applied only to the low-level items, leaving the constituent groups which contain them untagged. We distinguish the class of *top-level constituents* of dictionary entries, which can occur directly within the [entry](#) element, from the class of *phrase-level* constituents, which can normally occur only within top-level constituents. The top-level constituents of dictionary entries are described in section [9.2.2 Groups and Constituents](#), and documented more fully, together with their phrase-level sub-constituents, in section [9.3 Top-level Constituents of Entries](#).

In addition, however, dictionary entries often have a complex hierarchical structure. For example, an entry may consist of two or more sub-parts, each corresponding to information for a different part-of-speech homograph of the headword. The entry (or part-of-speech homographs, if the entry is split this way) may also consist of senses, each of which may in turn be composed of two or more sub-senses, etc. Each sub-part, homograph entry, sense, or sub-sense we call a *level*; at any level in an entry, any or all of the constituent parts of dictionary entries may appear. The hierarchical levels of dictionary entries are documented in section [9.2.1 Hierarchical Levels](#).

- Haupt-Elemente: u.a. <form>, <gramGrp>, <etym>, <def>, <cit>, <usg>.
- ■ Unterelemente: jeweils eine Reihe von Elementen

# Einfaches Beispiel

```
<entry>
  <form>
    <orth>competitor</orth>
    <hyph>com|peti|tor</hyph>
    <pron>k@m"petit@ (r) </pron>
  </form>
  <gramgrp>
    <pos>n</pos>
  </gramgrp>
  <def>person who competes.</def>
</entry>
```

# Haupt-Elemente

- `<form>`: Wortform (u.a. Schreibung, Aussprache)
- `<gramGrp>`: grammatikalische Informationen (bspw. Wortart) – flexibler Ort
- `<def>`: Definition(en) / Bedeutung(en)
- `<etym>`: Etymologie (Herkunftsgeschichte)
- `<cit>`: Zitate / Belege / Verwendungsbeispiele
- `<usg>` (usage): Informationen zur Verwendung
- u.a.m.

# Unterelemente von <form>

- <orth>: Schreibweise
- <pron>: Aussprache
- <hyph>: Angabe der Trennstellen
- <syll>: Angabe der Silbengrenzen
- <gramGrp>: Angabe der Silbengrenzen (aber: innerhalb von <form>)



# Beispiel für <form>

```
<entry>
  <form>
    <orth>biryani</orth>
    <orth>biriani</orth>
    <pron notation="ipa">ˌbɪrɪˈaːnɪ</pron>
  </form>
  [...]
</entry>
```

# Unterelemente von <gramGrp>

- <gramGrp> kann als Unterelement von <entry>, <form>, <sense>, oder <cit> vorkommen
- Unter anderem mit den Informationen zu:
  - <pos> (part of speech): Wortart
  - <gen>: Gender (maskulin / feminin / neutrum)
  - <number>: Numerus (Singular / Plural)

# Beispiel für <gramGrp>

```
<entry>
  <form>
    <orth>isotope</orth>
  </form>
  <gramgrp>
    <pos>adj</pos>
  </gramgrp>

  [...]
</entry>
```

# Element <def>

- Für die Angabe der Wortbedeutung(en)
- Hat keine notwendigen Unterelemente
- Kann direkt eine Bedeutungsangabe enthalten
- Oder, bei mehreren Bedeutungen, jeweils innerhalb von <sense>

# Beispiel für <def>

```
<form>
  <orth>demigod</orth>
</form>
<sense n="1">
  <def>a being who is part mortal, part god.</def>
</sense>
<sense n="2">
  <def>a lesser deity.</def>
</sense>
<sense n="3">
  <def>a godlike person.</def>
</sense>
```

# Unterelemente von `<etym>`

- Oft teilweise unstrukturiert, zudem speziell für einzelne Bestandteile:
- `<lang>`: Sprache
- `<date>`: Datums- / Zeitangabe
- `<usg>` (usage): Angaben zur Verwendung
- `<mentioned>`: Für Wörter, die als solche genannt werden, statt benutzt zu werden
- `<gloss>`: Umschreibung oder Definition eines Wortes

# Beispiel für <etym>

```
<entry>
  <form>
    <orth>neuma</orth>
  </form>
  <etym>
    <lang>F</lang> fr. <lang>ML</lang>
    <mentioned>pneuma</mentioned>
    <mentioned>neuma</mentioned> fr. <lang>Gk</lang>
    <mentioned>pneuma</mentioned>
    <gloss>breath</gloss>
  </etym>
  <sense>
    <def>any of various symbols used in the notation of Gregorian
  </sense>
</entry>
```

# Unterelemente von <cit>

- <cit>: Ein Zitat als Beispiel oder Beleg für eine Bedeutung, mit einer Quellenangabe
- Üblicherweise mit einem @type mit Werten wie "example" oder "translation"
- Darin für den zitierten Wortlaut selbst:
  - <q> (quoted - typographisch markiert) oder
  - <quote> (quotation - unmarkiert)
- Und <bibl> für die Quellenangabe



# Beispiel für <cit>

```
<entry>
  <form>
    <orth>valeur</orth>
  </form>
  <cit type="example">
    <quote>La valeur n'attend pas le nombre des années</quote>
    <bibl>
      <author>Corneille</author>
    </bibl>
  </cit>
</entry>
```


# Strukturierung mehrerer <entry>s

```
<superentry>
  <entry n="1" type="hom">
    <form>
      <orth>mouse</orth>
    </form>
    <def>Small quadruped animal.</def>
    <!-- ... -->
  </entry>
  <entry n="2" type="hom">
    <form>
      <orth>mouse</orth>
    </form>
    <def>Pointing device for computers.</def>
    <!-- ... -->
  </entry>
</superentry>
```

# (3) Weitere Perspektiven

# Trierer Wörterbuchnetz

Kooperationspartner Hinweise Kontakt



## Wörterbuchnetz

© 2011 Trier Center for Digital Humanities, Universität Trier

... Je weiter ich in diesem Studium fortgehe, desto klarer wird mir der Grundsatz: daß kein einzelnes Wort oder Wörtchen bloß eine Ableitung haben, im Gegenteil jedes hat eine unendliche und unerschöpfliche. Alle Wörter scheinen mir gespaltene und sich spaltende Strahlen eines wunderbaren Ursprungs, daher die Etymologie nichts tun kann, als einzelne Leitungen, Richtungen und Ketten aufzufinden und nachzuweisen, soviel sie vermag. Fertig wird das Wort nicht damit.

Jacob Grimm an Savigny. 20. Apr. 1815

Stichwort in allen Wörterbüchern suchen

### Die Wörterbücher und Nachschlagewerke

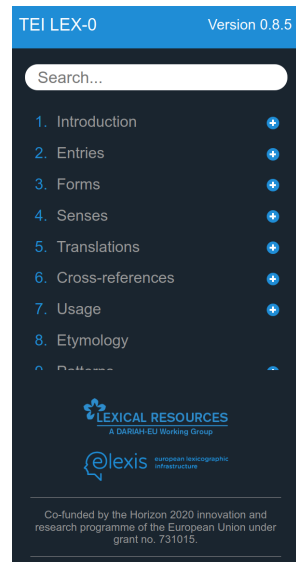
(Mit \* gekennzeichnete Wörterbücher sind externe Angebote.)

<b>AWB</b> Althochdeutsches Wörterbuch*	<b>Adelung</b> Grammatisch-Kritisches Wörterbuch der Hochdeutschen Mundart	<b>BMZ</b> Mittelhochdeutsches Wörterbuch von Benecke, Müller, Zarncke
<b>DFD</b> Digitales Familiennamenwörterbuch Deutschlands*	<b>DRW</b> Deutsches Rechtswörterbuch*	<b>DWB</b> Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm
<b>EisWB</b> Wörterbuch der elsässischen Mundarten	<b>FWB</b> Frühneuhochdeutsches Wörterbuch*	<b>Finckel</b> Findebuch zum mittelhochdeutschen Wortschatz
<b>GWB</b> Goethe-Wörterbuch	<b>Hederich</b> Gründliches mythologisches Lexikon von Benjamin Hederich	<b>Idiotikon</b> Schweizerisches Idiotikon / Wörterbuch der schweizerdeutschen Sprache*
<b>LLU</b> Lexicon der Luxemburger Umgangssprache*	<b>LWB</b> Luxemburger Wörterbuch*	<b>Lexa</b> Mittelhochdeutsches Handwörterbuch von Matthias Lexa
<b>LmL</b> Lexicon musicum Latinum medi aevi	<b>LothWB</b> Wörterbuch der deutsch-lothringischen Mundarten	<b>MHDBDB</b> Mittelhochdeutsche Begriffsdatenbank*
<b>MLW</b> Mittellateinisches Wörterbuch	<b>MWB</b> Mittelhochdeutsches Wörterbuch*	<b>Meyers</b> Meyers Großes Konversationslexikon
<b>NLexa</b> Nachträge zum Mittelhochdeutschen Handwörterbuch von Matthias Lexa	<b>NRWB</b> Nachträge zum Rheinischen Wörterbuch	<b>PfWB</b> Pfälzisches Wörterbuch
<b>REDE</b> Regionalsprache.de*	<b>RhWB</b> Rheinisches Wörterbuch	<b>SHW</b> Südhessisches Wörterbuch*
<b>WLM</b> Wörterbuch der Luxemburgischen Mundart*	<b>Wander</b> Deutsches Sprichwörter-Lexikon von Karl Friedrich Wilhelm Wander	<b>WdW</b> Wörterbuch der deutschen Winzersprache*

© 2011 Trier Center for Digital Humanities, Universität Trier  
Home | Impressum | Kontakt

<http://www.woerterbuchnetz.de/>

# TEI Lex-0



## TEI Lex-0

— A baseline encoding for lexicographic data

### 1. Introduction

#### 1.1. TEI Lex-0 in a nutshell

TEI Lex-0 is both a technical specification and a set of community-based recommendations for encoding machine-readable dictionaries. It is rooted in the [Guidelines of the Text Encoding Initiative](#) (TEI) and delivered as a customization of the TEI schema.

Following the spirit of TEI Analytics, developed in the context of the MONK project ([Zillig 2009](#)), TEI Lex-0 aims at establishing a baseline encoding and a target format to facilitate the interoperability of heterogeneously encoded lexical resources. This is important both in the context of building lexical infrastructures as such ([Ermolaev and Tasovac 2012](#)) and in the context of developing generic TEI-aware tools such as dictionary viewers and profilers.

- Austauschformat für die TEI-Kodierung von Wörterbüchern
- Stärker standardisiert als in den Guidelines
- Zielformat für Kodierung, Vergleichbarkeit und Tool-Entwicklung

# Beispieleintrag nach Lex-0

```
<entry xml:id="OALD.competitor" type="mainEntry" xml:lang="en">
  <form type="lemma">
    <orth>competitor</orth>
    <hyph>com|peti|tor</hyph>
    <pron>k@m"petit@ (r)</pron>
  </form>
  <gramgrp>
    <gram type="pos">n</gram>
  </gramgrp>
  <sense xml:id="OALD.competitor.1">
    <def>person who competes.</def>
  </sense>
</entry>
```

- Beachte: @xml:id und @xml:lang
- <form> und <gram> mit @type

# Lektürehinweise

## Referenzlektüre

- "TEI Lex-0. A baseline encoding for lexicographic data" [Abschnitte 1-4]. *ELEXIS*. <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>

## Weitere Empfehlungen

- "9. Dictionaries", in: *Guidelines of the Text Encoding Initiative*, P5, Version 4.1.0, 2020. <https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html>
- Piotr Bański, Jack Bowers, Tomaz Erjavec. "TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms." *Electronic Lexicography in the 21st Century: Proceedings of ELex 2017 Conference*, Sep 2017, Leiden, Netherlands. <https://hal.inria.fr/hal-01757108>
- Gerhard Budin, Stefan Majewski and Karlheinz Mörth. "Creating Lexical Resources in TEI P5. A Schema for Multi-purpose Digital Dictionaries". *Journal of the Text Encoding Initiative*, 3, 2012. <https://doi.org/10.4000/jtei.522>

# Danke!

---

Lizenz: [Creative Commons Attribution \(CC BY\)](#), 2020.

---