



# Text Encoding Initiative

Vorlesung *Einführung in die Digital Humanities*  
MSc Digital Humanities | Wintersemester 2020/21

Prof. Dr. Christof Schöch

---





# Überblick

1. Die Text Encoding Initiative
2. Modul: "Default Text Structure"
3. Modul: "Core"
4. Modul: "TEI Header"
5. Beispiele

# (1) Die Text Encoding Initiative

# Was ist die Text Encoding Initiative?

- Ein Standard zur Kodierung von Texten
  - Guidelines und Schemata
  - Tools (Roma, OxGarage)
  - Wird seit 1987 entwickelt und gepflegt!
- Eine Institution zur Pflege des Standards
  - Mitglieder (Institutionen, Privatpersonen)
  - Board of Directors, (Technical) Council, SIGs
- Eine Community rund um den Standard
  - Nutzer:innen rund um die Welt
  - Members Meeting and Conference
  - Mailing-Listen
  - TEI Journal

## Relevanz der TEI: warum das alles?

- Die TEI ist ein de facto Standard
  - Langfristige Interpretierbarkeit
  - Langfristige Nachnutzung
  - Texte und ihre Metadaten
  - In XML formuliert

# Ziele der TEI

- Universelle Einsetzbarkeit erreichen
  - allen Sprachen
  - alle Schriftsysteme
  - aus allen Epochen
  - alle denkbaren Textsorten
  - verschiedene editorische Schulen
- Hilfestellung
  - für Anfänger: Anwendung der Guidelines
  - für Spezialisten: Anpassung, Weiterentwicklung der Guidelines
  - Ergebnis: flexibler, modularer, vielschichtiger Standard

# Geburtsstunde der TEI (1987)



(Poughkeepsie Meeting; Bildquelle: [Oxford University Computing Centre](#))

## Die TEI-Architektur bietet

- Begriffe, Definitionen, Hinweise für hunderte textkritische Phänomene
- Dadurch: enzyklopädische Kodifikation aktueller editorischer Praxis
- Eine Art, darüber nachzudenken, was "Text" ist / was eine "Edition" leisten soll

## Der TEI-Werkzeugkasten: Software

- Roma: Schema-Erstellen (TEI-C): <http://www.tei-c.org/Roma>
- Oxgarage: Dokument-Konversion (TEI-C):  
<https://oxgarage.tei-c.org/>
- Publikationstools: Versioning Machine, SADE, TEI Publisher uvm.
- Editoren: oXygen, Atom mit Plugins, jEdit mit Plugins, TextGridLab

## (2) Modul: "Default Text Structure"

# TEI Guidelines (Module 4)

The screenshot shows the TEI Guidelines P5 page. At the top is a blue header with the TEI logo and the text <Text Encoding Initiative>. Below the header is a dark blue navigation bar with links for Home, Guidelines, Activities, Tools, Membership, Support, About, and News. To the right of the navigation bar is a search bar containing "P5 Guidelines — English" and a "Search" button. The main content area has a light gray background. At the top of the content area is a section titled "P5: Guidelines for Electronic Text Encoding and Interchange" with the subtitle "Version 4.1.0. Last updated on 19th August 2020, revision b414ba550". On the left side of the content area is a "Table of contents" sidebar with a list of sections under "4.1 Divisions of the Body". The main content area starts with a section titled "4 Default Text Structure". It contains a paragraph about the default high-level structure for TEI documents, mentioning the `teiHeader` element and other modules like `textstructure` and `transcr`. Below this is a bulleted list of five items describing different types of encoded resources. At the bottom of the content area is a note about grouping multiple resources sharing the same metadata.

Table of contents

- 4.1 Divisions of the Body
- 4.2 Elements Common to All Divisions
- 4.3 Grouped and Floating Texts
- 4.4 Virtual Divisions
- 4.5 Front Matter
- 4.6 Title Pages
- 4.7 Back Matter
- 4.8 Module for Default Text Structure

« 3 Elements Available in All TEI Documents

» 5 Characters, Glyphs, and Writing Modes

Home

## 4 Default Text Structure

This chapter describes the default high-level structure for TEI documents. A full TEI document combines metadata describing it, represented by a `teiHeader` element, with the document itself, represented by one or more `text` elements or other elements taken from the `model.resource` class. That is, the `TEI` element is used to group together metadata about an encoded resource (in `teiHeader`, specified by the `header` module, which is fully described in chapter [2 The TEI Header](#)) with an encoded resource. Possible encoded resources are

- a logical transcription of a source document in a `text` element; the `text` element is specified along with its high-level constituents in the `textstructure` module and described in the remainder of the current chapter
- a diplomatic transcription of a source document in a `sourceDoc` element, which is specified in the `transcr` module and described in chapter [11 Representation of Primary Sources](#)
- an encoded representation of a text-bearing object as images in a `facsimile` element, which is also specified in the `transcr` module and described in chapter [11 Representation of Primary Sources](#)
- a collection of contextual information or annotations that provides more detail about another encoded resource (whether in the same or a different TEI document) in a `standOff` element, which is specified in the `linking` module and described in section [16.10 The standOff Container](#)
- a feature system declaration which can be used to declare the use of `fs` elements in the rest of the document, which is specified in the `iso-fs` module and described in section [18.11 Feature System Declaration](#)

In a case in which more than one resource related to the same source document share the same metadata, they may be grouped together in a `TEI` element following a single `teiHeader`.

<https://tei-c.org/release/doc/tei-p5-doc/en/html/DS.html>

# Makrostruktur: Grundlegende Elemente

- <text>
- <front>, <body>, <back>
- <group>
- <titlePage>
- <graphic>
- <div>

# Makrostruktur: Beispiel

```
<text xml:id="v147">
  <front>
    [...]
  </front>
  <group>
    <text xml:id="I1914-07-01">
      <body> [... first issue (1 July) ...] </body>
    </text>
    <text xml:id="I1914-07-15">
      <body> [... second issue (15 July) ...]</body>
    </text>
    [...]
  </group>
  <back> [... index, appendix ...] </back>
</text>
```

## (3) Das Modul "Core"

# TEI Guidelines (Module 3)

The screenshot shows the TEI Guidelines (Module 3) page. At the top is the TEI logo and the text "< Text Encoding Initiative >". Below the logo is a navigation bar with links: Home, Guidelines, Activities, Tools, Membership, Support, About, and News. A search bar contains the text "P5 Guidelines — English" and a "Search" button. The main content area has a title "P5: Guidelines for Electronic Text Encoding and Interchange" and a subtitle "Version 4.1.0. Last updated on 19th August 2020, revision b414ba550". On the left is a "Table of contents" sidebar with sections 3.1 through 3.13. The main content area starts with a section titled "3 Elements Available in All TEI Documents". It describes elements like paragraphs, punctuation, highlighting, and quotation. It follows with sections on lists, notes, graphics, reference systems, bibliographic citations, and an overview of the core module. It concludes with sections on the TEI header and default text structure.

**P5: Guidelines for Electronic Text Encoding and Interchange**  
Version 4.1.0. Last updated on 19th August 2020, revision b414ba550

Table of contents	3 Elements Available in All TEI Documents
3.1 Paragraphs 3.2 Treatment of Punctuation 3.3 Highlighting and Quotation 3.4 Simple Editorial Changes 3.5 Names, Numbers, Dates, Abbreviations, and Addresses 3.6 Simple Links and Cross-References 3.7 Lists 3.8 Notes, Annotation, and Indexing 3.9 Graphics and Other Non-textual Components 3.10 Reference Systems 3.11 Bibliographic Citations and References 3.12 Passages of Verse or Drama 3.13 Overview of the Core Module	<p>This chapter describes elements which may appear in any kind of text and the tags used to mark them in all TEI documents. Most of these elements are freely floating phrases, which can appear at any point within the textual structure, although they should generally be contained by a higher-level element of some kind (such as a paragraph). A few of the elements described in this chapter (for example, bibliographic citations and lists) have a comparatively well-defined internal structure, but most of them have no consistent inner structure of their own. In the general case, they contain only a few words, and are often identifiable in a conventionally printed text by the use of typographic conventions such as shifts of font, use of quotation or other punctuation marks, or other changes in layout.</p> <p>This chapter begins by describing the <code>p</code> tag used to mark paragraphs, the prototypical formal unit for running text in many TEI modules. This is followed, in section <a href="#">3.2 Treatment of Punctuation</a>, by a discussion of some specific problems associated with the interpretation of conventional punctuation, and the methods proposed by these Guidelines for resolving ambiguities therein.</p> <p>The next section (section <a href="#">3.3 Highlighting and Quotation</a>) describes a number of phrase-level elements commonly marked by typographic features (and thus well-represented in conventional markup languages). These include features commonly marked by font shifts (<a href="#">3.3.2 Emphasis</a>, <a href="#">Foreign Words, and Unusual Language</a>) and features commonly marked by quotation marks (<a href="#">3.3.3 Quotation</a>) as well as such features as terms, cited words, and glosses (<a href="#">3.3.4 Terms, Glosses, Equivalents, and Descriptions</a>).</p> <p>Section <a href="#">3.4 Simple Editorial Changes</a> introduces some phrase-level elements which may be used to record simple editorial interventions, such as emendation or correction of the encoded text. The elements described here constitute a simple subset of the full mechanisms for encoding such information (described in full in chapter <a href="#">11 Representation of Primary Sources</a>), which should be adequate to most commonly encountered situations.</p>
» 2 The TEI Header » 4 Default Text Structure	

<https://tei-c.org/release/doc/tei-p5-doc/en/html/CO.html>

# 50 Elemente, die man fast immer braucht

- Allgemeines
  - <p>, <pb/>, <said>, <head>, <foreign>
  - <date>, <name>, <time>
- Textsorten-Spezifisches
  - Verstext: <lg> (linegroup), <l> (line)
  - Theater: u.a.: <sp> (speech), <speaker>, <stage>
- Sonstiges
  - Bibliographische Angaben: <bibl>, <author>, <title>, <publisher>, <pubPlace>
  - Editorische Interventionen: u.a. <abbr> und <expan>, <corr> und <reg>

## Einige globale Attribute

- Einige Merkmale lassen sich auf (fast) alles anwenden
  - Identität
  - Sprache
  - Aussehen
- TEI bietet dafür "globale Attribute" an
  - @type - Klassifikation
  - @xml:id - "unique identifier"
  - @n – Name oder Nummer
  - @xml:lang – Sprache (nach ISO Standard)

# Beispiel: Versdrama

```
<sp>
  <speaker>Mephistopheles</speaker>
  <lg type="stanza">
    <l>Ich möcht' mich gleich dem Teufel übergeben,</l>
    <l>Wenn ich nur selbst kein Teufel wär'!</l>
  </lg>
</sp>
<sp>
  <speaker>Faust</speaker>
  <lg type="stanza">
    <l>Hat sich dir was im Kopf verschoben?</l>
    <l>Dich kleidet's, wie ein Rasender zu tob'en!</l>
  </lg>
</sp>
```

# Beispiel: Bibliographische Angabe

```
<biblFull>
  <titleStmt>
    <title>Envisioning Information</title>
    <author>Tufte, Edward R[olf]</author>
  </titleStmt>
  <extent>126 pp.</extent>
  <publicationStmt>
    <publisher>Graphics Press</publisher>
    <pubPlace>Cheshire, Conn. USA</pubPlace>
    <date>1990</date>
  </publicationStmt>
</biblFull>
```

## (4) Das Modul "teiHeader"

# TEI Guidelines (Module 2)

The screenshot shows the TEI Guidelines website. At the top is a blue header bar with the TEI logo and the text "< Text Encoding Initiative >". Below the header is a navigation menu with links for Home, Guidelines, Activities, Tools, Membership, Support, About, and News. A search bar contains the text "P5 Guidelines — English" and a "Search" button. The main content area has a title "P5: Guidelines for Electronic Text Encoding and Interchange" and a subtitle "Version 4.1.0. Last updated on 19th August 2020, revision b414ba550". On the left, there is a "Table of contents" sidebar with links to various sections like "Organization of the TEI Header", "File Description", etc. The main content area starts with a section titled "2 The TEI Header". It includes a detailed description of what the header is used for, followed by three numbered points explaining the types of descriptions it contains: a file description, an encoding description, and a text profile.

## 2 The TEI Header

This chapter addresses the problems of describing an encoded work so that the text itself, its source, its encoding, and its revisions are all thoroughly documented. Such documentation is equally necessary for scholars using the texts, for software processing them, and for cataloguers in libraries and archives. Together these descriptions and declarations provide an electronic analogue to the title page attached to a printed work. They also constitute an equivalent for the content of the code books or introductory manuals customarily accompanying electronic data sets.

Every TEI-conformant text must carry such a set of descriptions, prefixed to it and encoded as described in this chapter. The set is known as the *TEI header*, tagged [teiHeader](#), and has five major parts:

1. a *file description*, tagged [fileDesc](#), containing a full bibliographical description of the computer file itself, from which a user of the text could derive a proper bibliographic citation, or which a librarian or archivist could use in creating a catalogue entry recording its presence within a library or archive. The term *computer file* here is to be understood as referring to the whole entity or document described by the header, even when this is stored in several distinct operating system files. The file description also includes information about the source or sources from which the electronic document was derived. The TEI elements used to encode the file description are described in section [2.2 The File Description](#) below.
2. an *encoding description*, tagged [encodingDesc](#), which describes the relationship between an electronic text and its source or sources. It allows for detailed description of whether (or how) the text was normalized during transcription, how the encoder resolved ambiguities in the source, what levels of encoding or analysis were applied, and similar matters. The TEI elements used to encode the encoding description are described in section [2.3 The Encoding Description](#) below.
3. a *text profile*, tagged [profileDesc](#), containing classificatory and contextual information about the text, such as its subject matter, the situation in which it was produced, the individuals described by or participating in producing it, and so forth. Such a text profile is of particular use in highly

<https://tei-c.org/release/doc/tei-p5-doc/en/html/HD.html>

## Grundlegende Metadaten

- Identifikation der Ressource: "Um was handelt es sich?"
- Zuordnung von Verantwortlichkeiten: "Wer hat wann was gemacht?"
- Angabe der Quelle(n): "Was wurde woher genommen?"
- Publikations-Angaben: "Wie wird diese Ressource zugänglich gemacht?"
- Dokumentation der Kodierungspraxis: "Was bedeutet der Markup?"

# TEI Header: 4 Komponenten

- <fileDesc> (**file description**)
  - notwendig (als einzige Komponente)
  - bibliographische Beschreibung des Dokuments
- <encodingDesc> (**encoding description**)
  - Beziehung zwischen kodiertem Text und Quelle
- <profileDesc> (**text-profile description**)
  - Viele weitere Aspekte: Sprache, Kontext, etc.
- <revisionDesc> (**revision description**)
  - Fasst den Bearbeitungsverlauf des Dokuments zusammen.

## Obligatorische Teile der <fileDesc>

- <titleStmt> mit <title>
- <publicationStmt>: Aussage zum Status des Dokuments bzgl. Publikation
- <sourceDesc>: Beschreibung der Quelle

# Beispiel: minimaler Header

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>My 'Strange Meeting' document</title>
    </titleStmt>
    <publicationStmt>
      <p>An exercise for learning TEI.</p>
    </publicationStmt>
    <sourceDesc>
      <p>The primary resource of this file is <ref
target="http://www.oucs.ox.ac.uk/ww1lit/
collections/item/3350"> Strange Meeting</ref> from
Jon Stallworthy's edition, available on the WWI
Poetry Digital Archive. </p>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```

## Erweiterungen für das <titlestmt>

- Standard-Rollen
  - <author> - Autor des kodierten Texts oder des „born digital“ Dokuments
  - <editor> - Herausgeber des kodierten Texts
- Für Spezifischere Rollen
  - <respstmt> mit <resp> und <name>
  - <resp> - frei definierbare Aufgabe, bspw. "Transkription" oder "Kodierung der Named Entities"

# Beispiel <titleStmt>

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>Monster: eine XML-Edition</title>
      <author>Wir sind Helden</author>
      <editor>Franz Sahle</editor>
      <respStmt>
        <resp>Kodierung</resp>
        <name>Patrick Fischer</name>
      </respStmt>
    </titleStmt>
    [...]
  </fileDesc>
</teiHeader>
```

# Beispieldaten

# Brief aus der A.W.-Schlegel-Edition

 Digitale Edition der Korrespondenz August Wilhelm Schlegels [Version-10-20]

Suchbegriff  Erweiterte Suche Register Zeitstrahl

Suchergebnisse 6/76

August Wilhelm von Schlegel an Eduard Weber

Absendeort: Bonn (GND) · Empfangsort: Bonn (GND) · Datum: 04.09.1820

Editionsstatus: Neu transkribiert und ausgezeichnet; zweimal kollationiert ⓘ

Parallelansicht Zitierempfehlung Editorische Richtlinien Vollständige Metadaten XML PDF

Volltext Handschrift Digitalisat Handschrift

[1] Ich schicke Ihnen hiebey für zwey Bogen ins Reine geschriebenes Manuscript. Ich habe gestern darauf gewartet, daß Sie mich würden wissen lassen, es sey in der Druckerey so weit, daß angefangen werden könne zu setzen.

Ich bitte Sie, mir hierüber eine Zeile Nachricht zu geben –

ergebenst

AWvSchlegel

Mont. früh

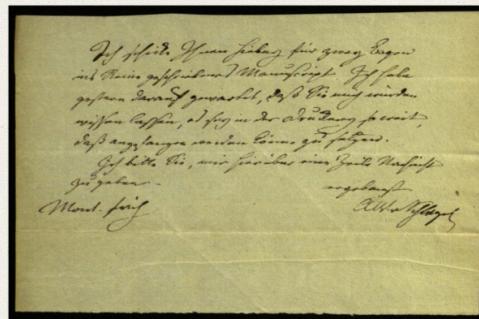
[2] Herrn  
Buchh. Weber

4/9 1820 v. Schlegel

Seite 1/2

Register  
Körperschaften ▾  
□ Bonner Universitäts-Buchdruckerei (GND)

Periodika ▾  
□ Indische Bibliothek. Eine Zeitschrift von August Wilhelm von Schlegel (GND)



<https://august-wilhelm-schlegel.de/briefedigital/letters/xml/7621>

# Roman aus ELTeC-eng (Gaskell)

The screenshot shows the GitHub repository page for COST-ELTeC / ELTeC-eng. The repository has 15 stars and 2 branches. The 'Code' tab is selected, showing a list of commits from user 'lb42'. The commits include adding report directories for levels 0, 1, and 2, changing language codes, and fixing header errors. The 'About' section describes the repository as containing TEI XML Sources for the English novel part of the ELTeC. It includes links to distantreading.github.io/eltec and a DOI for the collection. The 'Releases' section shows a release with 100 novels, the latest being 29 days ago. The 'Packages' section indicates no packages are published. The 'Contributors' section lists three contributors: lb42 Lou, CarolinOdebrecht, and christofs Christof Schöch.

Search or jump to... Pull requests Issues Marketplace Explore

COST-ELTeC / ELTeC-eng Watch 15 Star

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

master 2 branches 3 tags Go to file Add file Code

**lb42 add reports dir** 663193a 12 days ago 217 commits

- Reports add reports dir 12 days ago
- level0 add reports dir 12 days ago
- level1 add reports dir 12 days ago
- level2 change language code in md 2 years ago
- Makefile header errors 17 months ago
- README.md +current DOI 29 days ago

**README.md**

DOI 10.5281/zenodo.3462536

## ELTeC-eng

This is the English novel collection for the ELTeC, the European Literary Text Collection, produced by the COST Action Distant Reading for European Literary History (CA16204, <https://distant-reading.net>).

### Release notes

General information about ELTeC releases is available at <https://github.com/COST-ELTeC/ELTeC>.

The ELTeC-eng collection contains a total of 100 titles, 87 encoded at level 1, and 13 at level 0. The corpus composition criteria have been observed as far as possible, and a conscious effort has been made to maximise diversity in the types of novels included. Release 1.0.0 has the following DOI: <https://doi.org/10.5281/zenodo.4271630>

### About

TEI XML Sources for the English novel part of the ELTeC

distantreading.github.io/eltec

novel nineteenth-century tei-xml

xml literature

Readme

### Releases 3

Release with 100 novels. (Latest) 29 days ago

+ 2 releases

### Packages

No packages published Publish your first package

### Contributors 3

- lb42 Lou
- CarolinOdebrecht
- christofs Christof Schöch

[https://github.com/COST-ELTeC/ELTeC-eng/blob/master/level1/ENG18482\\_Gaskell.xml](https://github.com/COST-ELTeC/ELTeC-eng/blob/master/level1/ENG18482_Gaskell.xml)

# Abschluss

# Lektürehinweise

## Referenzlektüre

- Christof Schöch: "Ein digitales Textformat für die Literaturwissenschaften: die Richtlinien der Text Encoding Initiative und ihr Nutzen für Textedition und Textanalyse". *Romanische Studien* 4, 2016. <http://romanischedestudien.de/index.php/rst/article/view/58/517> (Open Access)

## Weitere Empfehlungen

- Lou Burnard: *What is the Text Encoding Initiative?* Marseille: Open Edition Press, 2014. <https://books.openedition.org/oep/679?lang=de> (Open Access)

Christof Schöch, 2020  
<http://www.christof-schoech.de>

---

Lizenz: Creative Commons Attribution 4.0