Text Encoding Initiative

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

Prof. Dr. Christof Schöch

Universität Trier

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

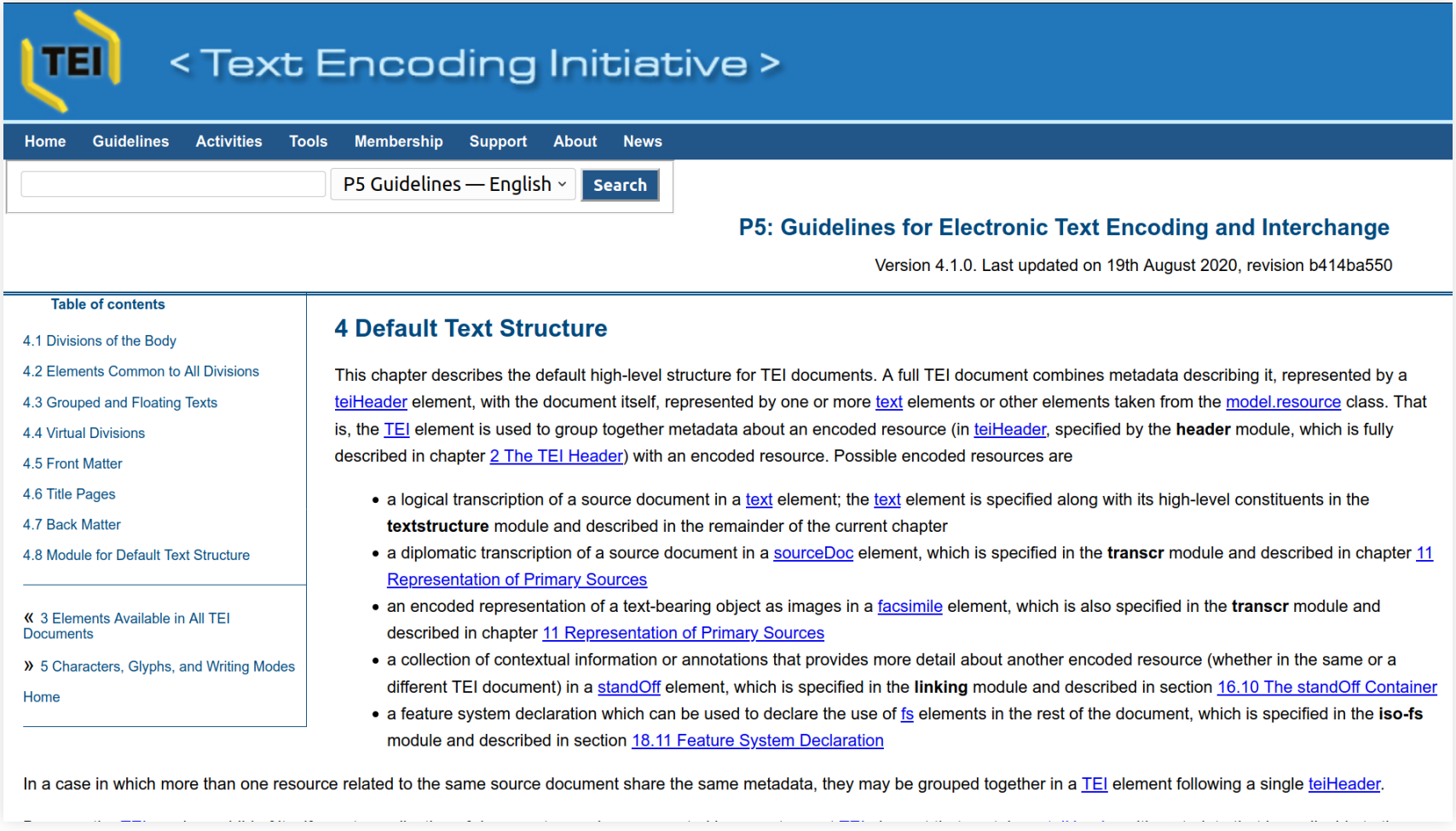
- Begriffe, Definitionen, Hinweise für hunderte textkritische Phänomene
- Dadurch: enzyklopädische Kodifikation aktueller editorischer Praxis
- Eine Art, darüber nachzudenken, was "Text" ist / was eine "Edition" leisten soll

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)



< Text Encoding Initiative >

Home Guidelines Activities Tools Membership Support About News

P5 Guidelines — English Search

P5: Guidelines for Electronic Text Encoding and Interchange

Version 4.1.0. Last updated on 19th August 2020, revision b414ba550

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

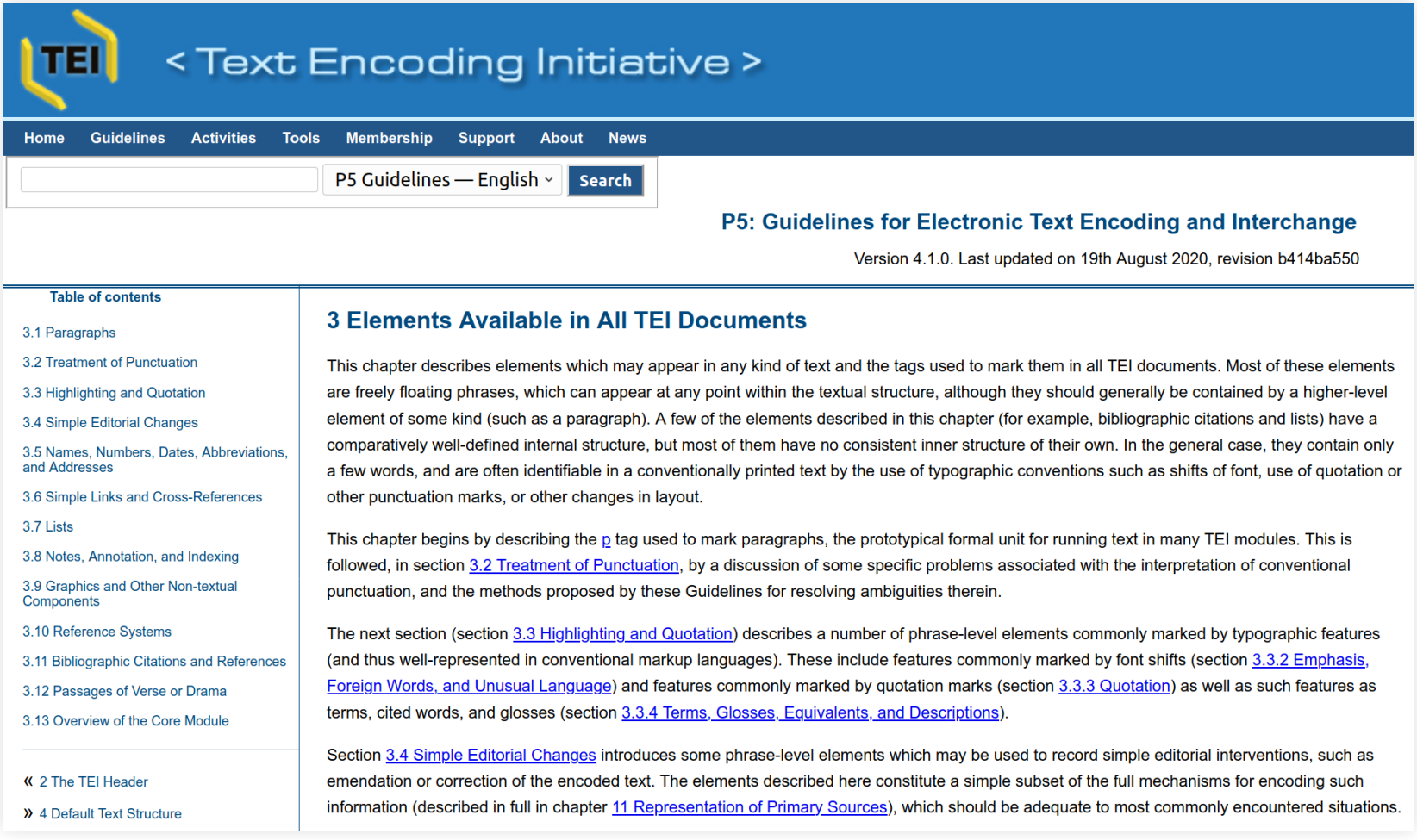
- `<text>`
- `<front>`, `<body>`, `<back>`
- `<group>`
- `<titlePage>`
- `<graphic>`
- `<div>`

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)



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

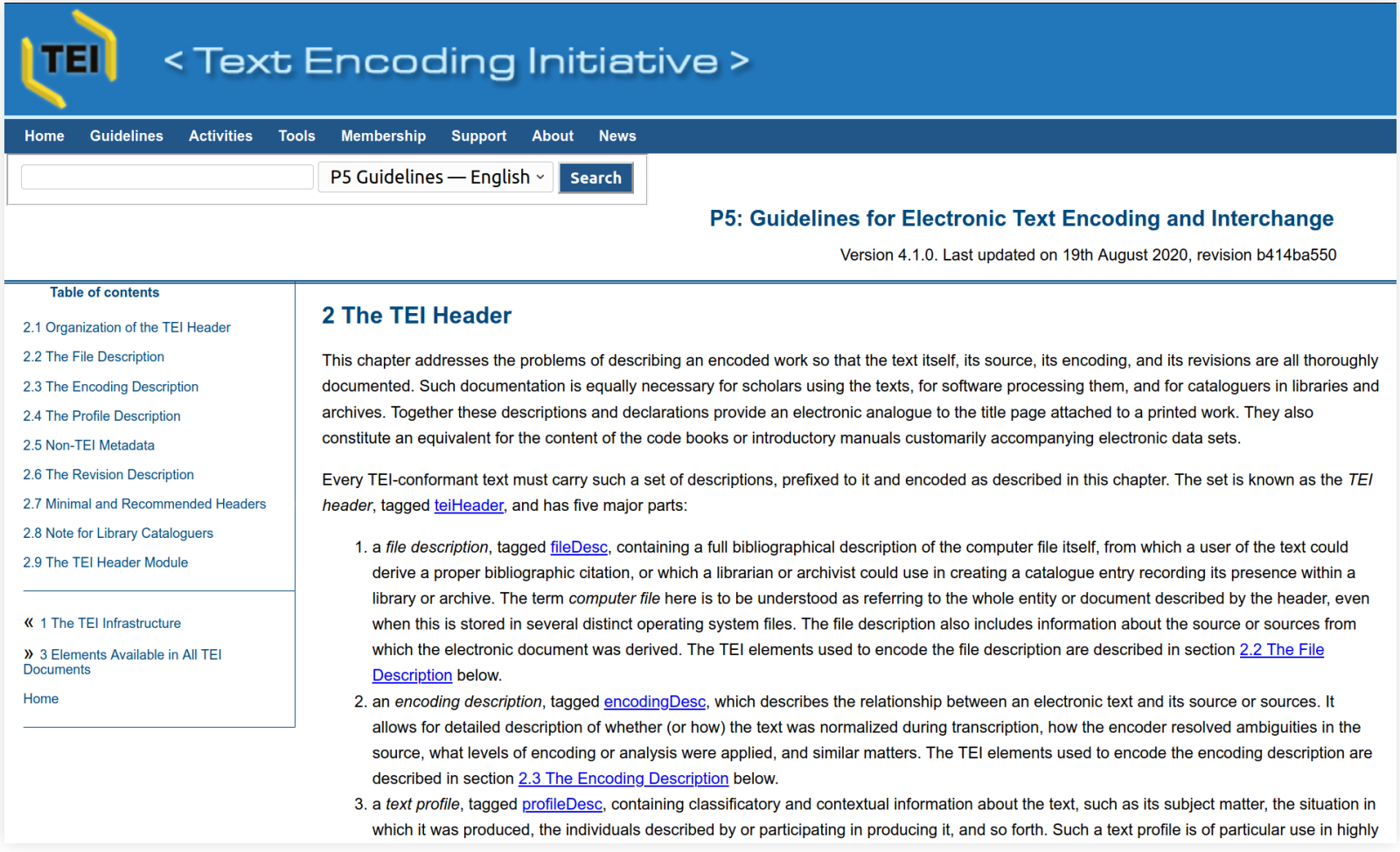
```
<sp>
  <speaker>Mephistopheles</speaker>
  <lg type="stanza">
     <l>Ich möcht’ mich gleich dem Teufel übergeben,</l>
     <l>Wenn ich nur selbst kein Teufel wär’!</l>
  </lg>
</sp>
<sp>
  <speaker>Faust</speaker>
  <lg type="stanza">
     <l>Hat sich dir was im Kopf verschoben?</l>
     <l>Dich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



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>`

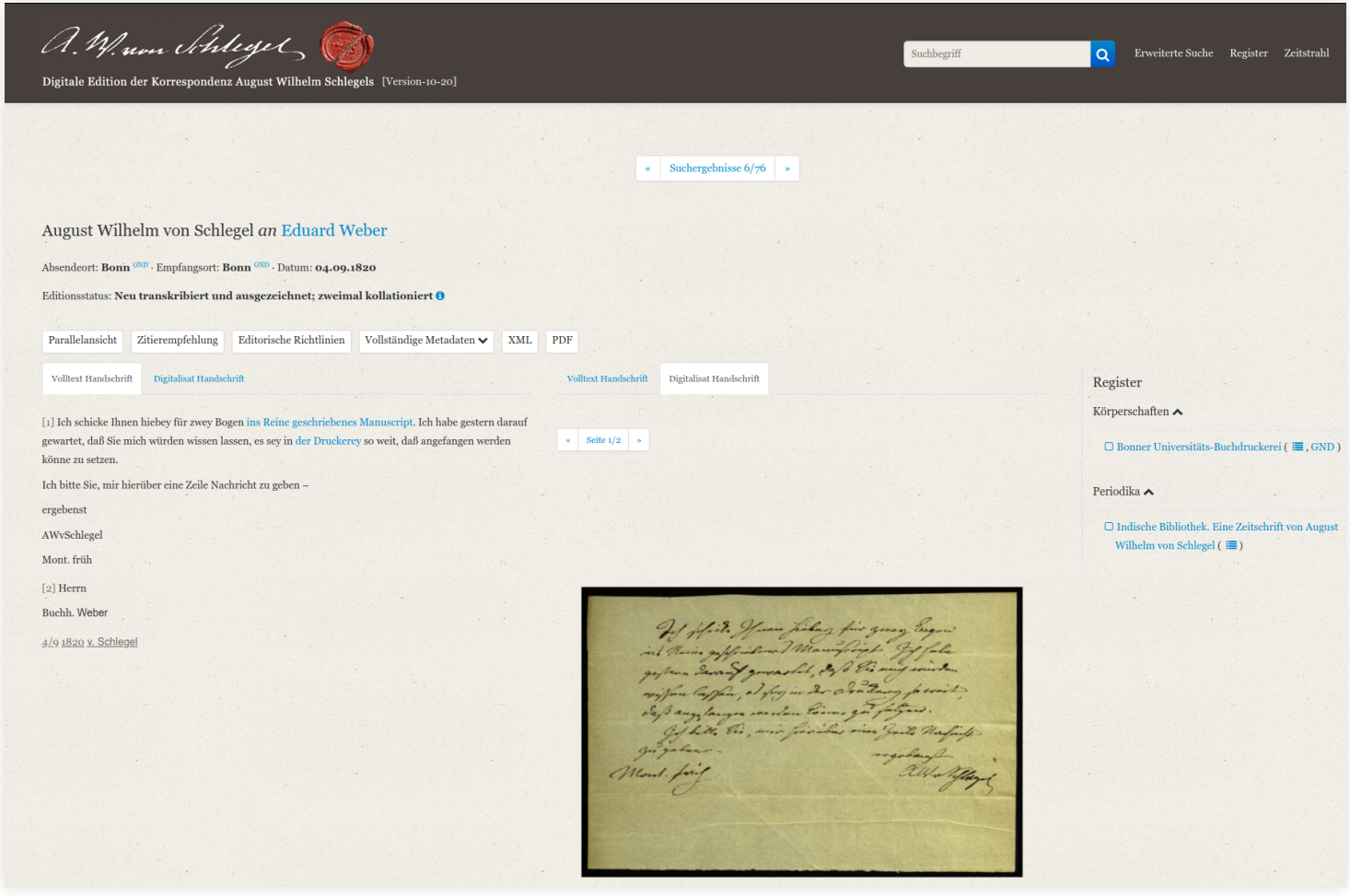
- Standard-Rollen
  - `<author>` - Autor des kodierten Texts oder des „born digital“ Dokuments
  - `<editor>` - Herausgeber des kodierten Texts
- Für Spezifischere Rollen
  - `<respStmt>` mit `<resp>` und `<name>`
  - `<resp>` - frei definierbare Aufgabe, bspw. "Transkription" oder "Kodierung der Named 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>
      </resptStmt>
    </titleStmt>
    [...]
  </fileDesc>
</teiHeader>
```

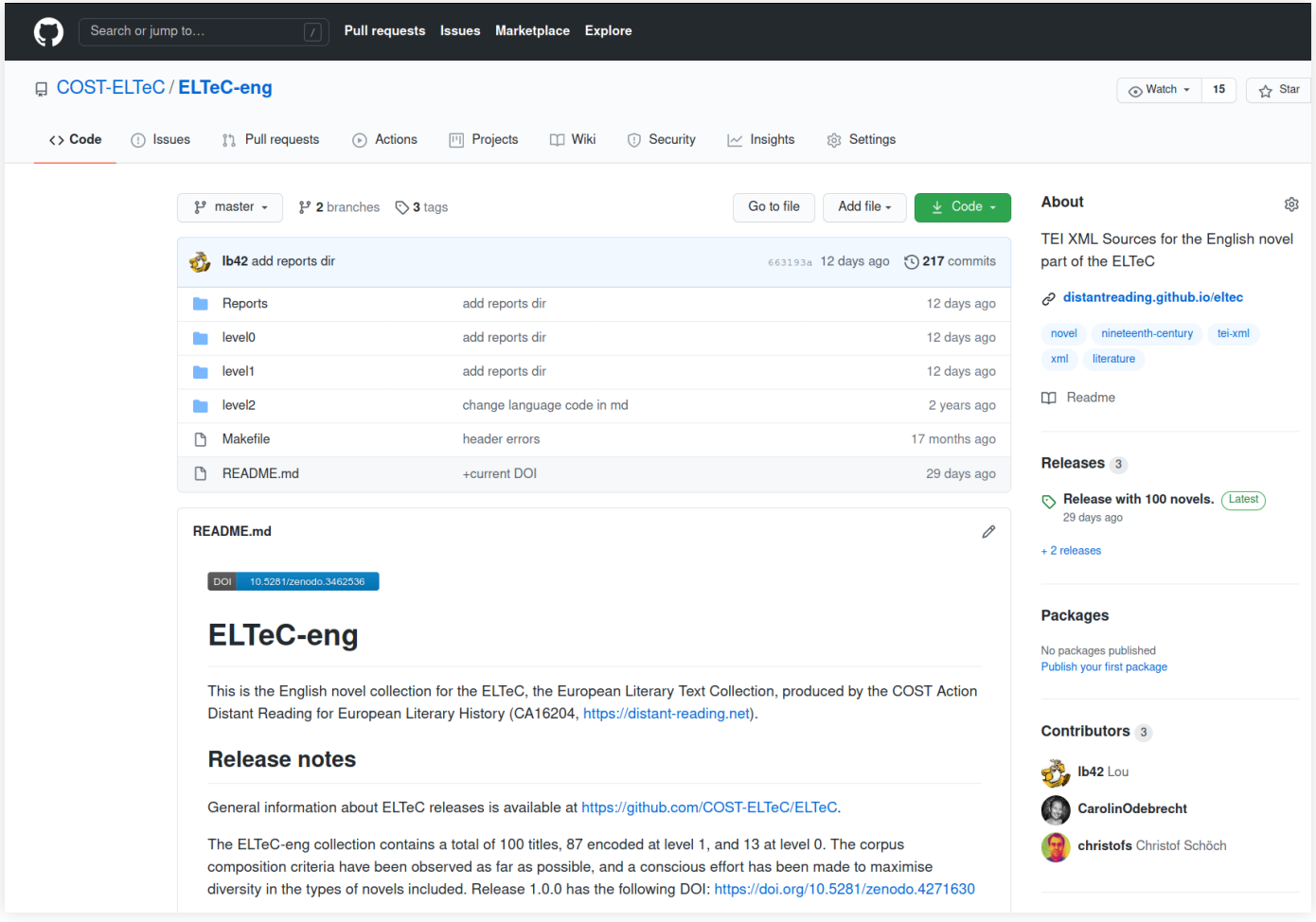
Beispieldaten

Brief aus der A.W.-Schlegel-Edition



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

Roman aus ELTeC-eng (Gaskell)



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://romanischestudien.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
