Winter School 2024

Bergische Universität Wuppertal

18.03 – 22.03.2024

Grundsätze der Textauszeichnung mit TEI

Nadine Sutor

Foto: Nadine Sutor

BERGISCHE UNIVERSITÄT WUPPERTAL

DH@BUW

Dokument Text Edition Graduiertenkolleg 2196

IZ ED Interdisziplinäres Zentrum für Editions- und Dokumentwissenschaft

i|d|e

Worum geht es heute?

- Wie kommen wir von XML nach TEI?
- Was ist TEI?
- Wofür wird TEI eingesetzt?
- Welche Bezugspunkte gibt es zu anderen Technologien?
- Was ist mit TEI (alleine) (nicht) möglich?
- Guidelines

Vorbemerkung: Warum technische Einführung in XML?

→ Grundlage für die Arbeit mit TEI

- TEI ist eine XML-Anwendung
- Verständnis für die Zusammenhänge und Begriffe in der XML-Welt (Elemente, Attribute, Prinzip der Wohlgeformtheit etc.)
- Grundlagenkenntnisse von XML essentiell für die Arbeit mit *oXygen*

Wir sind in der Lage XML-Dokumente zu erstellen…

Bisher:

- Verwendung der formalen XML-Regeln für den Entwurf *individueller Auszeichnungssprachen* (→ vgl. Postkarte)
- wir können (logische) *Strukturen* und *Inhalte* von Dokumenten (semantisch) auszeichnen
- auf Basis formaler Regeln für die XML-Syntax (→ „Wohlgeformtheit“)
- dabei haben wir *freie Wahl*, wie wir die Markup-Bestandteile benennen und hierarchisch anordnen
- Entwurf der Regeln für Dokumentstrukturen und Benennung nach *eigenen Vorstellungen* bzw. *eigener Sicht* auf das Dokument und deren Inhalte

ABER: Ist das wirklich so ???

Recap: XML

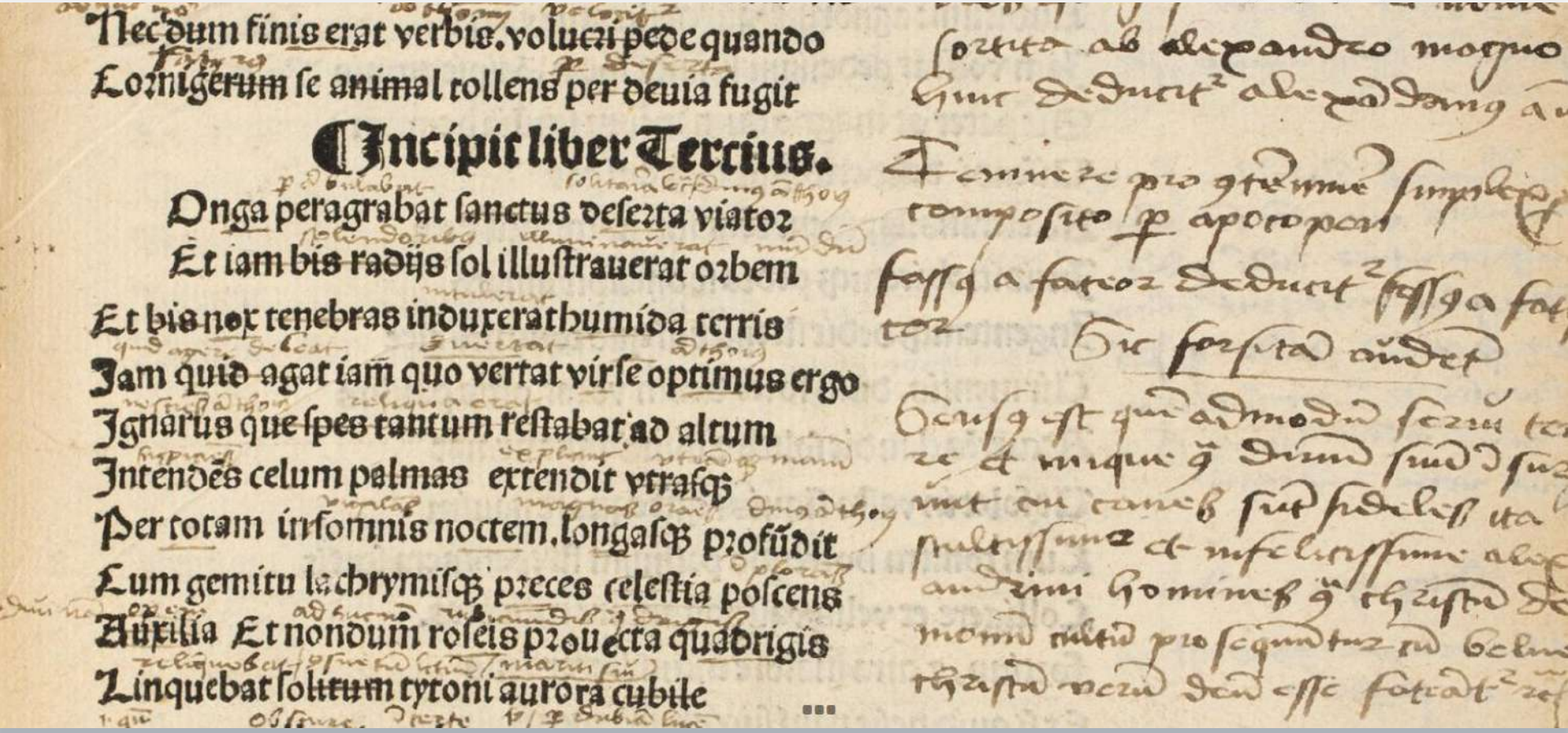
Für verschiedene Anwendungsfelder haben sich Standards für die Auszeichnung von Dokumenten etabliert:

- **z.B. GPS Exchange Format**
- Datenformat zur Speicherung von Geodaten (basiert auf XML)
- Ein XML-Schema beschreibt die Elemente und den Aufbau des GPS Exchange Formats
- Als Dateiendung wird die Abkürzung *.gpx* verwendet (Wikipedia)

Betrifft:

- Einheitliche Beschreibung von *Strukturen* und *Inhalten* (→ Benennung und hierarchischer Aufbau)
- Vereinfachung des Datenaustauschs und der Datenverarbeitung (→ Beschreibung von Dokumenten für das WWW = Verarbeitung durch verschiedene Browser möglich)

Wie sieht es mit einem Standard für die digitale Beschreibung von Dokumenten in den Textwissenschaften aus?


Nec dum finis erat verbis. volucri pede quando
Lornigerum se animal tollens per deuia fugit

Incipit liber Tercius.

Onga peragrabat sanctus deserta viator
Et iam bis radijs sol illustrauerat orbem
Et bis nox tenebras induxerat humida terris
Jam quid agat iam quo vertat vir se optimus ergo
Ignarus que spes tantum restabat ad altum
Intendes celum palmas extendit vtrasq;
Per totam insomnis noctem, longasq; profudit
Cum gemitu lachrymisq; preces celestia poscens
Auxilia Et nondum roseis prouecta quadrigis
Linquebat solitum tytoni aurora cubile ...

Text Encoding Initiative: **Anwendungsgebiet**

“Die Relevanz der TEI als des führenden Standards für Textcodierung ist innerhalb der Fachgemeinschaft der „Digital Humanities“ unbestritten.“ [Sahle 2013; S. 341]

die **TEI** …

… verfolgt weniger das Ziel des Entwerfens neuer Texte

… rückt vielmehr die ***Re****kodierung* von Texten und Dokumenten in den Fokus

… ermöglicht eine wissenschaftlich zulässige und informationsreiche digitale Textrepräsentation

… hat einen interdisziplinär und textübergreifenden Anspruch

… ist ein gut dokumentierter, etablierter Standard

… gewährleistet die Austauschbarkeit, langfristige Nutzbarkeit und Verarbeitbarkeit von TEI-Dokumenten (TEI-Guidelines)

Text Encoding Initiative: **Anwendungsgebiet**

- International anerkannter Standard zur Kodierung und zum Austausch von Texten in den Geisteswissenschaften
- Hat sich für die philologische Textauszeichnung und texttechnologische Erschließung geisteswissenschaftlicher Inhalte etabliert
- erleichtert und unterstützt Vorhaben, die über die statische Präsentation eines Textes hinausgehen. TEI-Daten:
  - dokumentieren die Entstehungsgeschichte
  - erlauben eine detaillierte Auszeichnung des dokumentarischen Befundes
  - verbinden Text und Bild
  - generieren variable Anzeigeformen (Single-Source-Publishing)

Vgl. Georg Vogeler (2012): Digitale Edition mit der TEI. Vortragsfolien zum Workshop *Texttechnologische Standards in den Geisteswissenschaften* am 17.10.2012 in Wien.

Text Encoding Initiative: **Entstehungsgeschichte**

1980er : Es existieren diverse Projekte, die sich mit der Verarbeitung und der geisteswissenschaftlichen Auswertung von digitalen Texten befassen:

- Unterschiedliche technische Lösungen für die Kodierung von Texten
- Abhängigkeiten von spezieller Hard- und Software

1987: Verabschiedung der *Poughkeepsie Principles*

- Treffen von Fachwissenschaftler*innen in Poughkeepsie, NY
- Entwurf von Prinzipien für ein allgemeines Textbeschreibungs- und -auszeichnungsformat

→ Grundlage für die Entwicklung des TEI-Regelwerks

http://projects.oucs.ox.ac.uk/ENRICH/Deliverables/Training/Graphics/poughkeepsie.png

Text Encoding Initiative: **Entstehungsgeschichte**

1988: Gründung der TEI als Forschungsprojekt
Einrichtung unterschiedlicher Arbeitsgruppen für bestimmte Aufgabenbereiche

1990: Verabschiedung eines ersten Vorschlags für "*Guidelines for the Encoding and Interchange of Machine-Readable Texts*"
(→ TEI-Guidelines Version „P1")

In den folgenden Jahren:

- Ergänzung und Weiterentwicklung der Richtlinien durch spezielle Arbeitsgruppen (*special interest groups*)
- Technische Anpassung, Weiterentwicklung der Markup-Technologien

Text Encoding Initiative: **Entstehungsgeschichte**

1990: Release der ersten Version P1 (damals noch SGML-basiert)
(Standard Generalized Markup Language)
ISO 8879:1986 (seit 1986)

2001: Umstellung auf XML-Technologie (Version P4)

Aktuell: P5 Version 4.7.0. Last updated on 16th November 2023

Inzwischen gibt es eine große Bandbreite von Anwender*innen
→ Bibliotheken, Museen, Verlage, Forschungsprojekte, etc.

Text Encoding Initiative
Organisation und Zielsetzung

TEI Consortium

"*a nonprofit membership organization composed of academic institutions, research projects, and individual scholars from around the world*"

Organisation, die Empfehlungen für die Kodierung von Texten erarbeitet und publiziert.

Zielsetzung

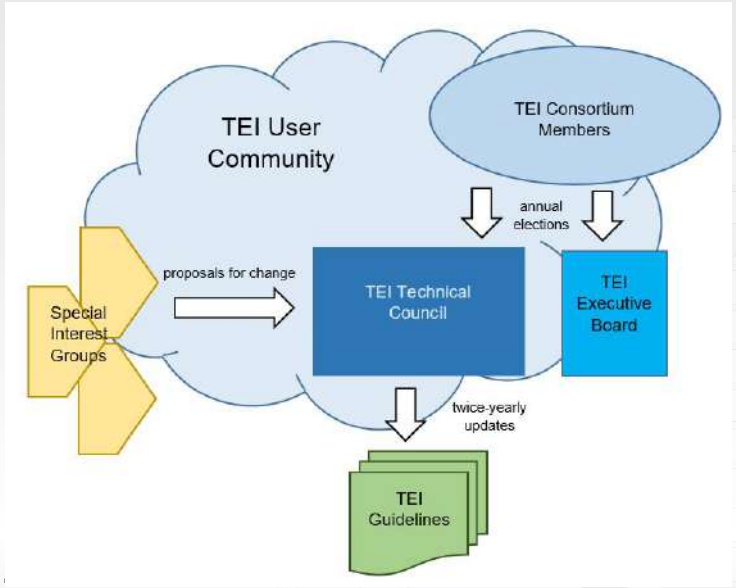
Entwicklung eines Standards für die digitale Repräsentation von (textbasierten) Dokumenten in den Geisteswissenschaften

"*addressed to anyone who works with any kind of textual resource in digital form*"

[https://tei-c.org/guidelines/]

Text Encoding Initiative

Begriffe und Arbeitsabläufe



TEI-Guidelines

- *Spezifikation* und *Dokumentation* des TEI-Markups
- Anwendungsempfehlung, *Beispiele*
- Einführung, Dokumentation Versions-Historie. http://www.tei-c.org/Guidelines/P5/
- TOC: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

TEI als Bezeichnung der Markup-Sprache

- Spezielle XML-Anwendung zur philologischen Textauszeichnung
- Entwickelt und gewartet durch das TEI Consortium
- Dokumentiert in den TEI Guidelines (Regeln für die Anwendung)

Text Encoding Initiative: **Community**

- die TEI ist eine Institution, die getragen wird von einer Gemeinschaft von Forschenden
- bietet neben den Guidelines auch dazugehörige Ressourcen (Tools, Handreichungen, Mailingliste)
- gibt eine Zeitschrift heraus: *Journal of the Text Encoding Initiative*

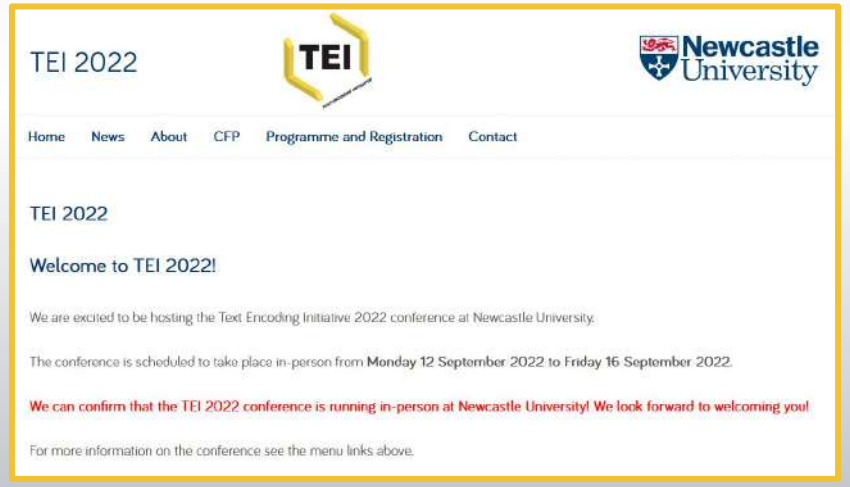
Text Encoding Initiative: **Community**

"***Tools*** *for creating, editing, transforming, and publishing TEI documents and schemas are essential to using the TEI Guidelines*"

- Roma → *Generation of schemas and documentation*
- OxGarage → *Conversion to and from TEI*
  - Dokumente: `.docx`, `.txt` (plain text) oder Präsentationen: `.pptx` oder Spreadsheets: `.csv`, `xls`
- Stylesheets → *Stylesheets for converting TEI documents to various formats*
  - z.B. HTML oder LaTeX
  - Dokumentation online, updated mit jedem neuen Release der Guidelines
  - GitHub, TEI-Wiki
- Zotero-Library, Zotero-Group ("markup theory")

Text Encoding Initiative: **Community**

Organisiert eine Jahreskonferenz → *Annual Conference and Members' Meeting*



TEI-Guidelines: Das Arbeitsergebnis der Text Encoding Initiative

- Regelwerk der TEI
- Modular aufgebaut (22 Module)
- definiert die Kodierung unterschiedlicher Genres
- liegen in ihrer fünften Version (P5) auf Basis von XML vor (4.7.0., letztes Update am 16.11.2023)
- Markup umfasst über 500 Elemente und über 200 Attribute
- *Special Interest Groups* → https://tei-c.org/activities/sig/
    - entwickeln Markup für Themenbereiche, die noch nicht in den TEI-Guidelines aufgenommen sind
    - *SIG Music*: Musiknotation, *SIG Correspondance*: Briefedition
    - *SIG Manuscripts*: Handschriften, *SIG Text and Graphics*

TEI-Guidelines: Das Arbeitsergebnis der Text Encoding Initiative

- formulieren **Regeln** für das TEI-Markup und **Empfehlungen** für die Vorgehensweise bei der digitalen Textauszeichnung:
  - ***Spezifikation*** von Markup für die elektronische Repräsentation von Text
    (formale Deklaration: Bezeichnung, Inhaltsmodelle)
  - ***Dokumentation*** der Markup-Komponenten
    (Beschreibung des angedachten Verwendungszwecks)
  - ***Beispiele*** für die Anwendung (in Kombination mit weiterem TEI-Markup)

TEI-Guidelines: **Flexibilität**

Die Text Encoding Initiative versteht sich als Plattform, um den Bedürfnissen verschiedener Fachgemeinschaften gerecht zu werden.

Das Markup soll die Freiheit bieten, die Sicht der Anwender*in auf den „Text“ / das „Dokument“ auszudrücken.

Allgemeine Elemente (Differenzierung über Attribute):

`<div>` *(text division) contains a subdivision of the front, body, or back of a text.*

Spezielle Elemente für konkrete Anwendungsfälle: z.B. in der Editionswissenschaft

`<listApp>` *(list of apparatus entries) contains a list of apparatus entries.*

```
<listApp xml:id="CA_Y-36"
   xml:lang="pal-Avst">
   <head>Variants from witnesses in Avestan script</head>
   <app from="#Y-36.01_L1_W-01">
    <rdg wit="#Pt4 #F2 #J2 #M1">ahiiā</rdg>
   </app>
   <app from="#Y-36.01_L1_W-02">
    <rdg wit="#Pt4 #F2 #J2 #M1">θβā</rdg>
   </app>
   <app from="#Y-36.01_L1_W-03">
    <rdg wit="#Pt4 #J2 #M1">āθrō</rdg>
    <rdg wit="#F2">āθrōi</rdg>
   </app>
<!-- ... -->
  </listApp>
```

TEI-Guidelines: **Flexibilität**

Das TEI-Markup ist also in seiner Gesamtheit sehr flexibel: Es ist vorgesehen, dass einzelne Befunde in unterschiedlicher Weise kodiert werden können.
Beispiel: *Auszeichnung von Personennamen*

`<name>` (name, proper noun) contains a proper noun or noun phrase
→ **(Modul core):**

```
<p>“If I asked you where the hell we were,” said
    <name type="person">Arthur</name> weakly,
            “would I regret it?”</p>
```

`<persName>` (personal name) contains a proper noun or proper-noun phrase referring to a person, possibly including one or more of the person's forenames, surnames, added names, etc → **(Modul namesdates)**

```
<p>“If I asked you where the hell we were,” said
<persName>Arthur</persName> weakly,
“would I regret it?”</p>
```

```
<p>“If I asked you where the hell we were,” said
<persName><forename>Arthur</forename></persName>
weakly, “would I regret it?”</p>
```

TEI-Guidelines: **Basic Modules**

Diese vier Module enthalten grundlegendes Markup (*Minimalkonfiguration*)

1. **The TEI Infrastructure (TEI Infrastructure, *tei*)**

   Beschreibt die Zusammenhänge zwischen den Markup-Komponenten und Datentypen für alle Module:
   [...] describes the infrastructure for the encoding scheme defined by these Guidelines. It introduces the conceptual framework within which the following chapters are to be understood [...]

2. **The TEI Header (Common Metadata, *header*)**

   Deklarationen für Elemente und Attribute zum Aufbau des TEI-Headers

3. **Elements Available in All TEI Documents (Common Core, *core*)**

   Deklarationen für allgemeine Elemente und Attribute, die in jedem Dokumenttyp benötigt werden

4. **Default Text Structure (Default Text Structure, *textstructure*)**

   Deklaration von grundlegenden Elementen zur hierarchischen Strukturierung von Dokumenten

TEI-Guidelines

Module in der Übersicht

Module name	Formal public identifier
analysis	Analysis and Interpretation
certainty	Certainty and Uncertainty
core	**Common Core**
corpus	Metadata for Language Corpora
dictionaries	Print Dictionaries
drama	Performance Texts
figures	Tables, Formulae, Figures
gaiji	Character and Glyph Documentation
header	**Common Metadata**
iso-fs	Feature Structures

TEI-Guidelines
Module in der Übersicht

Module name	Formal public identifier
linking	Linking, Segmentation, and Alignment
msdescription	Manuscript Description
namesdates	Names, Dates, People, and Places
nets	Graphs, Networks, and Trees
spoken	Transcribed Speech
tagdocs	Documentation Elements
tei	TEI Infrastructure
textcrit	Text Criticism
textstructure	**Default Text Structure**
transcr	Transcription of Primary Sources
verse	Verse

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 sourcecode

- Getting and Using the TEI Sources.
- TEI GitHub Repository
- Bug Reports, Feature Requests, etc.

Introducing the guidelines:
https://tei-c.org/guidelines/

table of contents:
https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html

TEI in der Anwendung

Warum TEI und nicht HTML ?

TEI ermöglicht die Erschließung von Dokumenten nach *textwissenschaftlichen Kriterien*

- Unabhängig von der Darstellung (Konzeption von HTML war ursprünglich präsentationsorientiert)
- TEI ist *nicht* zur Beschreibung der Präsentation in einem Ausgabemedium gedacht (hierzu ist der Einsatz von sog. CSS-Stylesheets notwendig)

TEI in der Anwendung

TEI ist komplexer als HTML und bietet 500+ *Tags* zur Auszeichnung von …

- logischen Textstrukturen **(Überschriften, Absätze etc.)**
- (typo)grafischen Informationen **(Hervorhebungen im Text etc.)**
- inhaltlichen / semantischen Informationen **(z.B. zur deskriptiven Auszeichnung und inhaltlichen Analyse)**
- sprachwissenschaftlichen Informationen **(grammatische Strukturen etc.)**

TEI in der Anwendung

TEI ist komplexer als HTML und bietet 500+ *Tags* zur Auszeichnung von …

- Metainformationen **(z.B. Textträgerbeschreibung, Entstehungskontext, Informationen zur Bearbeitung des digitalen Texts)**
- Relationen zwischen Textbestandteilen und/oder Metainformationen **(z.B. Verweissysteme im Dokument / der digitalen Edition)**
- Verknüpfung von Metainformationen mit dem Text, etc.

TEI in der Anwendung: **Potenzial**

Welches Einsatzpotenzial bietet TEI konkret für textwissenschaftliche Zwecke?

TEI als XML-basierte Anwendung:

- Zukunftsfähigkeit: XML ist eine etablierte, weit verbreitete Technologie
- Nachhaltigkeit: einfaches, offenes Datenformat
- Publishing-Workflows: in der heutigen Zeit primär digital und i.d.R. auf Basis von XML

TEI in der Anwendung: **Potenzial**

Möglichkeiten und Chancen im digitalen Raum:

- Aufhebung der für den Buchdruck charakteristischen Beschränkungen
- *Single Source Publishing*
- Material kann nach Interessen der NutzerInnen aufbereitet und selektiv bereitgestellt werden
- Synergie-Effekte durch Vernetzung (z.B. gemeinsame Nutzung von Datenbanken)
- Tendenz: Print-Ausgaben als Derivate (1:1 “Übersetzung/Übertragung”) digitaler Ausgaben (?)

  “***Eine digitalisierte Edition ist keine digitale Edition” (Sahle)***

  Eine digitale Edition folgt einem anderen Paradigma und geht inhaltlich und funktional über eine gedruckte Edition hinaus!

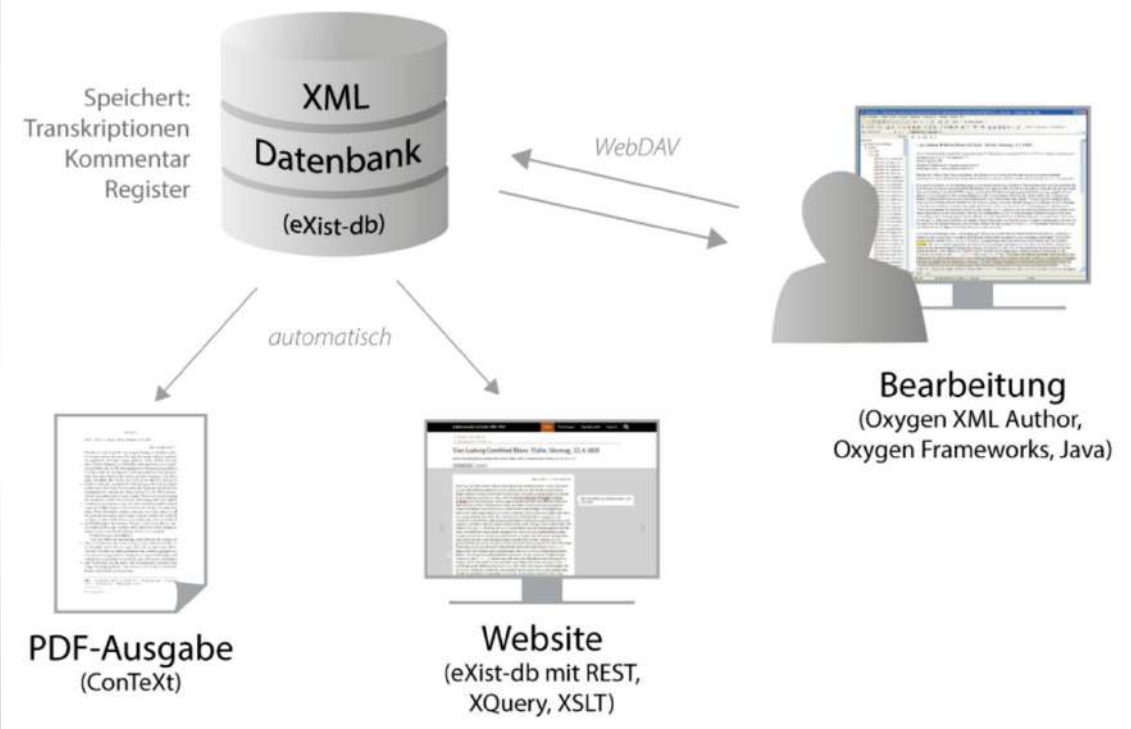
TEI in der Anwendung: **Herausforderungen**

- TEI: Standardisierung vs. Mehrdeutigkeiten bzw. Flexibilität und Umfang des Markups
- XML-Daten/digitale Projekte allgemein: Einstiegshürden in die Langzeitarchivierung
- Digitale Editionen: „Flüssigkeit“ des elektronischen Mediums vs. Forderung nach stabiler, zitierbarer Textgrundlage
- XML: Repräsentation von multiplen Hierarchien

TEI in der Anwendung: **Beispielworkflow**

Datenarchitektur & Workflow

TEI-Dokumente im elektronischen Publishing



Beispiel für einen TEI-basierten Publishing-Workflow mit *ediarum*

https://www.bbaw.de/bbaw-digital/telota/forschungsprojekte-und-software/ediarum

ediarum -
Digitale Arbeits- und Publikationsumgebung für Editionsvorhaben

Allgemeines zur Arbeit mit TEI

Wichtig: Vorgaben für Dokumente machen

- Schritte überlegen: Welche Inhalte sollen nach welchen Vorgaben kodiert und anschließend abgebildet werden?
- Was kommt alles in eine Datei? Wann splittet man Inhalte auf? Nach welchen Kriterien? (Nur trennen wenn es einen Grund dafür gibt (z.B. untersch. Bearbeiter*innen))

Allgemeines zur Arbeit mit TEI

Wichtig: Vorgaben für Dokumente machen

- Digitale Edition: Nur ausgewählte Textinhalte transkribieren die interessant/relevant sind
- anschließend in TEI modellieren und ins digitale Medium „überführen“. Dann folgen weitere Schritte:

  Schemaeinbindung → Transformation → Visualisierung
- Verwendung des TEI-Markups: Elemente, Attribute, Werte so nutzen, wie sie semantisch am besten “passen”
  → (pb = page break vs. milestone)

Literatur

- TEI by Example: Module 0: Introduction to Text Encoding and the TEI (Abschnitt 5 - 8)
- TEI Guidelines: The TEI Infrastructure (Intro und 1.1)
- Schöch, Christof (2016): Ein digitales Textformat für die Literaturwissenschaft: Die Richtlinien der Text Encoding Initiative und ihr Einsatz bei Textkonstitution und Textanalyse. Romanische Studien, Nr. 4, S. 325–364. Digital verfügbar unter: http://romanischestudien.de/index.php/rst/article/view/58/517. Date accessed: 05 März 2024.

Vielen Dank für Ihre Aufmerksamkeit!

BERGISCHE
UNIVERSITÄT
WUPPERTAL