Winter School 2024

Bergische Universität Wuppertal

18.03 – 22.03.2024

Die weite(re) Welt der TEI

Customization / Ausblicke

Andreas Mertgens

Foto: Nadine Sutor

BERGISCHE UNIVERSITÄT WUPPERTAL

DHa BUW

Dokument Text Edition Graduiertenkolleg 2196

IZ ED Interdisziplinäres Zentrum für Editions- und Dokumentwissenschaft

i|d|e

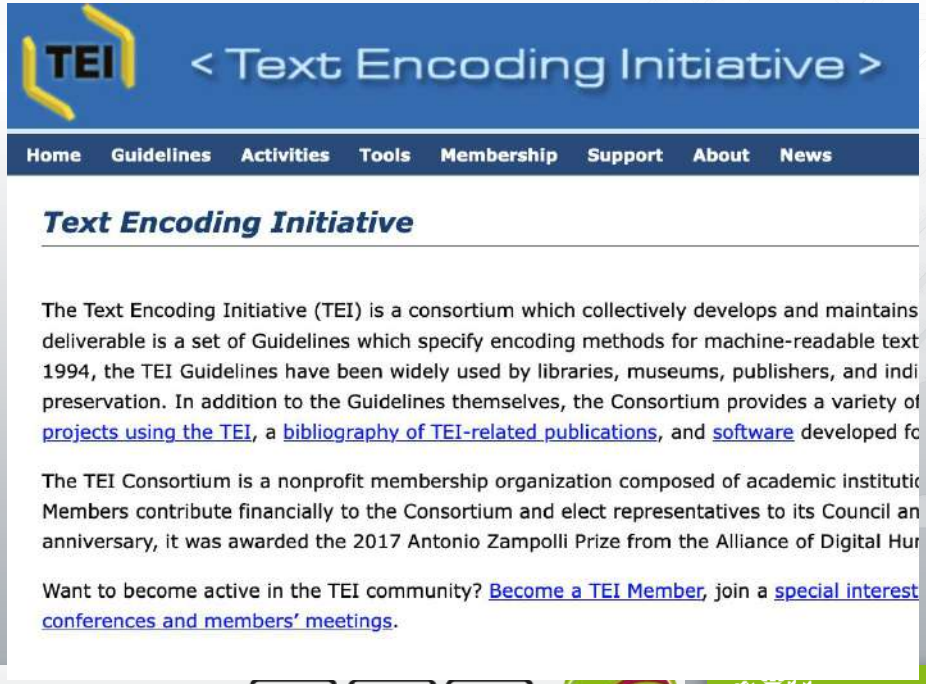
Wir schauen nun in zwei Richtungen über den Tellerrand dieser School:

1. **Customizing TEI: Ein kurzer Blick in den weiteren Werkzeugkasten**

   (nur damit sie es mal gehört haben, müssen sie nicht jetzt lernen)

2. **Ausblicke**
   - DraCor
   - EpiDoc
   - MEI

TEI Homepage als Anlaufpunkt für weitere Exploration:
Module, SIG (Special Interest Groups), Tools etc.



https://tei-c.org/

Kontrolle von Dokumentstrukturen: Schemata und Schemasprachen

Schemasprachen

I.d. R. ist für XML-Dokumente eine bestimmte Struktur für den Dokumentaufbau vorgesehen. Diese Struktur kann unter Hinzunahme formaler Grammatiken *beschrieben*, *dokumentiert* und *festgelegt* werden:

- Regeln für den Aufbau von Dokumentstrukturen definieren
- zulässige Markup-Komponenten deklarieren (“Inventar”)
  → betrifft das Inhaltsmodell bzw. den Attributtyp und -werte

Dokumentstrukturen *kontrollieren*:

- Entspricht die *Benennung* der Markup-Komponenten den Festlegungen im Schema?
- Entspricht die *Dokumentstruktur* den in dem Schema festgelegten Inhaltsmodellen?
- → **validieren** = überprüfen, ob ein XML-Dokument den aufgestellten Regeln im Schema entspricht -> Hinweise in Oxygen

Kontrolle von Dokumentstrukturen: Schemata

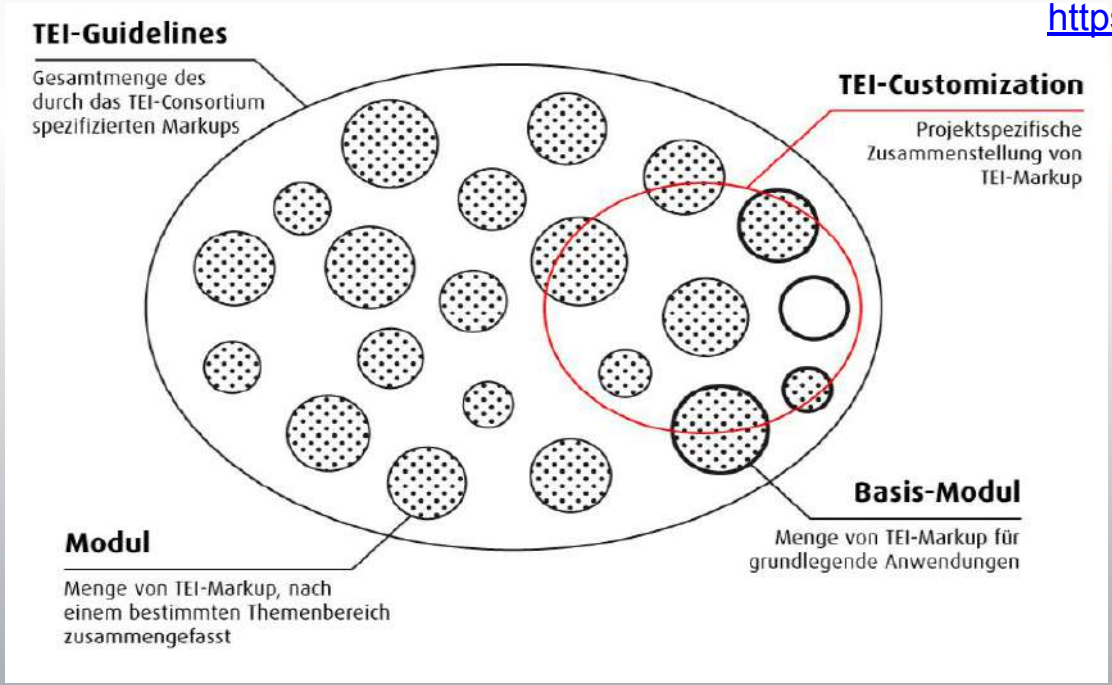
TEI-Lite

"includes basic elements for simple documents"

"TEI Lite is a specific customization of the TEI tagset, designed to meet "90% of the needs of 90% of the TEI user community""

→ enthält ca. 150 Elemente aus den TEI-Guidelines

https://tei-c.org/guidelines/customization/lite/



Kontrolle von Dokumentstrukturen: Schemata

TEI-All

Das andere Ende des Spektrums. Alles ist erlaubt!

Vorteil: sehr flexibel, gut für Experimente

Nachteil: Für ein Textphänomene gibt es meistens mehrere mögliche valide Auszeichnungen. (haben wir Dienstag bei der Postkarte erlebt)

Element: <rs> (referencing string) → <rs type="PersName">

Element: <name> (name) → <name type="person">

Element: <persName> (personal name)

Kontrolle von Dokumentstrukturen: Schemata

“[...] it is almost impossible to use the TEI scheme without customizing or personalizing it in some way.” (cf. P5 Guidelines of the TEI, ch. 23.2)

“Customization is a central aspect of TEI usage and the Guidelines are designed with customization in mind. (http://www.tei-c.org/Guidelines/Customization/)

“From the start, the TEI was intended to be used as a set of building blocks for creating a schema suitable for a particular project. This is in keeping with the TEI philosophy of providing a vocabulary for describing texts, not dictating precisely what those texts must contain or might have contained.”

(http://www.tei-c.org/Guidelines/Customization/odds.xml) → Die TEI Guidelines unterstützen die Anpassung des Schemas.

Kontrolle von Dokumentstrukturen: Schemata

Customizations provided by the TEI Consortium

Lite	TEI Lite, the most widely used TEI customization; includes basic elements for simple documents	ODD DTD RNG XSD HTML PDF
TEI Tite	A constrained customization designed for use by keyboarding vendors.	ODD DTD RNG XSD HTML PDF
Bare	TEI Absolutely Bare, a very barebones schema with the absolute minimum of elements	ODD DTD RNG XSD
All	TEI with all modules included	ODD DTD RNG XSD
Corpus	TEI for Linguistic Corpora, includes the modules for encoding linguistic corpora	ODD DTD RNG XSD
MS	TEI for Manuscript Description, includes the elements for describing manuscripts and complex physical aspects of documents	ODD DTD RNG XSD
Drama	TEI with Drama, includes the TEI drama module	ODD DTD RNG XSD
Speech	TEI for Speech Representation, includes the TEI module for spoken language	ODD DTD RNG XSD
Dictionaries	TEI for Dictionaries	ODD DTD RNG XSD

http://www.tei-c.org/Guidelines/Customization/

Beispiel einer externen Customization: Basisformat des Deutschen Textarchivs
http://deutschestextarchiv.de/doku/basisformat

DTD

```
<!--doc:(abbreviation) contains an abbreviation of any sort. [3.6.5. Abbreviations and Their Expansions] -->
<!ELEMENT abbr %macro.phraseSeq;>
<!ATTLIST abbr xmlns CDATA "http://www.tei-c.org/ns/1.0">
<!ATTLIST abbr
 %att.global.attributes;
 %att.typed.attribute.subtype;
type %teidata.enumerated;  #IMPLIED >
<!--doc:(addition) contains letters, words, or phrases inserted in the source text by an author, scribe, or a p
<!ELEMENT add %macro.paraContent;>
<!ATTLIST add xmlns CDATA "http://www.tei-c.org/ns/1.0">
<!ATTLIST add
 %att.global.attributes;
 %att.transcriptional.attributes;
 %att.placement.attributes;
 %att.typed.attributes;
 %att.dimensions.attributes; >
<!--doc:(address line) contains one line of a postal address. [3.6.2. Addresses 2.2.4. Publication, Distributio
<!ELEMENT addrLine %macro.phraseSeq;>
<!ATTLIST addrLine xmlns CDATA "http://www.tei-c.org/ns/1.0">
<!ATTLIST addrLine
 %att.global.attributes; >
```

XSD

```
<xs:element name="add">
  <xs:annotation>
    <xs:documentation>(addition) contains letters, words, or phrases ins
  </xs:annotation>
  <xs:complexType>
    <xs:complexContent>
      <xs:extension base="tei:macro.paraContent">
        <xs:attributeGroup ref="tei:att.global.attributes"/>
        <xs:attributeGroup ref="tei:att.transcriptional.attributes"/>
        <xs:attributeGroup ref="tei:att.placement.attributes"/>
        <xs:attributeGroup ref="tei:att.typed.attributes"/>
        <xs:attributeGroup ref="tei:att.dimensions.attributes"/>
      </xs:extension>
    </xs:complexContent>
  </xs:complexType>
</xs:element>
```

ODD: One Document Does it All

TEI-Markup zur Beschreibung von TEI-Customizations

Ein TEI-basiertes Metaformat zur Generierung multipler Outputs.

Ein TEI Dokument, das *Schemaspezifikationen* (spezifische Dokumentationselemente) enthält. Diese bestehen aus:

- einer *formalen* Dokumentation von verwendeten Modulen, Elementen und Attributen sowie deren Werteinschränkungen
- einer *deskriptiven* Dokumentation
  - dazu werden Elemente aus dem TEI-Modul [22 Documentation Elements] verwendet
  - beinhaltet TEI-Markup für die Definition einer TEI-Customization und deren Dokumentation

→ Begründung: Welche Modul und Attribute brauche ich und warum? Was brauche ich nicht?

Warum habe ich diese Konfiguration so gewählt?

ODD: One Document Does it All

TEI-Markup zur Beschreibung von TEI-Customizations

- Dokumentation, Kontrolle und Festlegung der Kodierung in XML/TEI-Dokumenten
- erlaubt die Einbindung von CSS-Anweisungen
- Um die Technologien sauber voneinander zu trennen, sollte die ODD ausschließlich TEI-Markup umfassen

  → Schema nicht zu stark “überstrapazieren”

  → Auslagerung nicht relevanter Informationen (z.B. externes CSS-Stylesheet)

  → Stil

Die in Roma vorgenommenen Einstellungen werden in ODD-Files gespeichert.

ODD: Aufbau

- Es gibt einen **<teiHeader>** mit Metadaten (kennen wir schon)
- **<publicationStmt>**
- <body> mit der Beschreibung der Spezifikation
- mit <schemaSpec> startet die ODD-Customization (“Wurzelelement”), umfasst mehrere sog. <moduleRef>-Elemente
- <moduleRef> definiert die in einer ODD-Customization eingebundenen **Module**, inklusive deren Elemente, Attribute und zulässigen Werte

ODD: One Document Does it All

<schemaSpec> (schema specification)

- Container für Markup zur formalen Definition einer TEI-Customization
- Ein ODD-File muss genau ein schemaSpec-Element enthalten
- kann weiteres Markup zur Dokumentation der Customization beinhalten

<moduleRef> (module reference)

- Beschreibung der in die TEI-Customization eingebundenen **Module**.
  Wichtige Attribute:
  - @key : “the name of a TEI module“
  - @except : “supplies a list of the elements which *are not to be copied* from the specified module into the schema being defined“
  - @include : “supplies a list of the elements which *are to be copied* from the specified module into the schema being defined“

step-by-step Beispiel: http://www.tei-c.org/Guidelines/Customization/odds.xml

ODD: One Document Does it All

Beschreibung einer TEI-Customization ist auch **innerhalb des TEI-Headers eines XML/TEI-Dokuments** möglich:

<schemaSpec> lässt sich **inline** in der <encodingDesc> kodieren

Auf diese Weise können alle Informationen über die verwendete TEI-Customization im Metadatenbereich eines TEI-Dokuments festgehalten werden

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
      <fileDesc>
      <titleStmt>
            <title>..</title>
      </titleStmt>
      <publicationStmt>
            <p>...</p>
      </publicationStmt>
      <sourceDesc>
            <p>...</p>
      </sourceDesc>
      </fileDesc>
      <encodingDesc>
      <schemaSpec ident="engels" start="TEI">
            <moduleRef key="header"/>
            <moduleRef key="core"/>
            <moduleRef key="tei"/>
            <moduleRef key="textstructure"/>
      </schemaSpec>
      </encodingDesc>
  </teiHeader>
  <text>
      <body>
      <p>...</p>
      </body>
  </text>
</TEI>
```

RelaxNG : Schemabeschreibungssprache für XML

- Wird aus der ODD erzeugt (ROMA-Tool oder Transformationsszenario in oXygen) **strukturelle Validierung**
- ist jedoch begrenzt mit Blick auf die **inhaltliche Validierung**.
- Würde das Schema überfrachten, zu unsystematisch, Vermischung der Ansätze
- Speziellere Dinge müssen über andere Schemata geregelt werden → **Schematron**

  → “Regelsatz”

  → projektspezifische Modifikationen

  → Vorteil: Einfache Syntax, gute Lesbarkeit, Datentypen von W3C XML Schema übernommen, XML-basiert

Schematron: “A language for making assertions about the presence or absence of patterns in XML documents” (https://schematron.com/)

- zur Validierung von **Struktur** und **Inhalt** von XML-Dokumenten
- ermöglicht die Abfrage und Prüfung bestimmter Regeln, die mit anderen Sprachen *nicht* möglich sind
- ist “mächtiger” in seinen Funktionen und stellt eine Ergänzung zu anderen Sprachen wie DTD, XML Schema oder RelaxNG dar. (Z.B. Markup reglementieren, reduzieren und erweitern)
  - Einführung auf data2type: Schematron kurz und bündig
- keine Beschränkung auf die TEI
  (“Business rules validation, data reporting, general validation, quality control, quality assurance, firewalling, filtering, constraint checking, naming and design rules checking, statistical consistency, data exploration, transformation testing, feature extraction, house-style-rules checking”)

Schematron: Validierung von Struktur und Inhalt von XML-Dokumenten

Inhaltsvalidierung

- zusätzliche Validierung (“layer”), die man über die Daten legt
- überprüft Strukturen, die mit Hilfe des RNG-Schemas nicht kontrollierbar sind

Beispiel

- Briefe sollen an bestimmten Stellen mit einer Anrede beginnen
- automatische Warnung, wenn diese Art der Kodierung im XML-File nicht vorhanden ist
  → “Connection” via *Einbindung* des Schemas in die XML/TEI-Daten
- Wenn kein <opener> vorkommt, kann mit einem Befehl automatisch an dieser Stelle ein Element eingefügt werden, wo es das Schema vorsieht

Schematron: Validierung von Struktur und Inhalt von XML-Dokumenten

- Zeigt Warnungen an den Stellen im Dokument an, wo etwas fehlt:
  - Quick-Fix: Mit einem Klick ergänzen
  - ist über das RNG-Schema nicht möglich!
- erlaubt die Anwendung von *RegEx*
- Via repair-Funktionen fehlende Elemente automatisch einfügen oder eine Sequenz / Abfolge bestimmter Elemente generieren
- Auslagerung projektspezifischer Informationen (Workflow)
  - im Schema neue Namespaces erstellen und einbinden
  - separat von TEI-Daten halten und losgelöst im Schema deklarieren und bearbeiten
  - via Funktion in das RNG-Schema includen und jederzeit wieder excluden

→ **Nachhaltigkeit // Nachnutzung**

Jedes unterhalb von muss type = letter und subtype=full/substitute haben

```
<rule context="tei:body/tei:div[@type = 'part']/tei:div">
    <assert test="@type" sqf:fix="addType">Each &lt;<name/>&gt; needs to have a @type.</assert>
    <assert test="@subtype" sqf:fix="addSubtype">Each &lt;<name/>&gt; needs to have a @subtype.</assert>
    <sqf:fix id="addType">
        <sqf:description>
            <sqf:title>Adding the @type attribute</sqf:title>
            <sqf:p>In order to validate, you need to complete it with 'letter'.</sqf:p>
        </sqf:description>
        <sqf:add match="." target="type" node-type="attribute">
            <value-of select="'letter'"/>
        </sqf:add>
    </sqf:fix>
    <sqf:fix id="addSubtype">
        <sqf:description>
            <sqf:title>Adding the @subtype attribute</sqf:title>
            <sqf:p>In order to validate, you need to complete it with 'full' or 'substitute'.</sqf:p>
        </sqf:description>
        <sqf:add match="." target="subtype" node-type="attribute"/>
    </sqf:fix>

    <assert test="matches(@xml:id, concat($docID, '_letter_\d{5}$')) or matches(@xml:id, concat($docID, '_
    <assert test="@sameAs">The @sameAs should be used to record the numbering of the letters in the editic
    <report role="warn" test="matches(@xml:id, concat($docID, '_letter_noEntryExists$'))">This entry has c
</rule>
```

Für divs mit type = letter und subtype=full/substitute gibt es unterschiedliche verpflichtende Elemente

```
<pattern id="full">
    <rule context="tei:div[@type = 'letter'][@subtype = 'full']">
        <assert test="tei:head">A document element must contain a &lt;head&gt; element.</assert>
        <assert test="tei:pb">A document element must contain a page break (&lt;pb&gt;) element.</assert>
        <assert test="tei:note">A document element must contain a &lt;note&gt; element.</assert>
        <assert test="tei:p or tei:div[@type = 'letterSection']">A document element must contain a &lt;p&gt; element.</assert>
   </rule>
</pattern>
<pattern id="substitute">
    <rule context="tei:div[@type = 'letter'][@subtype = 'substitute']">
        <assert test="tei:head">A document element must contain a &lt;head&gt; element.</assert>
        <assert test="tei:note">A document element must contain a &lt;note&gt; element.</assert>
        <assert test="tei:p">A document element must contain a &lt;p&gt; element.</assert>
    </rule>
</pattern>
```

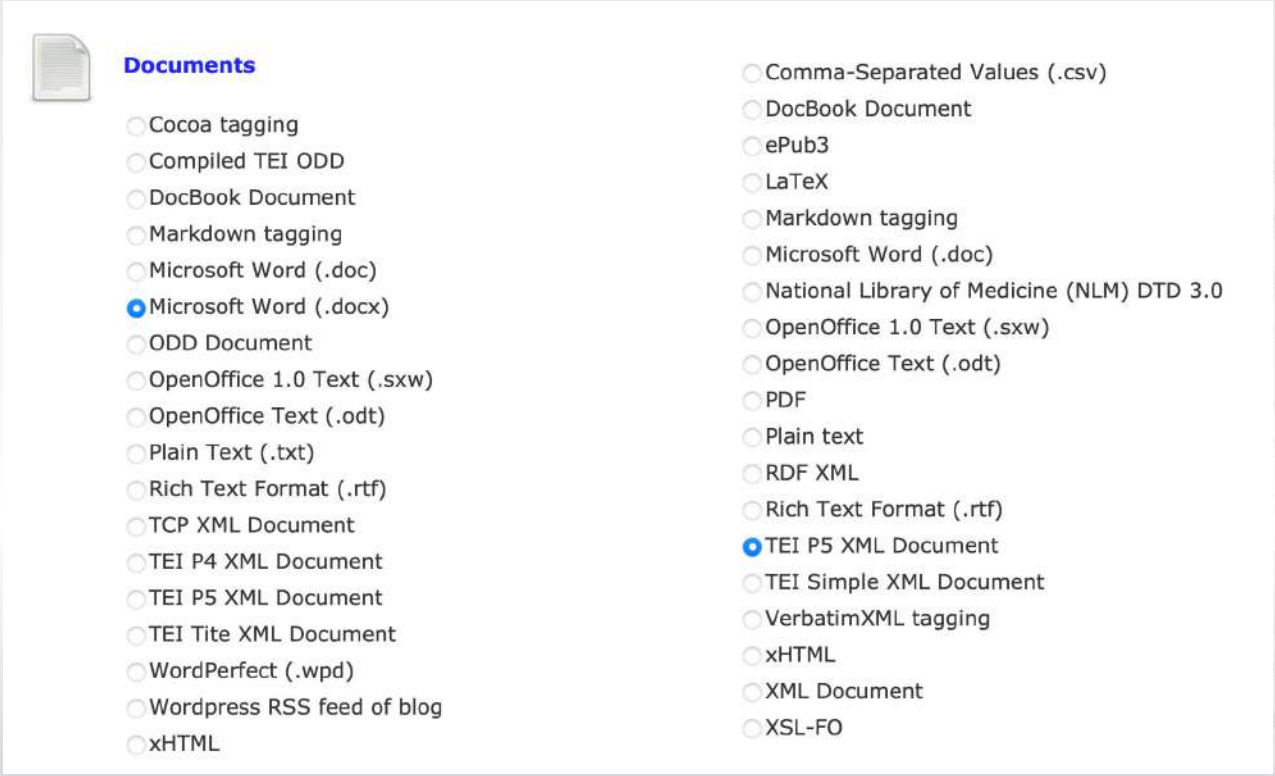
ROMA: „generating customizations / validators for the TEI“

Roma ist eine webbasierte Anwendung zur Erstellung und Bearbeitung sog. *TEI-Customizations*

- neu oder auf Basis bereits bestehender TEI-Customizations
- Kernfunktion: Zusammenstellen von TEI-Markup
- Änderung der Inhaltsstrukturen des TEI-Markup-Modells
- Automatische Ausgabe von:
  - **ODD-File** (TEI-Dokument zur Beschreibung der TEI-Customization)
  - **RelaxNG** (Schema-Sprache für XML zur Beschreibung der Struktur eines XML-Dokuments)
  - **Schematron**-Regelsatz: Inhaltsvalidierung, “mächtiger” in seiner Funktion, Ergänzung zu anderen Schemasprachen
  - Dokumentation in untersch. Formaten

https://roma.tei-c.org/

TEI Tools: TEIGarage (früher OXGarage)



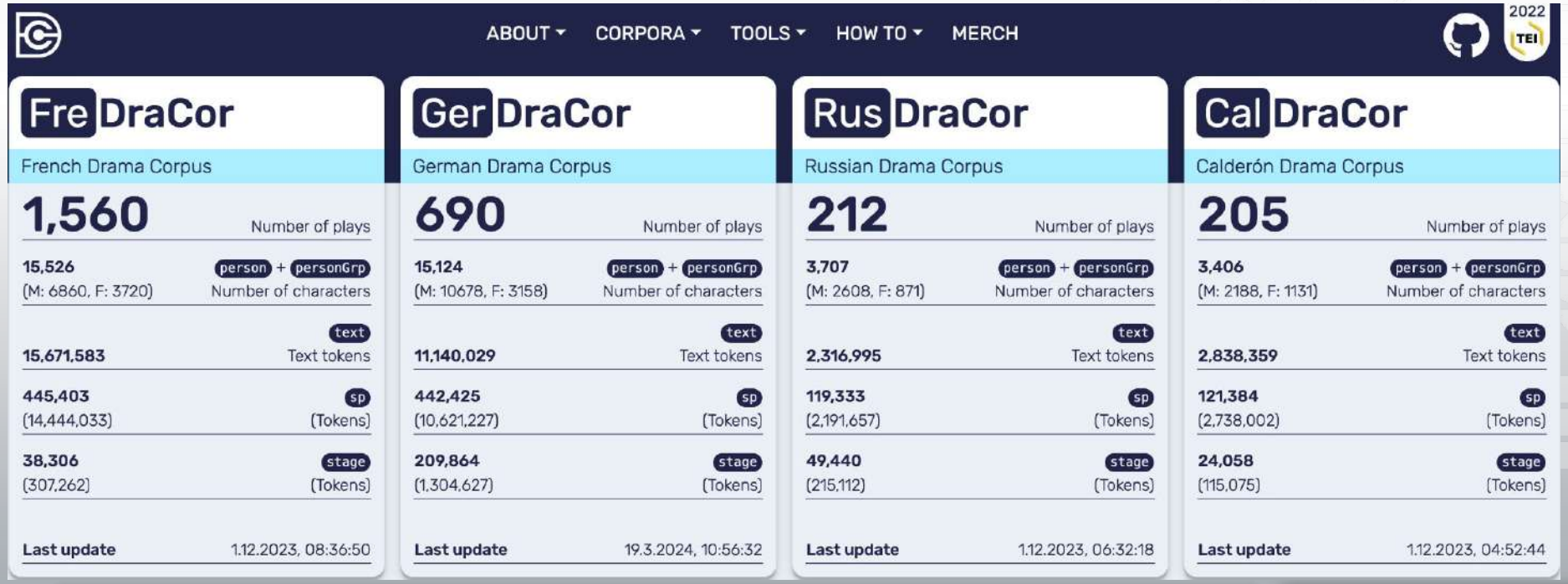
Konvertierung aller möglicher Dateiformate: https://teigarage.tei-c.org/

- Ausgangspunkt für weitere Transformationen

z.B. TEI Modul 7 “Performance Texts”

https://tei-c.org/release/doc/tei-p5-doc/en/html/DR.html

Anwendungsfall: DraCor ("**drama corpora**") https://dracor.org/



DraCor uses a TEI customization that contains only selected TEI elements supported by the wider DraCor system and restricts the use of XML attributes and its values.

However, homogeneity is something that has to be created first, because even if corpora are available in the target format TEI, they often differ in the way TEI is applied.

The cast lists or dramatis personae that are contained in most dramatic texts are an insufficient source in this regard, because they tend to be incomplete. Speaker labels contained in the proper text are also often misleading, because they are often not stable enough to serve as an identifier.

Therefore the plays encoded for the DraCor plattform have an **additional section in their metadata in the <teiHeader>** that lists all characters as <person> elements in a <listPerson> and assigns them an unique identifier xml:id, that is then used in the attribute who to link the individual speech acts <sp> with their respective speakers.

https://dracor.org/doc/odd#div_intro

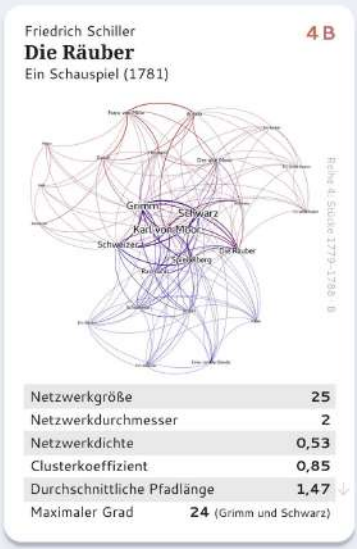
Anwendungsfall: DraCor ("drama corpora",)

https://dracor.org/

Emilia Galotti – original DraCor project t-shirt

Card games

Dramenquartett



Eine sehr große spezialisierte Nische:

EpiDoc: Epigraphic Documents in TEI XML

- provides guidelines and tools for encoding scholarly and educational editions of ancient documents.
- uses **a subset of the Text Encoding Initiative**'s standard
- developed initially for the publication of digital editions of ancient inscriptions (e.g. Inscriptions of Aphrodisias, Vindolanda Tablets).
- domain has expanded to include the publication of papyri and manuscripts (e.g. Papyri.info).
- addresses not only the transcription and editorial treatment of texts themselves, but also the history and materiality of the objects on which the texts appear

https://epidoc.stoa.org/

EpiDoc: Epigraphic Documents in TEI XML

https://papyri.uni-koeln.de/stueck/tm44629

```
<lb n="1"/><supplied reason="lost">το</supplied>ῖς τὸ <num value="35">λε</num> <expan><ex>ἔτος<

<lb n="2"/><unclear>π</unclear>αρὰ Ἱππ<unclear>άλ</unclear>ου καὶ Θεοδότου καὶ Πο<unclear>λ</un

<lb n="3"/><supplied reason="lost">τ</supplied>ῶν <num value="3">γ</num> Ἰουδαίων τῶν ἐκ Πειμπα

<lb n="4"/><supplied reason="lost">με</supplied>μ<unclear>ι</unclear>σ<unclear>θ</unclear>ωμένω
<lb n="4"/><gap reason="lost" extent="unknown" unit="line"/>
   </ab></div><div n="v" type="textpart"><ab>

<lb n="5"/><gap reason="lost" extent="unknown" unit="character"/><gap reason="illegible" quanti

<lb n="6"/><gap reason="lost" extent="unknown" unit="character"/><gap reason="illegible" quanti
      <div type="translation">
```

TEI substandard EPIDOC (für Inschriftenkunde, Papyrologie etc.)

```
<body>
    <p>Some <space extent="2cm" quantity="2"></space>text
            <supplied>here</supplied>.</p>
</body>
```

mit TEI_all Schema: Validierung erfolgreich

mit tei-epidoc:

test.xml, schema "**tei-epidoc.rng**" (3 Elemente)

- E [Jing] element "supplied" missing required attribute "reason"
- E [ISO Schematron] space may have @quantity (a figure) or @extent (a descriptive text value) but not both
- E [ISO Schematron] If space has @quantity then @unit is required

Probleme

Users/andreasmertgens/Downloads/test.xml — Validierung fehlgeschlagen. Fehler: 3.

Eine sehr große spezialisierte Nische:

EpiDoc: Epigraphic Documents in TEI XML

Magica Levantina 8 Oxygen

Eine kleine Nische

Computer-Mediated Communication and Social Media Corpora.

CMC-TEI

```
<message id="74" type="utterance" creator="muji_mit_mops" color="#4A29D6">
    <messageHead>
        <nickname>muji_mit_mops</nickname>
    </messageHead>
    <messageBody> Es werden immer mehr <emoticon>;-)</emoticon>
    </messageBody>
</message>
```

https://cmc-corpora.org/ckcmc/docs/tei/

Wir verlassen die Welt der TEI -> MEI

You are now leaving TEI.

Music Encoding Initiative

- 1999 erstes XML Schema von Perry Roland für Musik Notation
- MEI bewusst als analoge Organisation zur TEI strukturiert
- Community von Researchers, Developers -> MEI Schema
- Seit 2013 an der Akademie der Wissenschaften und Literatur in Mainz gehostet.

https://beethovens-werkstatt.de/glossary/mei/

https://music-encoding.org/guidelines/v5/MEI_Guidelines_v5.0.pdf

Wir verlassen die Welt der TEI -> MEI

You are now leaving TEI.

Music Encoding Initiative

- Enger Kontakt zur TEI Community, Überschneidungen
- Encoding Cultures – joint MEC and TEI Conference 2023
- Editionsprojekte mit TEI und MEI

  https://weber-gesamtausgabe.de

Vielen Dank für Ihre Aufmerksamkeit!

BERGISCHE
UNIVERSITÄT
WUPPERTAL

BERGISCHE
UNIVERSITÄT
WUPPERTAL

Xproc

```
<p:declare-step xmlns:p="http://www.w3.org/ns/xproc"
                name="xinclude-and-validate"
                version="3.0">
  <p:input port="source"/>
  <p:input port="schemas" sequence="true"/>
  <p:output port="result"/>
  <p:choose>
    <p:when test="/*[@version &lt; 2.0]">
      <p:validate-with-xml-schema>
        <p:with-input port="schema" href="v1schema.xsd"/>
      </p:validate-with-xml-schema>
    </p:when>

    <p:otherwise>
      <p:validate-with-xml-schema>
        <p:with-input port="schema" href="v2schema.xsd"/>
      </p:validate-with-xml-schema>
    </p:otherwise>
  </p:choose>

  <p:xslt>
    <p:with-input port="stylesheet" href="stylesheet.xsl"/>
  </p:xslt>
</p:declare-step>
```