Erstellen und Verwenden eines projektspezifischen TEI-Datenschemas

Susanne Haaf
Deutsches Textarchiv, BBAW Berlin
www.deutschestextarchiv.de
haaf@bbaw.de

Schema allgemein

- Inventar von Elementen
- Verhältnis der Elemente zueinander
- Regeln, in welchen Umgebungen welche Annotationen möglich sind

- Umsetzung: DTD, XSD, RNG

Beispiele:

- TEI-Schema: tei_all
  http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng

- TEI-Element: <fileDesc>
  http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-fileDesc.html

Schema-Spezifikation für TEI/P5

- Problem: Das tei_all-Schema ist (intentional) sehr flexibel

- Beispiel: Eigennamentagging

  → verschiedene Möglichkeiten, in TEI Personennamen auszuzeichnen
  - Element: <rs> (referencing string) → <rs type="propNounPersName">
  - Element: <name> (name, proper noun) → <name type="person">
  - Element: <persName> (personal name)

Schema-Spezifikation für TEI/P5

- "[...] 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.

TEI-Empfehlungen für Spezifikationen

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

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

The following customizations use features which are not available in the DTD and XSD schema formats.

Odds	TEI for authoring ODD, includes the TEI module for creating ODD files and customizations	ODD RNG
allPlus	TEI with with all modules included, plus all external additions	ODD RNG
TEI + SVG	TEI with SVG	ODD RNG
TEI + Math	TEI with MathML	ODD RNG
TEI + XInclude	TEI with XInclude	ODD RNG

Customizations provided by the TEI community

To have your customization listed, please contact web@tei-c.org.

Best Practices for TEI in Libraries

A guide for mass digitization, automated workflows, and promotion of interoperability with XML using the TEI (website, prose documentation, ODD files)

EpiDoc: Epigraphic Documents in XML

A P5 customization for encoding epigraphic materials (Guidelines, SourceForge site, ODD)

Digital Archive of Letters in Flanders (DALF)

A P4 customization for detailed encoding of letters (Project site and DTD files)

TEI: Customization

- ODD: „One document does it all“
- Grundlage für die TEI Guidelines, bestehend aus Dokumentation, Beispielen und den formalen Deklarationen (für Elemente, Attribute, Module, ...)
- ODD nutzt das **tagdocs**-Modul → Inventar zur Anpassung der TEI Guidelines an die eigenen Bedürfnisse
- kein direkter Eingriff in das Schema der TEI, sondern übergeordnete Anpassungsregeln

- Hilfe beim Erstellen eines ODD-Dokuments: Roma
  - als Webapplikation: http://www.tei-c.org/Roma/
  - zum Herunterladen: http://wiki.tei-c.org/index.php/Roma, http://wiki.tei-c.org/index.php/Vesta

TEI: Customization

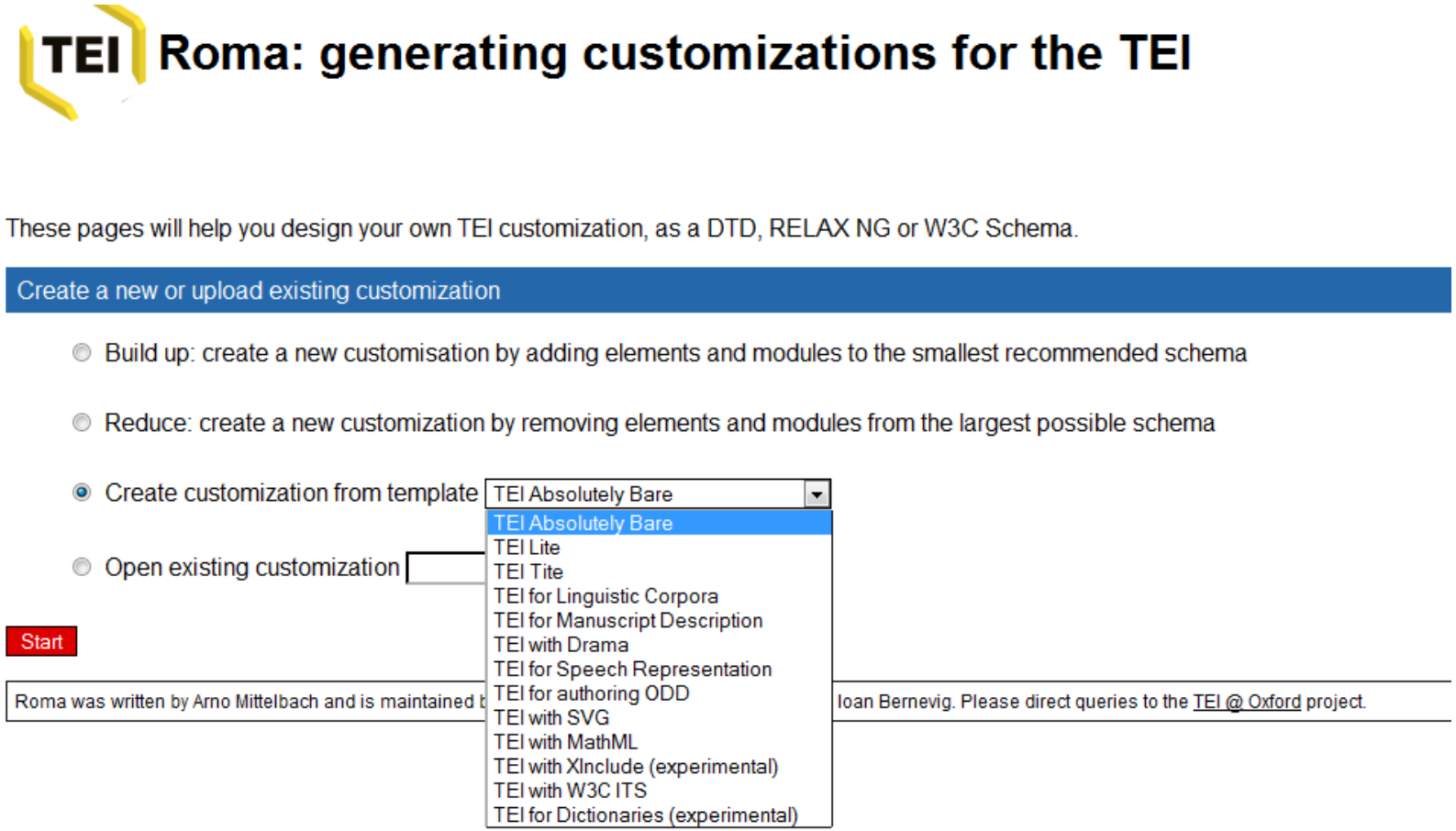
1. Elemente entfernen
2. Elemente umbenennen
3. Content models von Elementen oder Klassen modifizieren
4. Attribut- und Werteauswahl für Elemente und Klassen modifizieren
5. Klassenzugehörigkeiten modifizieren
6. neue Elemente (zu bestehenden Klassen) hinzufügen

(Vgl. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/USE.html#MDMD)

Mögliche Grundaktionen: add, delete, change, replace

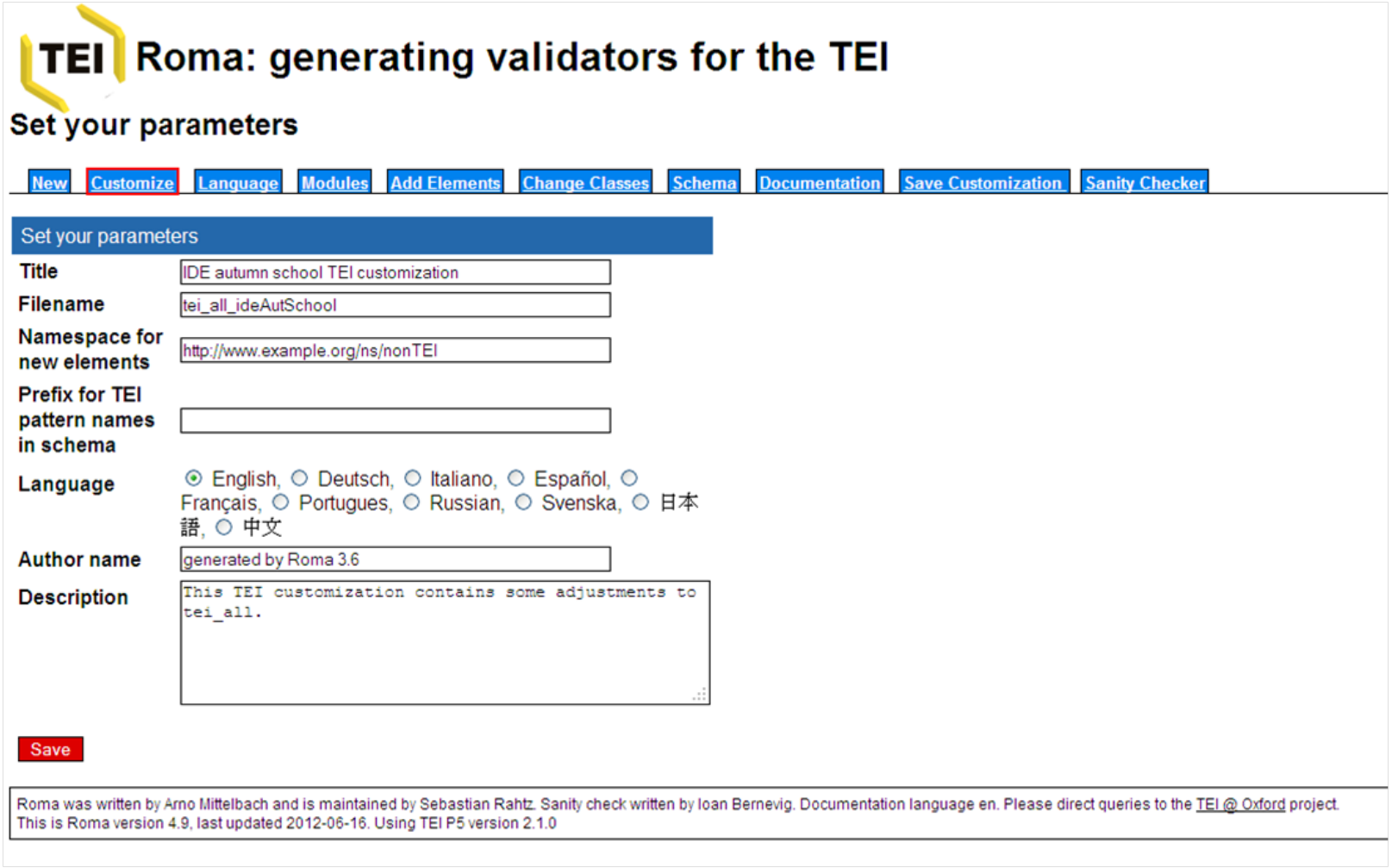
Achtung! Spezifikation vs. Veränderung des TEI-Schemas: Für die Austauschbarkeit von TEI-Dokumenten ist es wünschenswert, dass TEI-Dokumente in Bezug auf tei_all gültig sind.

Roma: Startpunkt auswählen



→ Bsp.: Ausgangspunkt tei_all (= „Reduce: ...“)

Roma: Metadaten für das ODD spezifizieren



Module und Elemente auswählen

- Roma:
  - Module: „Modules“ > „List of selected Modules“
  - Elemente: „List of elements in module:[Modulname]“
- ODD:
  - Module: <moduleRef> (module reference) mit @key
  - Elemente: @except oder @include in <moduleRef>

- Beispiel:

```
<moduleRef key="textstructure" except="div1 div2 div3 div4 div5 div6 div7 group"/>
<moduleRef key="drama" except="camera caption epilogue move performance prologue set sound spGrp tech view"/>
```

- *Achtung!* Module ohne <moduleRef> im ODD werden nicht eingebunden in das Schema.

Roma: Module auswählen

Modules

New | Customize | Language | Modules | Add Elements | Change Classes | Schema | Documentation | Save Customization | Sanity Checker

List of TEI Modules

	Module name		A short description	Changes
add	analysis	?	Simple analytic mechanisms	
add	certainty	?	Certainty and uncertainty	
add	core	?	Elements common to all TEI documents	
add	corpus	?	Corpus texts	
add	dictionaries	?	Dictionaries	
add	drama	?	Performance texts	
add	figures	?	Tables, formulæ, notated music, and figures	
add	gaiji	?	Character and glyph documentation	
add	header	?	The TEI Header	
add	iso-fs	?	Feature structures	
add	linking	?	Linking, segmentation and alignment	
add	msdescription	?	Manuscript Description	
add	namesdates	?	Names and dates	
add	nets	?	Graphs, networks, and trees	
add	spoken	?	Transcribed Speech	
add	tagdocs	?	Documentation of TEI modules	
add	textcrit	?	Critical Apparatus	
add	textstructure	?	Default text structure	
add	transcr	?	Transcription of primary sources	
add	verse	?	Verse structures	

List of selected Modules

tei

remove	core
remove	analysis
remove	certainty
remove	corpus
remove	dictionaries
remove	drama
remove	figures
remove	gaiji
remove	header
remove	iso-fs
remove	linking
remove	msdescription
remove	namesdates
remove	nets
remove	spoken
remove	textcrit
remove	textstructure
remove	transcr
remove	verse
remove	tagdocs

Liste der verfügbaren TEI-Module

Liste der für die eigene Spezifikation ausgewählten TEI-Module

Roma: Elemente auswählen

Change module

New | Customize | Language | Modules | Add Elements | Change Classes | Schema | Documentation | Save Customization | Sanity Checker

back

List of elements in module:namesdates

	Include	Exclude	Name		Description	Attributes
addName	●	○	addName	?	(additional name) contains an additional name component, such as a nickname, epithet, or alias, or any other descriptive phrase used within a personal name.	Change attributes
affiliation	○	●	affiliation	?	(affiliation) contains an informal description of a person's present or past affiliation with some organization, for example an employer or sponsor.	Change attributes
age	○	●	age	?	(age) specifies the age of a person.	Change attributes
birth	○	●	birth	?	(birth) contains information about a person's birth, such as its date and place.	Change attributes
bloc	○	●	bloc	?	(bloc) contains the name of a geo-political unit consisting of two or more nation states or countries.	Change attributes
climate	○	●	climate	?	(climate) contains information about the physical climate of a place.	Change attributes
country	○	●	country	?	(country) contains the name of a geo-political unit, such as a nation, country, colony, or commonwealth, larger than or administratively superior to a region and smaller than a bloc.	Change attributes
death	○	●	death	?	(death) contains information about a person's death, such as its date and place.	Change attributes
district	○	●	district	?	contains the name of any kind of subdivision of a settlement, such as a parish, ward, or other administrative or geographic unit.	Change attributes
education	●	○	education	?	contains a description of the educational experience of a	Change attributes

Liste der Elemente eines gewählten Moduls

Module und Elemente auswählen

- Nötiges Vorwissen:
  - In welchen Modulen sind Elemente enthalten, die ich benötige?
    Welche Module kann ich hingegen ganz eliminieren?
  - In welchen der von mir ausgewählten Module sind die Elemente enthalten, die ich verändern möchte?
- Hilfreich dabei: TEI-Elemente gruppiert nach Modulen
  http://www.tei-c.org/release/doc/tei-p5-doc/en/html/REF-ELEMENTS.html
  → „Show by Module“

Module und Elemente auswählen – Beispiel

Beispiel: Eigennamen (vgl. Folie 3)

- Modul „namesdates“ (und somit auch das in diesem Modul enthaltene Element <persName>) entfernen
  - Roma: „Modules“ → in der „List of selected modules“: „remove namesdates“
  - ODD: auf eine <moduleRef key="namesdates"/> verzichten

- Element <rs> (aus dem Modul „core“) entfernen
  - Roma: „Modules“ → in der „List of selected modules“ das Modul „core“ auswählen; die „List of elements in module:core“ erscheint; dort mittels „Exclude“ das Element <rs> aus dem Modul entfernen
  - ODD: <moduleRef key="core" except="rs"/>

Module und Elemente auswählen – Beispiel

Beispiel: Eigennamen (vgl. Folie 3)

- Element <name> (aus dem Modul „core“) für die Auszeichnung von Eigennamen erhalten
  - Roma: „Modules“ → in der „List of selected modules“ das Modul „core“ auswählen; die „List of elements in module:core“ erscheint; dort mittels „Include“ das Element <name> in dem Modul erhalten
  - ODD:
    - <moduleRef key="core"/>: erlaubt sämtliche Elemente des Moduls „core“
    - <moduleRef key="core" insert="name p quote"/>: erlaubt nur die genannten Elemente des Moduls „core“ (also auch <name>)
    - <moduleRef key="core" except="rs q said"/>: erlaubt alle Elemente des Moduls „core“, außer den genannten (also auch <name>)

Attribute und Werte auswählen

- Die laut TEI erlaubten Attribute können eliminiert oder hinsichtlich ihrer Werte konkretisiert werden

  → Klassenweise (und somit Element-übergreifend)

  - Roma: „Change Classes“ → für jede Klasse unter „Change attributes“ mittels „Include“ Attribute einschließen oder sie mittels „Exclude“ entfernen; nähere Festlegungen zur Attributauswahl können von dort aus für jedes Attribut unter „Change attribute“ vorgenommen werden
  - ODD: <classSpec> (class specification)

    Beispiel:

```
<classSpec ident="att.datcat" module="tei" type="atts" mode="delete"/>

<classSpec ident="att.typed" module="tei" type="atts" mode="change">
  <attList>
    <attDef ident="subtype" mode="delete"/>
  </attList>
</classSpec>
```

Attribute und Werte auswählen

- Die laut TEI erlaubten Attribute können eliminiert oder hinsichtlich ihrer Werte konkretisiert werden

  → Elementweise

  - Roma: „Modules“ → ... „List of elements in module:[X]“ → „Change Attributes“
  - ODD: <elementSpec> (element specification) → <attDef> (attribute definition)

    Beispiel:

```
<elementSpec ident="figure" module="figures" mode="change">
    <attList>
        <attDef ident="place" mode="delete"/>
        <attDef ident="type" mode="change" usage="opt">
            <valList type="closed" mode="replace">
                <valItem ident="notatedMusic"/>
            </valList>
        </attDef>
        <attDef ident="n" mode="delete"/>
    </attList>
</elementSpec>
```

Attribute und Werte auswählen

- Hilfreich dabei: die Auflistung sämtlicher Elemente, die von einer bestimmten Attributklasse Gebrauch machen

  → z.B. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-att.global.html

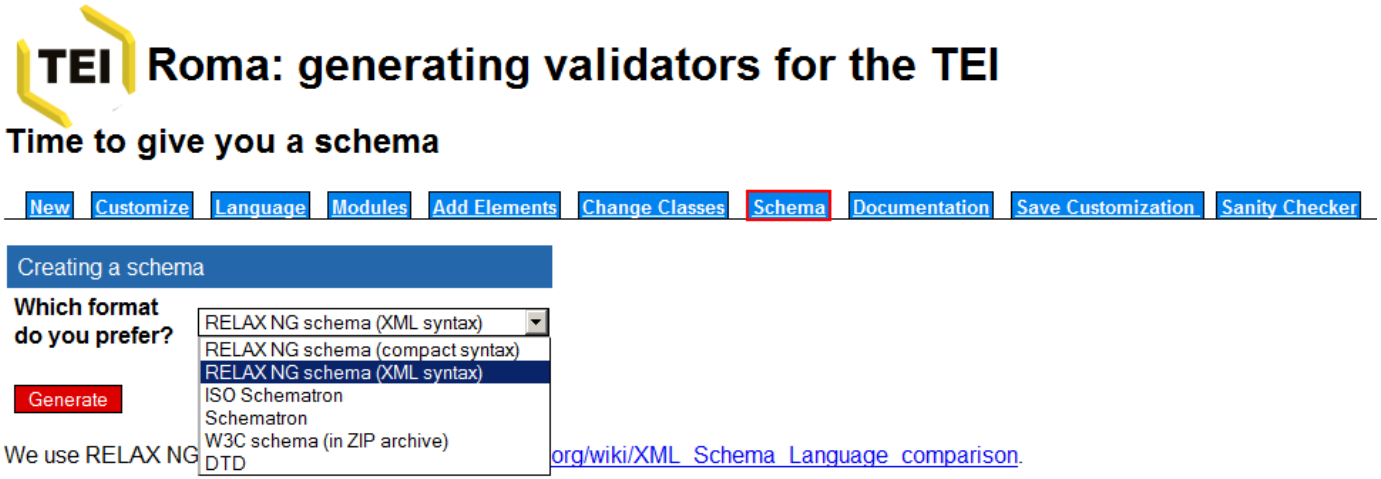
- *Übrigens*: Übersicht über die Attributklassen in den TEI Guidelines
  http://www.tei-c.org/release/doc/tei-p5-doc/en/html/REF-CLASSES-ATTS.html

Schema erstellen mit Roma

- ODD auf Fehler überprüfen mit dem „Sanity Checker“
- ODD speichern unter „Save customization“

  Achtung! Das ODD sollte immer abgespeichert werden, damit die bestehende Spezifikation späteren möglichen Anpassungen zugrunde gelegt werden kann.
- Schema erstellen unter „Schema“

  → Auswahl des bevorzugten Schema-Formats



Arbeiten mit dem erstellten Schema

- Das Schema wird an zweiter Stelle (nach der XML-Deklaration) in die TEI-Datei eingebunden mit dem Ausdruck:

  <?xml-model href="[URL zum Schema]" schematypens="[Namensraum]"?>

- z.B. das RNG von tei.all:

  <?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng" schematypens="http://relaxng.org/ns/structure/1.0"?>

- Im oXygen kann das Schema auch durch Verwendung des Menüs „Dokument“ > „Schema“ > „Schema zuweisen“ eingebunden werden:

  → Pfad zum Schema angeben, Schematyp festlegen, OK

  → Resultat: Schema wird an der richtigen Stelle in der Datei eingebunden

ODD: erster Überblick

- <schemaSpec> (schema specification):
  - Wurzelelement für die Spezifikationen
- <moduleRef> (module reference):
  - Referenz auf ein Modul, welches somit in das Schema eingebunden wird
  - Möglichkeit, die Auswahl der Elemente je Modul zu bestimmen
- <classSpec> (class specification):
  - Klassenweise über Elemente oder Attribute entscheiden
- <elementSpec> (element specification):
  - Ein Element spezifizieren hinsichtlich seiner Attribute (<attList><attDef>), Klassenzugehörigkeit (<classes><memberOf>), seines Inhalts (<content>)
- Elemente zur Dokumentation, z.B. <desc>, <remarks>, <exemplum>

→ mehr: Kapitel 22 und 23.2 der TEI/P5 Guidelines

Aufgabe

- Erstellen Sie eine Element-Spezifikation für das Element !
- Es soll die folgenden Attribute enthalten:
  - @type
  - @n
- Das Attribut @type soll eine festgelegte Menge an möglichen Werten erhalten, um verschiedene Typen von Textabschnitten zu definieren: Brief, Akt und Szene eines Dramas, Register, Inhaltsverzeichnis, Vorwort, Anhang.
- Für das Attribut @n soll es keine feste Werteliste geben. Allerdings soll der mögliche Datentyp für Werte festgelegt werden: nur ganze Zahlen sollen als Werte möglich sein.

Lösung – Teil 1

```
<elementSpec ident="div" module="textstructure" mode="change">
  <attList>
    <attDef ident="n" mode="change">
      <datatype minOccurs="1" maxOccurs="1">
        <rng:ref name="data.count"/>
      </datatype>
    </attDef>
    <attDef ident="ana" mode="delete"/>
    <attDef ident="change" mode="delete"/>
    <attDef ident="copyOf" mode="delete"/>
    <attDef ident="corresp" mode="delete"/>
    <attDef ident="[weitere Attribute]" mode="delete"/>
```

d.h. es darf nur genau ein Wert für das Attribut angegeben werden (keine Werteliste)

Achtung! Gegebenenfalls können einige der Attribute bereits innerhalb der <classSpec> zur jeweiligen Attributklasse aus dem Schema entfernt werden, sodass auf Elementebene nicht noch einmal über sie entschieden werden muss.

Lösung – Teil 2

```
<attDef ident="type" mode="change" usage="req">
 <valList type="closed" mode="replace">
  <valItem ident="letter"/>
  <valItem ident="act"/>
  <valItem ident="scene"/>
  <valItem ident="index"/>
  <valItem ident="contents"/>
  <valItem ident="preface"/>
  <valItem ident="appendix"/>
 </valList>
</attDef>
</attList>
</elementSpec>
```

d.h. das Attribut @type ist obligatorisch bei Verwendung des Elements ; für fakultativ: usage="opt"

d.h. es dürfen nur Werte aus der angegebenen Werteliste verwendet werden; für eine offene Wertemenge: type="open"

→ Vgl. auch die Spezifikation für das Element im ODD des DTA-Basisformats:
www.deutschestextarchiv.de/basisformat.odd

Literatur und Links

- TEI Guidelines zu ODD und Customization: Kapitel 22 und 23
- Einführung zu ODD: http://www.tei-c.org/Guidelines/Customization/odds.xml
- ODD-Wiki: http://wiki.tei-c.org/index.php/ODD
- Roma: http://www.tei-c.org/Roma/

- Zum DTA-Basisformat im Vergleich mit anderen TEI-Formaten: *Geyken/Haaf/Wiegand*, The DTA 'base format': A TEI-Subset for the Compilation of Interoperable Corpora (Proceedings der Konvens 2012, http://www.oegai.at/konvens2012/proceedings.pdf#page=383)
- Dokumentation des DTA-Basisformats inkl. ODD und RNG: www.deutschestextarchiv.de/doku/basisformat

Vielen Dank!