TEI für Register und Sacherschließung

- Einheit ist eine Vertiefung in TEI
- Regularisierung und Normalisierung weiterer Schwerpunkt
- Es sollen Möglichkeiten zur Verarbeitung der so bearbeiteten Dokumenten deutlich werden
- Übungsmaterialien für die Schulung unter: http://nowalkowski.de/share/springschool2012/

Viel Spaß!

Überblick

- Regularisierung im Text
- Regularisierung von Entitäten und Sacherschließung (Einschub: Pointing)
- Indexerstellung
- Normalisierung
- Normdateien und Anbindung an externe Ressourcen

Regularisierung im Text

- Ausgangssituation: Regularisierung als Teil des Editionsprozesses
- 2 Möglichkeiten
  - Dokumentation im TEI Header in der `<editorialDecl />` in der `<encodingDesc />`
  - `correction hyphenation interpretation normalization quotation segmentation`


```
<editorialDecl>
  <normalization>
    <p>All words converted to Modern American spelling using
       Websters 9th Collegiate dictionary
    </p>
  </normalization>
</editorialDecl>
```

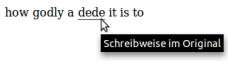
Verwendung der Elemente

- `<reg>` Kennzeichnung einer vorgenommenen Regularisierung
- `<orig>` Kennzeichnung einer Schreibweise im Original
- `<choice>` Gruppierung unterschiedlicher Schreibweisen für eine Textstelle

<orig>

```
<p>how godly a <orig>dede</orig> it is to <orig>overthrowe</orig>
    so wicked a race the world may judge: for my part I
    <orig>thinke</orig> there <orig>canot</orig>
    be a greater<orig>sacryfice</orig> to God
</p>
```

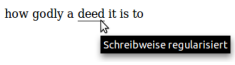
- Kennzeichnung des Bearbeitungsstandes
- Kennzeichnung bei der Anzeige



<reg>

```
<?xml version="1.0" encoding="UTF-8"?>
<p>how godly a <reg>deed</reg> it is to <reg>overthrow</reg>
    so wicked a race the world may judge: for my part I
    <reg>think</reg> there <reg>cannot</reg>
    be a greater <reg>sacrifice</reg>
    to God.
</p>
```

- Kennzeichnung einer Schreibweise als ediert
- Interpretation, unleserliche Quelle



<choice>

- Anbieten von Alternativen
- Rekonstruierbarkeit
- Anbieten unterschiedlicher Editionen im Web


```
<?xml version="1.0" encoding="UTF-8"?>
<p>how godly a
    <choice>
        <orig>dede</orig>
        <reg>deed</reg>
    </choice> it is to
    <choice>
        <orig>overthrowe</orig>
        <reg>overthrow</reg>
    </choice> so wicked a race the world may judge: for my part I
    <choice>
        <orig>thinke</orig>
        <reg>think</reg>
    </choice>there <choice>
        <orig>canot</orig>
        <reg>cannot</reg>
    </choice> be a greater <choice>
        <orig>sacryfice</orig>
        <reg>sacrifice</reg>
    </choice> to God.
</p>
```


@cert und @resp

- Qualifizierung des Editionsprozesses
- @cert Angaben zur Sicherheit der Lesart
  - mögliche Werte: unknown; low; medium; high
- @resp Angabe des Bearbeiters oder einer Bearbeitenden Einrichtung

```
<reg cert="unknown" resp="Niels-Oliver Walkowski">deed</reg>
```


Identifizierung von Entitäten und Sacherschließung

- Auszeichnung von Entitäten durch <name type="xxx" />
- Schwierigkeiten bei der Identifikation von Entitäten
- Identifizierung durch @key
  - Be

```
<p><name type="person" key="P103">Mühsam</name> wurde in
    Berlin als Kind jüdischer Eltern geboren und wuchs in Lübeck auf.
    .
    .
    .

Am 11. Januar 1896 wurde <name type="person" key="P103">Erich</name>
    von der Schule wegen „sozialdemokratischer Umtriebe“ verwiesen.</p>
```

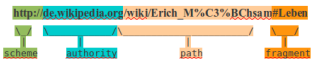

Identifizierung, Regularisierung und Erschließung durch

Einschub: Pointing

- Verweisen von einer Stelle im Dokument auf eine andere
  - Interne Stelle
  - Externe Stelle
- Begriffsverwirrung: Pointing vs. Linking
- Abbildung von Strukturen im Text jenseits der linearen Textstruktur

Der URI und seine Bestandteile

- Die URL als Beispiel einer URI
  - http://de.wikipedia.org/wiki/Erich_M%C3%BChsam
- Bestandteile einer URI (für unseren Zweck)



- Verkürzung bei internen Links #Leben
- aber wo ist das Leben? xml:id

```
<head n="2" xml:id="Leben">Leben</head>
```

… zurück zum Thema: `@ref...`

- `@ref` verweist auf eine Ressource (intern oder extern), die die Entität beschreibt
- Wert von `@ref` ist eine URI

```
<p><name type="person" key="P103" ref="#e_muehsam">Mühsam</name>
```

- Vorteile einer zentralen Beschreibung

```
<profileDesc>
    <particDesc>
        <listPerson type="historical">
            <person xml:id="e_muehsam">
                <persName>Erich Mühsam</persName>
            </person>
        </listPerson>
    </particDesc>
</profileDesc>
```

- Gruppen von Personen: `@ref="#e_muehsam #b_brecht"`

…und `@nymRef`

- Der kanonische Name einer Person
- Beschreibung der Namensformen statt der Biografie

```
<forename nymRef="#N123">Tony</forename> Blair.
```

- Erfassung in unterschiedlichen Listen in der `<sourceDesc />` im TEI Header (`model.ListLike`)

```
<listNym>
    <nym xml:id="N123">
        <form>Antony</form>
    </nym>
</listNym>
```


Sprachen und Etymologie

- Namensformen in verschiedenen Sprachen: `<orth xml:lang="de-DE" />`

```
<nym xml:id="J451">
    <form>
        <orth xml:lang="en-US">Ian</orth>
        <orth xml:lang="en-x-Scots">Iain</orth>
    </form>
```

- Etymologische Beschreibungen durch Vokabular aus dem `model.dictionary`

Sacherschließung im TEI Header

- `<listPerson />` in der `<particDesc />` der `<profileDesc />` zur Beschreibung von Personen Text
- `<listPlace />` in der `<settingDesc />` der `<profileDesc />` zur Beschreibung von Orten
- Mehrere mittels `@type` typisierte Listen möglich
- Beziehungen zwischen Personen oder Orten mittels `<relation />`

`<listPerson />`

- Beschreibung mittels Elementen aus model.personLike: `affiliation age birth death education event faith floruit langKnowledge nationality ….`

```
<person xml:id="karlIV">
    <persName>Karl IV., getauft auf den Namen Wenzel</persName>
    <birth>1316</birth>
    <death>1378</death>
    <faith>römisch-katholisch</faith>
</person>
```


`<listPlace />`

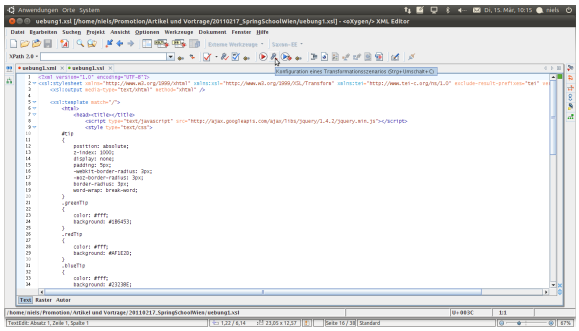
- Beschreibung mittels Elementen wie: `bloc climate country district event geogName listPlace location place placeName population region settlement state terrain`

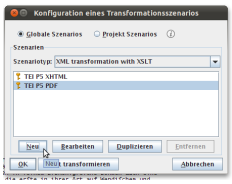
```
<place xml:id="prag">
    <placeName>Prag</placeName>
    <country>Tschechien</country>
</place>
```

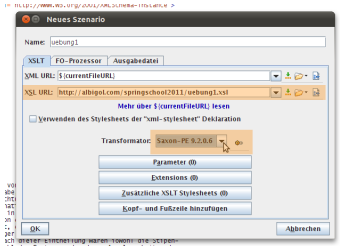
Übung 1

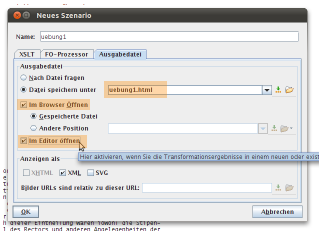
- Der Textausschnitt enthält veraltete Schreibweisen, Personennamen und Ortsnamen
- Arbeiten Sie den Text durch und zeichnen Sie aus:
  - Alte Schreibweisen unter Angabe der gängigen Schreibweise
  - Personen und Orte mit dem <name /> Element
  - Erstellen Sie eine Personen und Ortsliste im TEIHeader und referenzieren Sie mit @ref dorthin

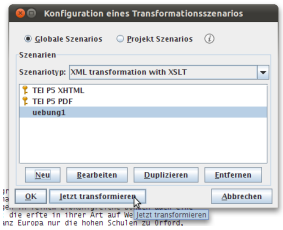
Anwenden eins Skripts











Indizes und Register

- Zwei verschieden Ausgangssituationen:
  - Abbildung eines Registers aus der Quelle
  - Erstellung eines Register während der Erstellung der digitalen Edition

Abbildung eines bestehenden Registers

- Kapselung des Registers in einem `<div />` Element (`@type="index"` möglich)
- Ein Register ist ein Liste. Gebrauch von `<list />` `<item />`

Verweise im `<ref />` Element

```
<div type="index">
  <!--...-->
  <list type="index">
    <item>Ahlhaus<ref>193</ref></item>
    <item>Bendtler<ref>443</ref></item>
    <item>Brachinger<ref>181</ref></item>
    <item>Dürrematt<ref>212</ref></item>
  </list>
</div>
```

Pointing

- Auf die entsprechende Passage kann direkt gezeigt werden.
  - Voraussetzung: identifizierbare Pagebreaks
  - Voraussetzung: Gebrauch des `@target` Attributs von `<ref />`

```
<list type="index">
    <item>Faas-Hardegger, Margarethe<ref target="#p346">343</ref></item>
</list>
<p>sondern auch die Mutterschaftsversicherung und die
    Idee von bezahlter Hausarbeit.<pb xml:id="p346" />
    Margarethe Hardegger machte eine Lehre als Telefonistin,
    gleich anschliessend holte sie mit Unterstützung ihres
    späteren Ehemanns, August Faas, die Matura nach.</p>
```


Mehrdimensionale Register

- Register mit mehreren Hierarchiestufen
- Lösungsansatz: eine verschachtelte Liste

```
<list type="index">
  <item>Agriculture
    <list type="indexentry">
      <item>ancient policy of Europe unfavourable to, <ref>371</ref></item>
      <item>cattle and tillage mutually improve each other, <ref>325</ref></item>
      <item>wealth arising from more solid than that which proceeds
        from commerce <ref>520</ref></item>
    </list>
  </item>
</list>
```


Übung 2

- Schauen Sie sich die Personenregistervorlage an und versuchen Sie diese in den <back /> Abschnitt des XML Dokuments einzuarbeiten
- Verwenden Sie Pointing um auf die angegebenen Pagebreaks zu verweisen.

Automatische Generierung eins Registers

- Vorgehen
  - Auszeichnung der zu indizierenden Terme oder Passagen
  - Automatische Erstellung des Registers durch ein Skript
- Verortung manuell oder automatisch nachträglich ersteller Register in der `<back />` Sektion des TEI Dokuments
- Benutzung von `<divGen />` statt `<div />` für automatisch erstellte Register

Auszeichnung zu indexierender Textstellen

- `<index />` Element zur Auszeichnung von zu indexierenden Textstellen
- `<term />` Element zur Angabe des Index Terms

```
<p>David's other principal backer, Josiah ha-Kohen
  <index>
    <term>Josiah ha-Kohen b. Azarya</term>
  </index> b. Azarya, son of one of the last gaons of Sura.
</p>
```


Auszeichnung von Textpassagen

- Vorgehen
  - Verweis auf eine Markierung im Text, bis zu der die Passage geht
- Verweis im `<index />` Element durch `@spanTo`
- Verweis auf einene gesetzte Markierung mittels des `<anchor />` Elements

```
<index spanTo="#ALAMEND">
  <term>Lemmatization, Arabic</term>
</index> concerning which it is important to note .....

<!-- much learned material omitted here -->
and now we can build our parser.<anchor xml:id="ALAMEND"/>
```

Sonderzeichen und mehrere Rgister

- Problem: Sortierung von Sonderzeichen
- Lösung: `@sortKey` im `<item />` Element
- Problem: Registerzugehörigkeit bei mehreren Registern
- Lösung: `@indexName` im `<index />` Element

```
<p>The Svenska Sångarförbundet
    <index sortKey="Sang" indexName="INDEX-INSTITUTIONS">
      <term>Sångarförbundet</term>
    </index> was,coincidentally, founded
  </p>
```


Mehrdimensionale Register

- Problem: Wie werden Terme innerhalb einer Registerhierarchie im Text kenntlich gemacht?
- Lösung: Abbildung der Hierarchie durch Verschachtelung von `<index />` Elementen

```
<p>The students understand procedures for Arabic lemmatisation
    <index>
        <term>lemmatization</term>
        <index>
            <term>arabic</term>
        </index>
    </index>
</p>
```


Übung 3

- Identifizieren Sie im folgenden Text Schlagworte und markieren Sie sie mittels des <index /> Elements
- Benennen Sie mittels @indexName das Register in das die Schlagworte aufgenommen werden sollen
- Verwenden Sie @xml:id um die Auszeichnung referenzierbar zu machen

Normalisierung von Datentypen

- Zeitpunkte, Perioden und Orte
  - 24.12.2010, 24. Dezember 2010, der erste Tag des Beth
  - die attische Zeit, das erste römische Triumvirat
  - 7km nördlich von Berlin, 10 Meilen westlich von Newcastle
- Normalisierung zu Erläuterungszwecken
- Normalisierung zu Prozessierungszwecken

Datum und Zeit

- Kenntlichmachung durch `<date />` und `<time />`
- `@when, @notBefore, @notAfter, @from, @to`
- Notation im W3C Standard: `YYYY-MM-DDTHH:MM:SS`

```
<p>
    Camus fuhr am <date when="1960-01-04">4. Januar</date>
    gegen einen Baum und war sofort tot. Der Krankenwagen traf
    um <date when="1960-01-04T15:32:00">15:32</date> am Unfallort ein.
</p>
```


`@period`

- Bestimmung von Epochen und Perioden
- Benutzung ähnlich wie bei Entitäten und `@ref`
- Inhalt ist eine URI die auf eine interne oder externe Ressource verweist

```
<placeName period="#christian">Stauropolis</placeName>
```

Deklaration von Perioden

- Beschreibung des Zeitabschnitts in der `<encodingDesc />` <classDecl /> im TEI Header

```
<taxonomy>
    <category xml:id="christian">
        <catDesc> The Christian period technically starts at the
            birth of Jesus, but in
            practice is considered to date from the conversion of Constantine
            in <date when="0312">312 AD</date>. </catDesc>
    </category>
</taxonomy>
```


Orte

- Benennungsproblem und Bestimmungsproblem
- Unterelement von `<place />`
- Bestimmung der Position durch `<location />`
- Kontextualisierung mit `<block />`, `<country />` und `<settlement />`

```
<location>
    <bloc>EU</bloc>
    <country>France</country>
</location>
```

```
<location>
    <country key="FR"/>
    <settlement type="city">Lyon</settlement>
    <district type="arrondissement">Perrache</district>
    <placeName type="street">Rue de la Charité</placeName>
</location>
```


Geolokalisierung

- Verwendung von Geokoordinaten nach WGS84

```
<location>
    <geo>45.769559 4.834843</geo>
</location>
```

- Verwendung anderer Geolokalisierungssysteme wie KML durch Auszeichnung in der `<geoDecl />` im TEI Header

```
<encodingDesc>
    <geoDecl datum="KML" />
</encodingDesc>
```


Einheiten `<measure>`

- Auszeichnung von Maßen
- Verknüpfung von Mengenangaben mit Werten in gebräuchlichen Maßeinheiten

```
<measure type="currency">£10-11-6d</measure>
<measure type="area">2 merks of old extent</measure>
<measure quantity="1.6" unit="km">1 Meile</measure>
```


Normdateien: PND und GeoNames

- Warum Normdateien?
- Schwierigkeiten bei der Verwendung von Normdateien

Personennormdatei - PND

- Bereitstellung eines Identifikators für die eindeutige Indentifikation von Personen
- Bestimmung der Person durch Metadaten
- Namensansetzungen
- Ursprung: bibliothekarischer Kontext: DNB
- Harmonisierung durch VIAF
- PND im nationalen Kontext etabliert, z.B Wikipedia
- Link-Resolver: http://toolserver.org/~apper/pd/person/pnd-redirect/de/PND
- Bsp.: Adolf Friedrich (schwedischer König)

GeoNames

- Frei zugängliche DB geographischer Orte mit
- Namensvariationen, Geodaten, GoogleMaps und Wikipediaverlinkung, Umgebungssuche und mehr
- Ambiguitäten: http://www.geonames.org/search.html?q=Berlin
- Eindeutige Identifizierung:
- http://www.geonames.org/2950159/

Übung 4

- Recherchieren Sie für die im TEI Dokument enthaltenen Orte und Personen die jeweilige GeoNames ID, bzw. PND im Internet
- Annotieren Sie die Personen und Orte mit <name type="person" /> oder <name type="place" />
- Tragen Sie die ID in ein hinzuzufügendes @key Attribut ein