Coding the Sources

Digitales Edieren in den Geisteswissenschaften

Online Summerschool und Workshop

Lübeck/Online, 17.08. – 27.08.2020

Forschungsstelle
für die Geschichte
der Hanse und des Ostseeraums

i|d|e

Christian-Albrechts-Universität zu Kiel
Historisches Seminar

Metadaten in TEI

<teiHeader>

Torsten Roeder
torsten.roeder@leopoldina.org
Lübeck/Online, 21.08.2020

Forschungsstelle
für die Geschichte
der Hanse und des Ostseeraums

Christian-Albrechts-Universität zu Kiel
Historisches Seminar


teiHeader

Inhalte

- Was sind Metadaten?
- Wie ist der teiHeader aufgebaut?
- Welche Elemente umfasst der teiHeader?

Voraussetzungen

- Was ist XML?
- Was ist TEI?

Metadaten?

- strukturierte Daten
- enthalten Informationen zu Merkmalen anderer Daten
- können aus den Daten direkt hervorgehen
- oder auf anderen Wegen ermittelt werden

Metadaten!

- gut miteinander vergleichbar
- fast immer standardisiert
- meistens gut maschinenlesbar
- geeignet für Weiterverarbeitung
- erzeugen viel Geld und/oder Macht
  - nur nicht in der Wissenschaft, aber:
    - sie sind Teil der Dokumentation
    - gehören zur guten wissenschaftlichen Praxis
    - tragen zum Erkenntnisprozess bei

Forschungsdaten?!

- Daten, die im Zusammenhang mit wissenschaftlichem Erkenntnisgewinn entstehen
- Forschungsdaten-Management (FDM): gute Forschungsdaten brauchen gute Metadaten!

Brainstorming

- was könnten Metadaten einer TEI-Datei sein?
- welche Sorten von Metadaten ließen sich voneinander abgrenzen?

(nutzen Sie den Chat)

Der TEI-Header

- <teiHeader> ist obligatorisch
- ist immer das erste Kind von <TEI>
- ist immer das erste Kind von <teiCorpus>

Vergleich des <teiHeader>s mit anderen Strukturelementen

- <group> mit mehreren <text>en:
  - keine Metadaten zu den einzelnen Texten
- <front> und/oder <back>:
  - diese gehörten zur Quelle
  - <front> oder <back> sagen:
    "der Text auf dem Buchumschlag lautet ..."
  - der <teiHeader> hingegen sagt:
    "der Titel lautet ..."

Aufbau des <teiHeader>s

- **<fileDesc>** [obligatorisch]
- <encodingDesc>
- <profileDesc>
- <xenoData>
- <revisionDesc>

1. <fileDesc>

- **<titleStmt>** [*]
- <editionStmt>
- <extent>
- **<publicationStmt>** [*]
- <seriesDesc>
- <notesStmt>
- **<sourceDesc>** [*]

1.1 <titleStmt>

- identifiziert das Dokument und daran Beteiligte
  - <title>, <author>, <editor> ...
  - spezifisch: <principal>, <funder>, <sponsor>

1.2 <editionStmt>

- beschreibt die vorliegende Ausgabe des Textes
- ergänzend zu <titleStmt>
- optional, wird aber empfohlen

1.3
<extent>

- beschreibt den Umfang des Textes
- Maßeinheit ist frei wählbar
- sinnvoll zum Dateivergleich

1.4 <publicationStmt>

- Informationen zum Veröffentlichungsstatus und den Beteiligten
- kann und sollte auch Informationen zur Lizenz enthalten

1.5 <seriesDesc>

- beschreibt eine Reihe, zu welcher der vorliegende Text gehört
- ergänzend zu <titleStmt>

1.6
<notesStmt>

- Anmerkungen
- alles was wichtig ist, aber woanders keinen Platz gefunden hat

1.7 <sourceDesc>

- Quellenbeschreibung
  - <bibl>
    gedruckte Quellen
  - <msDesc>
    handschriftliche Quellen
  - u.v.a.

    (Details später)

2. <encodingDesc>

- beschreibt editorische Methoden und Richtlinien
- dokumentiert die eigene Systematik
- als internes Referenzsystem nutzbar
- hilfreich für Nachnutzung sowie für automatische Verarbeitung

2.1 <editorialDecl>

- <correction> Eingriffe am Text
- <normalization> Schreibweisen
- <punctuation> Satzzeichen
- <quotation> Anführungszeichen
- <hyphenation> Worttrennung
- <segmentation> Textuntergliederung
- <stdVals> Standardeinheiten
- <interpretation> Analyseschema

2.2
<refsDecl>

- beschreibt die verwendeten Referenzsysteme
- hilfreich, wenn mit Normdaten, IDs oder Registern gearbeitet wird
- Standort des Referenzsystems lässt sich schablonenartig beschreiben
- im Editionstext muss dann nicht mehr die vollständige URL genannt werden

2.3 encodingDesc/*

- <projectDesc> Projekt
- <samplingDesc> Textselektionsmethode
- <classDecl> Klassifikationssystem
- <geoDecl> geograph. Referenzsystem
- <unitDecl> z.B. historische Maße
- <fsdDecl> linguistische Features
- <metDecl> Versmetrik-Notation
- <variantEncoding> Varianz-Kodierung

u.v.a.

3. <profileDesc>

- Angaben über den Text
  - <abstract> zum Inhalt
  - <creation> zur Textentstehung
  - <langUsage> zum Sprachgebrauch
  - <textClass> zur Textgattung
  - <correspDesc> zu schematischen Beschreibung von Absender und Adressaten (Korrespondenzen, Briefe)
  - <calendarDesc> zum verwendeten Kalender
  - speziell zu Korpora:
    - <textDesc>, <particDesc> und <settingDesc>
  - speziell zu Handschriften:
    - <handNotes> zu unterschiedlichen Schreibern
    - <listTranspose> zu Verschiebungen im Text

4. <xenoData>

- Metadaten in anderen Formaten
- z.B. importierte Daten aus Katalogen
- auch geeignet für den späteren Export
- Data Loop:
  - Metadaten importieren
  - Metadaten verbessern
  - Metadaten zurückspielen

5. <revisionDesc>

- Beschreibung des Bearbeitungsverlaufs
- unterschiedliche Methoden möglich
- sehr empfohlen!

<bibl>

- Beschreibung gedruckter Quellen
- Varianten:
  - <bibl> = praktisch ein Freitextfeld, einzelne Elemente können bei Bedarf getaggt werden
  - <biblStruct> = stärker durchstrukturierte Variante, interessant für systematische Verarbeitung
  - <biblFull> = Variante analog zum teiHeader mit <fileDesc> etc. etc.

<msDesc>

- Beschreibung handschriftlicher Quellen
- Gliederung:
  - <msIdentifier> Standort, Signatur etc.
  - <msContents> inhaltliche Beschreibung
  - <physDesc> materielle Beschreibung
  - <history> genetische Beschreibung
  - <additional> sonstiges

Fazit

- der Header ist sehr umfangreich
- viele Elemente sind sehr speziell
  - vieles wird man vielleicht nie brauchen
- mehrere Beschreibungsmethoden werden unterstützt
  - prosaischer Ansatz
  - vollständig durchstrukturierter Ansatz
- nutzbar als eigenes Referenzsystem

Wann kümmern?

- vermutlich mindestens zweimal:
  - bei der Erstellung eines Dokuments
  - beim Abschluss eines Dokuments
  - ggf. auch während der Bearbeitung eines Dokuments
- Metadatenerfassung als Teil des Erschließungsprozesses

Übung

- Verwenden Sie eine Datei, die Sie bereits bearbeitet haben und gut kennen.
- Prüfen Sie die obligatorischen Informationen im Header. Ist alles zufriedenstellend beschrieben? Können Sie die vorhandenen Informationen noch tiefer strukturieren?
- Nehmen Sie sich nun ein Metadaten-Element vor, welches noch nicht vorhanden ist, beispielsweise <encodingDesc>.
- Beginnen Sie mit der Prosa-Methode: Nutzen Sie zum Ausfüllen zunächst nur <p> und schreiben Sie hinein, was der TEI-Dokumentation zufolge dort passen könnte.
- Fahren Sie fort, indem Sie die <p> auflösen und stattdessen die spezifischen Header-Elemente nutzen. Vertiefen Sie den Detailgrad mithilfe der TEI-Dokumentation weitestmöglich. An welchen Stellen treten Schwierigkeiten auf?
- Versuchen Sie für Ihr Vorhaben oder für Ihre Herangehensweise eine pragmatische Grenze zwischen prosaischen und strukturierten Metadaten-Angaben zu finden. An welchen Stellen hilft Ihnen die Strukturierung?
- Vergleichen und erläutern Sie untereinander Ihre Ansätze.

Weiterführende Literatur:

https://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD11

Fragen?

torsten.roeder@leopoldina.org

Forschungsstelle
für die Geschichte
der Hanse und des Ostseeraums

i|d|e

Christian-Albrechts-Universität zu Kiel
Historisches Seminar