Datenmodellierung

Woche 6: Textmodellierung

Institut für Digital Humanities, Historisch-Kulturwissenschaftliche Informationsverarbeitung | Prof. Dr. Øyvind Eide

Was ist Textmodellierung?

Textstrukturen und Inhalte:

- Unterschiedliche Strukturen und Ebenen
  - Inhalt vs. Struktur
- Inhaltliche Aspekte können sehr komplex sein
  - Erzählebenen
  - Metapher
  - historische Informationen
  - topografische Informationen
  - ...

Was ist Textmodellierung?

Beispiele modellierbarer Elemente:

- Brieftexte
  - Absendedatum
  - Absendeort
  - Sender
  - Empfänger
  - Grußformeln
  - erwähnte Personen
  - Orte
- Gedichte
  - Strophen
  - Verse
- Tagebuchtexte
  - Tagebuch
  - Einträge mit Datumsangaben
  - erwähnte Entitäten
    - Personen
    - Orte
    - Institutionen
- Linguistische Forschung
  - Tagging einzelner Wörter und Wortteile

Was ist Textmodellierung?

Zweck der Textmodellierung:

- Identifikation und explizite Darstellung impliziter Strukturen
- Explizierung für die Weiterverarbeitung
  - Analyse
  - Visualisierung
  - Auswertung
- Zweck und Forschungsziel im Fokus behalten
  - Balance zwischen Detailtiefe und Forschungsrelevanz

OK, aber wissen wir eigentlich, was Text ist?

Nach dem Duden...

„schriftlich fixierte im Wortlaut festgelegte, inhaltlich zusammenhängende Folge von Aussagen"

???

Einige Zweifel

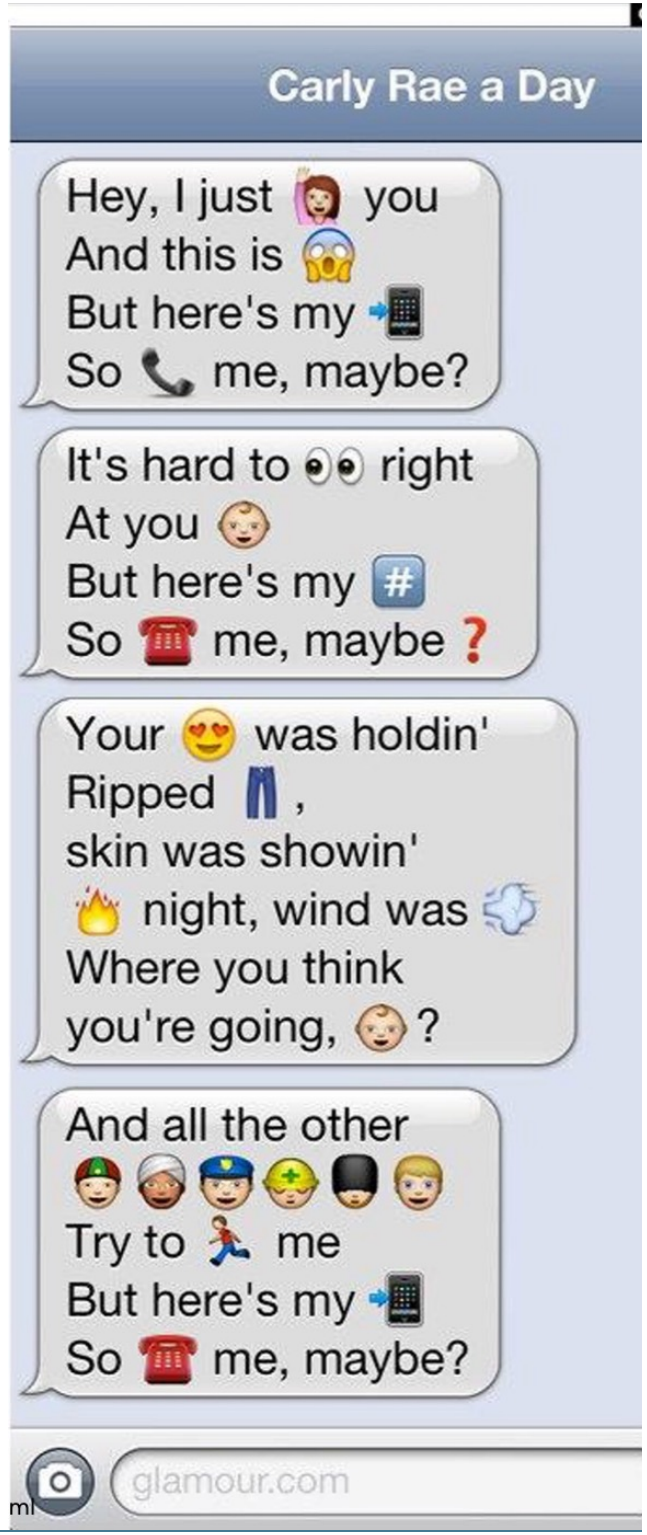
- schriftlich?
- fixiert?
  - ehm... https://taper.badquar.to/5/a_storm_in_2k.html
- im Wortlaut? Was ist mit den Smileys?
- inhaltlich zusammenhängend?
  - Ist „Der Computer trank den Sonnenuntergang, während die Melodie Bücher schrieb” kein Text?
- Folge von Aussagen? Ist das Gedicht von Strickland Gillilan (1869-1954) kein Text?

Titel: Lines on the Antiquity of Microbes (“Fleas”)

Adam

Had ‘em

- Sind „Aussagen” nötig? Ist eine Einkaufsliste kein Text?



Eine bessere Definition

Text als **Repräsentation** von **Information** in einem **semiotischen System**

Semiotik → Wissenschaft von den Zeichenprozessen in Kultur und Natur. Zeichen, wie zum Beispiel Bilder, Wörter, Geste und Gerüche vermitteln Informationen aller Art in Zeit und Raum. In Zeichenprozessen (Semiosen) werden Zeichen konstituiert, produziert, in Umlauf gebracht und rezipiert.

Ein Spiel: Wer bin ich?

- Gruppen von 3 Personen bilden
- Jede Gruppe zieht 2 Zettel von der Textartenbox
- In den Gruppen werden für jede gewählte Textart 8 Kategorien gelistet, die diese Textart charakterisieren

→ 10–15 Minuten

- Wenn die Gruppen fertig sind, nennt jede Gruppe jeweils zwei oder drei der gelisteten Kategorien. Die anderen Gruppen müssen raten, um welche Textart es geht!

- → 10–15 Minuten

Die Text Encoding Initiative

- Die TEI ist ein Konsortium, das das gleichnamige Format für die Kodierung von Texten entwickelt
- Die Auszeichnungssprache TEI basiert auf XML und ist heute als de-facto Standard zu betrachten
- In TEI XML kann man strukturelle und semantische Einheiten kodieren

https://tei-c.org/

TEI: Lexikon

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 Computer-mediated Communication
- ⊞ 10 Dictionaries
- ⊞ 11 Manuscript Description
- ⊞ 12 Representation of Primary Sources
- ⊞ 13 Critical Apparatus
- ⊞ 14 Names, Dates, People, and Places
- ⊞ 15 Tables, Formulæ, Graphics, and Notated Music
- ⊞ 16 Language Corpora
- ⊞ 17 Linking, Segmentation, and Alignment
- ⊞ 18 Simple Analytic Mechanisms
- ⊞ 19 Feature Structures
- ⊞ 20 Graphs, Networks, and Trees
- ⊞ 21 Non-hierarchical Structures
- ⊞ 22 Certainty, Precision, and Responsibility
- ⊞ 23 Documentation Elements
- ⊞ 24 Using the TEI

TEI sourcecode

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

[English] [Deutsch] [Español] [Italiano] [Français] [日本語] [한국어] [中文]

Beispiel: “performance texts” https://tei-c.org/release/doc/tei-p5-doc/en/html/DR.html

Textmodell

TEI ist auf dem OHCO Modell, nachdem Text eine „Ordered Hierarchy of Content Objects" ist, basiert.

DeRose, Steven J., David G. Durand, Elli Mylonas, et Allen H. Renear. 1990. «What is text, really?» Journal of Computing in Higher Education 1 (2): 3-26. https://doi.org/10.1007/BF02941632

ABSTRACT

THE WAY IN WHICH TEXT IS represented on a computer affects the kinds of uses to which it can be put by its creator and by subsequent users. The electronic document model currently in use is impoverished and restrictive. The authors argue that text is best represented as an ordered hierarchy of content object (OHCO), because that is what text really is. This model conforms with emerging standards such as SGML and contains within it advantages for the writer, publisher, and researcher. The authors then describe how the hierarchical model can allow future use and reuse of the document as a database, hypertext, or network.

Syntax: XML

- TEI basiert auf der Auszeichnungssprache XML

- Geschichte Auszeichnungssprachen:

ML → IBM GML (1969) → SGML (ISO 8879:1986) → XML (W3C 1998)

- Geschichte TEI:

P1 (SGML 1990), P3 (SGML 1994), P4 (XML 2002), P5 (XML 2007)

Syntax: XML

- TEI basiert auf der Auszeichnungssprache XML

Was ist XML?

- XML steht für eXtensible Markup Language
- Es ist eine Markup-Sprache, die zum Speichern und Transportieren von Daten verwendet wird
- XML ist sowohl für Menschen als auch für Maschinen lesbar und verständlich
  - obwohl unterschiedlich verständlich
- XML ist flexibel, erweiterbar und tool-unabhängig
- XML beruht auf einer hierarchischen Baumstruktur

Elemente

- Grundbausteine einer XML-Datei, die Daten enthalten und strukturieren
- Elemente werden durch Tags markiert.

  Start-Tag: <elementName>

  End-Tag: </elementName>
- Inhalt: Der Text oder andere Elemente, die zwischen den Tags stehen.

Universität zu Köln

Attribute

- Zusätzliche Informationen oder Eigenschaften eines Elements
- Attribute werden im Start-Tag eines Elements definiert
- Format:

  <elementName attributName="attributWert">INHALT</elementName>

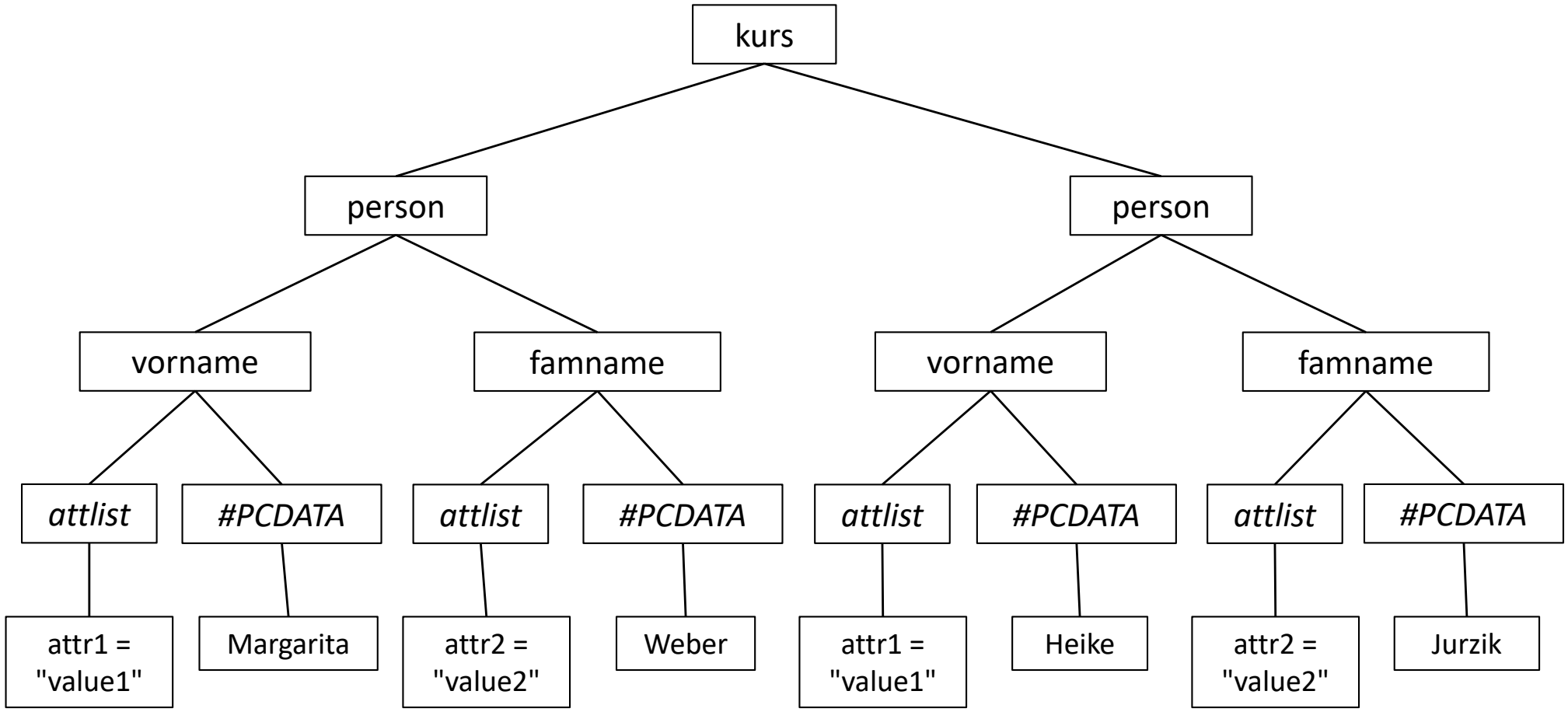
- Einsatz: Attribute sollten verwendet werden, um Metadaten oder zusätzliche Informationen über ein Element bereitzustellen, die keine komplexen Strukturen erfordern

Beispiel: XML als Zeichenkette (Fragment)

```
<kurs>
  <person>
    <vorname attr1="value1">Margarita</vorname>
    <famname attr2="value2">Weber</famname>
  </person>
  <person>
    <vorname attr1="value1">Heike</vorname>
    <famname attr2="value2">Jurzik</famnme>
  </person>
</kurs>
```


Universität zu Köln

…und als Baum



Universität zu Köln

Regeln

- Alle **Elemente** müssen geöffnet und geschlossen werden

```
<p>Some text here.</p>
```

- **Attribute** werden nur einmal definiert

```
<p n="1">Some text here.</p>
```

- Overlapping ist nicht erlaubt

```
*<l><placeName>Vened</l>ig</placeName>
```

- Self-contained elements lösen viele Probleme:

```
<lb/><placeName>Vened <lb/>ig</placeName> ist schön
```

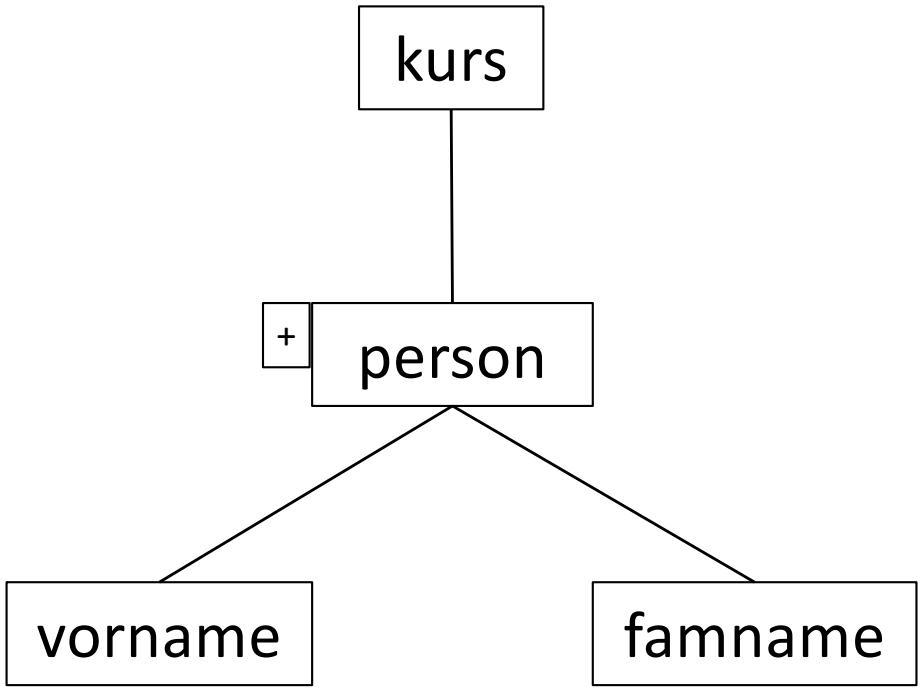
- Die Struktur ist hierarchisch!

Universität zu Köln

Regeln

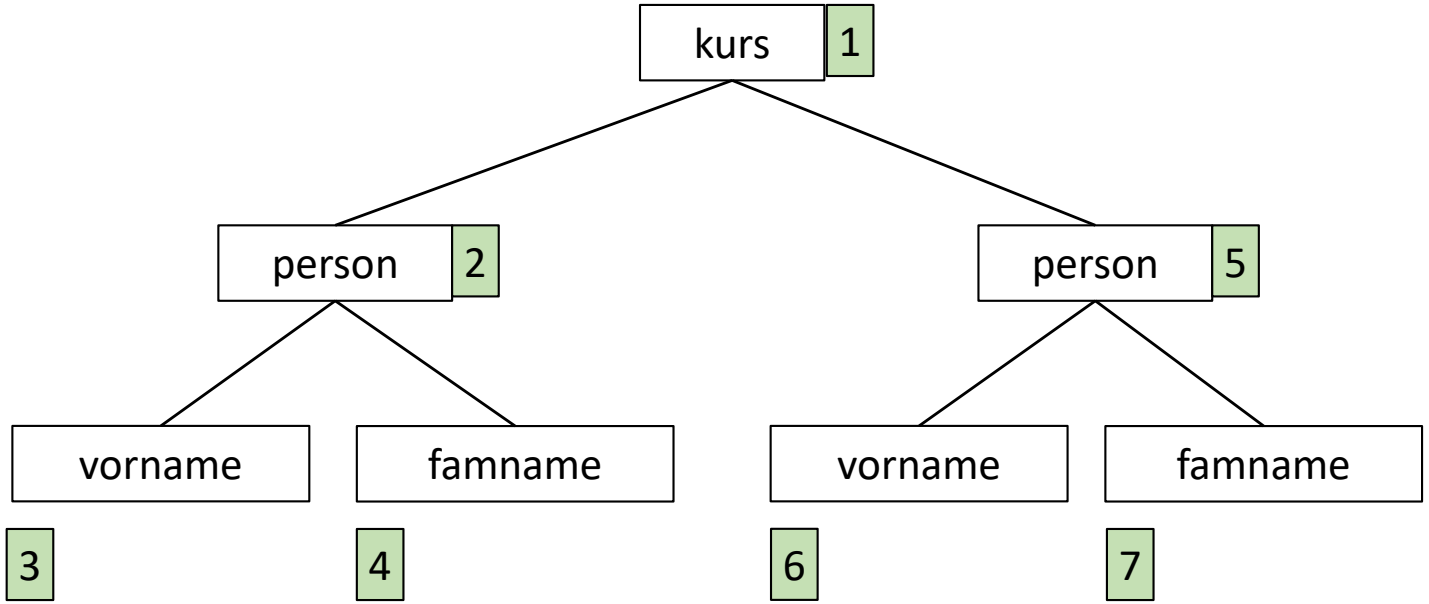
- Wohlgeformt
- Validiert
  - DTD (Document Type Definition)
  - Schema
  - RNG (RelaxNG)

DTD als Baum

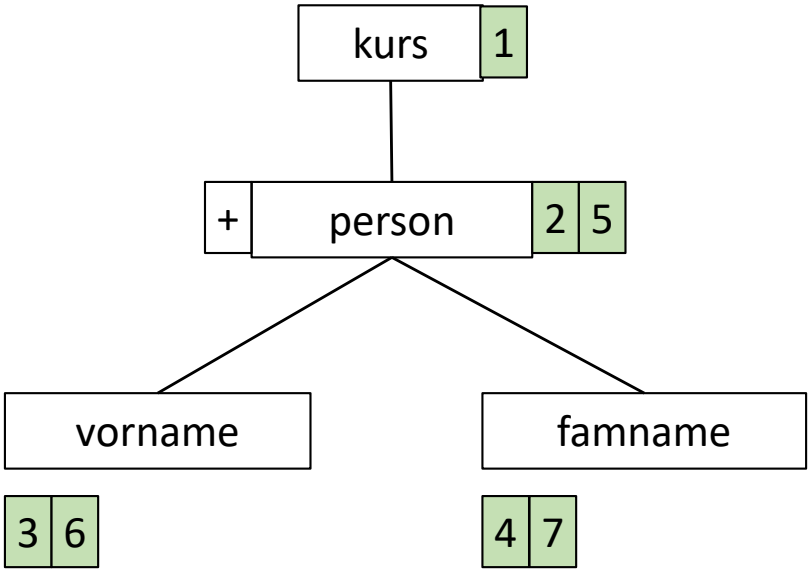


Universität zu Köln

Validierende Baumtraversierung



XML



DTD

Hausaufgabe 1: Oxygen XML Editor installieren

- Hier Demo Version downloaden:

Mac:
https://www.oxygenxml.com/xml_editor/download_oxygenxml_editor.html?os=macOS

Windows:
https://www.oxygenxml.com/xml_editor/download_oxygenxml_editor.html?os=Windows

Linux:
https://www.oxygenxml.com/xml_editor/download_oxygenxml_editor.html?os=Linux

- Musterdatei buchdernatur.xml (Ilias) in Oxygen öffnen

https://www.ilias.uni-koeln.de/ilias/goto.php/crs/6743345

Hausaufgaben 2

Die Kategorien aus der Gruppenaktivität über Textarten auf Ilias einreichen

Hier: https://www.ilias.uni-koeln.de/ilias/goto.php/exc/6898124

Universität zu Köln

Hausaufgaben 2: Vorbereitung auf Diskussion

- Was sind Aspekte von Text, die nicht hierarchisch sind?
- In welchen Bereichen könnte man Overlap von Hierarchien haben?

Universität zu Köln