Textkodierung und Digitale Edition: Teil 1

Vorlesung *Einführung in die Digital Humanities*

Prof. Dr. Christof Schöch
Wintersemester 2020/21

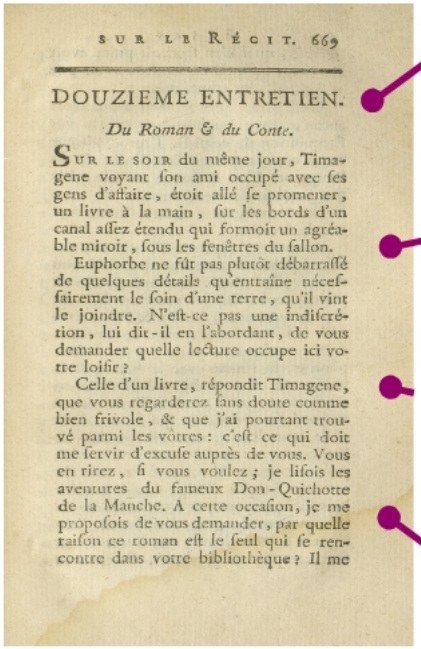
Universität Trier

Sitzung im Überblick

1. Was ist Text?
2. Textauszeichnung
3. Textkodierung mit XML
4. Praxis der Textkodierung

1. Was ist Text?
(= konzeptuelles Datenmodell)

Aspekte des digitalen Texts


SUR LE RÉCIT. 669

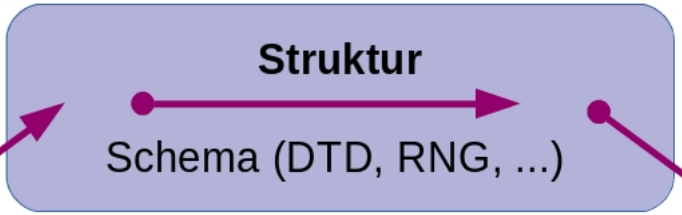
DOUZIEME ENTRETIEN.

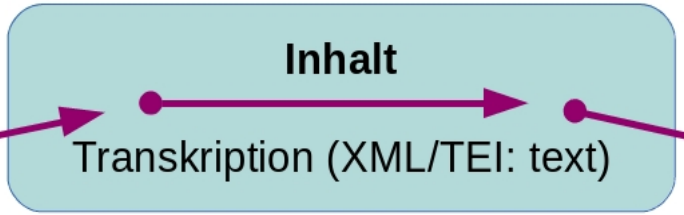
Du Roman & du Conte.

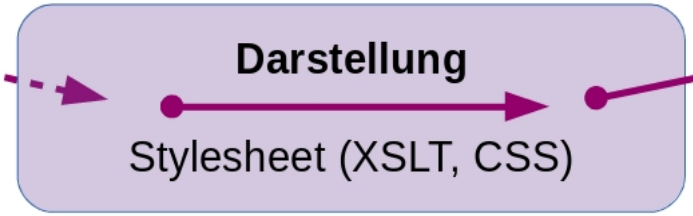
SUR LE SOIR du même jour, Timagene voyant ſon ami occupé avec ſes gens d'affaire, étoit allé ſe promener, un livre à la main, ſur les bords d'un canal aſſez étendu qui formoit un agréable miroir, ſous les fenêtres du ſallon.

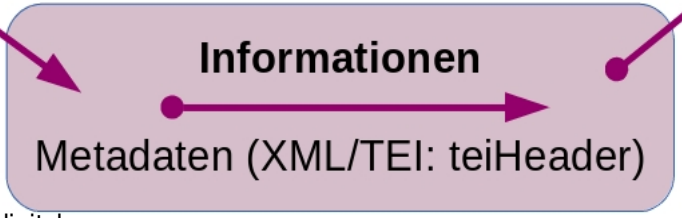
Euphorbe ne fût pas plutôt débarraſſé de quelques détails qu'entraîne néceſſairement le ſoin d'une terre, qu'il vint le joindre. N'eſt-ce pas une indiſcrétion, lui dit-il en l'abordant, de vous demander quelle lecture occupe ici votre loiſir?

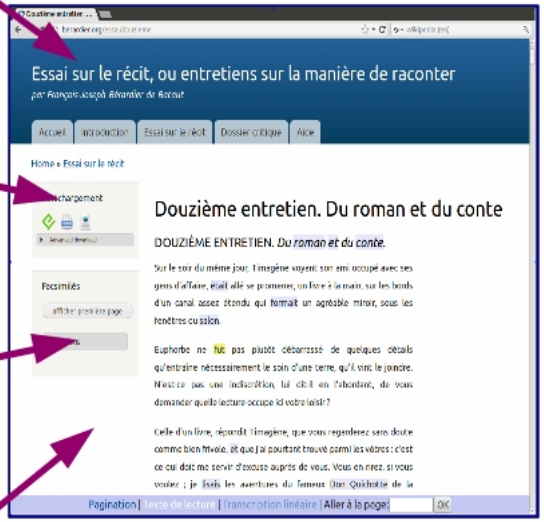
Celle d'un livre, répondit Timagene, que vous regarderez ſans doute comme bien frivole, & que j'ai pourtant trouvé parmi les vôtres: c'eſt ce qui doit me ſervir d'excuſe auprès de vous. Vous en rirez, ſi vous voulez; je liſois les aventures du fameux Don-Quichotte de la Manche. A cette occaſion, je me propoſois de vous demander, par quelle raiſon ce roman eſt le ſeul qui ſe rencontre dans votre bibliothèque? Il me











2. Textauszeichnung (Markup)

Was ist Markup?

- Ursprung im Druckwesen: Anweisungen für den Setzer
- “markup”
  - (Subst.) = Auszeichnung, Markierung
  - (Verb) = auszeichnen, markieren, kodieren
- Macht Eigenschaften explizit
  - benennt und/oder charakterisiert Teile der Zeichenkette
  - auf formalisierte und kohärente Weise

Zwei Typen von Markup

- "**prozedural**": visuell, typographisch:
  - Anweisung, wie ein Stück Text dargestellt werden soll
  - Schwerpunkt auf Darstellung, Aussehen
  - häufig mehrdeutig, unflexibel

- "**deskriptiv**": semantisch, funktional, strukturell
  - Explizieren, welche Funktion ein Stück Text hat
  - Schwerpunkt auf Struktur und Bedeutung
  - separat davon die Darstellung definieren
  - mehr, eindeutigere Information
  - Darstellung leichter anpassbar

Prozeduraler Markup

Woody Allens Urteil
Krieg und Frieden ist ein fantastisches, aber *viel* zu langes Buch.

"prozedural"

```
<markup>
  <fett>Woody Allens Urteil</fett><umbruch/><kursiv>Krieg und
Frieden</kursiv> ist ein fantastisches, aber
<kursiv>viel</kursiv> zu langes Buch.
</markup>
```

Deskriptiver Markup

Woody Allens Urteil

Krieg und Frieden ist ein fantastisches, aber *viel* zu langes Buch.

"deskriptiv"

```
<markup>
  <titel>Woody Allens Urteil</titel><absatz> <buch>Krieg und
Frieden</buch> ist ein fantastisches, aber <emph>viel</emph> zu
langes Buch.</absatz>
</markup>
```

Stylesheet

+

```
<stylesheet>
  titel: fett + Umbruch; buch: kursiv;
  absatz: normal, Umbruch; emph: kursiv;
</stylesheet>
```

3. Textkodierung mit XML

Was ist XML?

- eXtensible Markup Language
- W3C-Standard

- Metasprache zur Definition von XML-Formaten
- Standard für digitale Repräsentation von Daten
- Prinzipien + Syntax

- einfach (wenige, mächtige Mechanismen)
- anwendungs- und plattformunabhängig

X-Technologien im Überblick

- Technologien zur Definition von XML-Formaten
  - DTD (Document Type Definition)
  - XSD (XML Schema Declaration)
  - Relax NG
- Verwandte Technologien für Transformation und Analyse
  - XSL/XSLT
  - XPath (=> wichtig für Suche mit lxml!)
- Mit XML definierte spezielle Markup Languages
  - xHTML
  - MathML, MusicXML
  - TEI, MEI , CEI, ...

XML: Elemente, Attribute, Werte, Strings

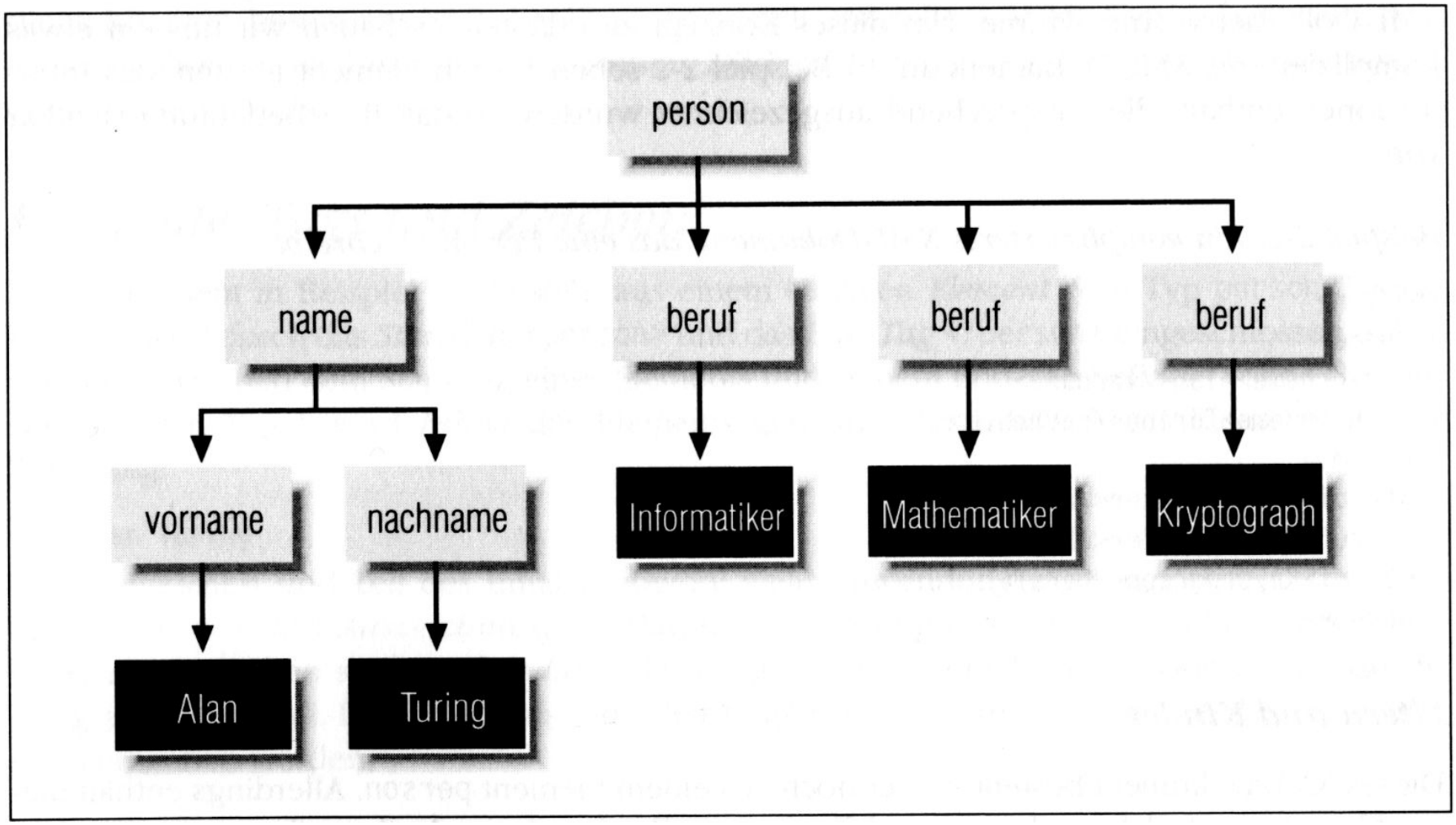
XML-Elemente-Attribute-Werte.xml

```xml
<?xml version="1.0" encoding="UTF-8"?>
<Krimi-Sammlung>
    <Krimi n="1">
        <titel>Piège pour Cendrillon</titel>
        <autor status="bekannt">
            <name>Japrisot</name>
            <vorname>Sébastien</vorname>
        </autor>
    </Krimi>
    <Krimi n="2">
        <titel>Meurtres pour mémoire</titel>
        <autor status="berühmt">
            <name>Daeninckxs</name>
            <vorname>Didier</vorname>
        </autor>
    </Krimi>
</Krimi-Sammlung>
```

Kodierung in XML

```
<person>
    <name>
      <vorname>Alan</vorname>
      <nachname>Turing</nachname>
    </name>
    <beruf>Informatiker</beruf>
    <beruf>Mathematiker</beruf>
    <beruf>Kryptograph</beruf>
</person>
```

Baumstruktur als Diagramm



Dokumentstruktur

Festgelegt in einem Schema, nach verschiedenen Standards:

- Document Type Definition
  - http://www.w3.org/TR/REC-xml/#dt-doctype
- Relax NG
  - http://standards.iso.org/ittf/PubliclyAvailableStandards/c052348_ISO_IEC_19757-2_2008(E).zip
- W3C Schema
  - http://www.w3.org/standards/techs/xmlschema

Beispiel Gedicht

Gaiku 123

early dew
the water contains
teaspoons of honey

(2009)

(Quelle: Yael Netzer et al., „Gaiku: Generating Haiku with word association norms, 2009)

Konzeptionelles Datenmodell

- Titel
- Gedicht
- Zeilen
- Jahr

Relax NG-Schema für „Haiku-Sammlung“

```
start = haikus
haikus = element haiku+
haiku = element haiku {head?, body}
head = element head {title?, year?}
title = element title {text}
year = element year {xsd:integer}
body = element body {stanza}
stanza = element stanza {line+}
line = element line {text}
```

Nach Schema kodiertes Gaiku

haikus.xml

```xml
<haikus>
  <haiku>
    <head>
      <title>Gaiku 123</title>
      <year>2009</year>
    </head>
    <body>
      <stanza>
        <line>early dew</line>
        <line>the water contains</line>
        <line>teaspoons of honey</line>
      </stanza>
    </body>
  </haiku>
</haikus>
```

well-formed vs. valid

- „well-formed“ (wohlgeformt)
  - Dokument entspricht den allgemeinen Prinzipien von XML
  - die Kriterien sind immer gleich
  - die Kriterien sind allgemein
- „valid“ (valide)
  - Dokument entspricht der Syntax und dem Lexikon eines spezifischen XML-Formats
  - Kriterien hängen von der jeweiligen Definition (DTD, Schema) ab
  - Kriterien sind meist sehr detailliert

Kriterien für "Wohlgeformtheit"

- Prolog: XML-Version, Zeichensatz
- Nur ein Element auf oberster Ebene

- Jedes Element hat Anfangs- und Endtag
- Hierarchische Struktur: keine überlappenden Elemente
- Elemente können Unterelemente haben
- Elemente können Attribute haben
- Attribute können Werte haben
- Die Werte sind in Anführungszeichen gesetzt
- Alle Zeichen entsprechen dem ang. Zeichensatz

Kriterien für "Validität"

- wohlgeformt
- Definition (Schema/DTD) vorhanden: intern/extern
- Dokument entspricht der Definition

- Alle notwenigen, nur erlaubte Elemente
- Alle notwendigen, nur erlaubte Attribute
- Alle Werte haben eine gültige Form/Ausprägung
- Elemente und Attribute kommen nur dort vor, wo sie auch erlaubt sind

Literaturhinweise

Lektürehinweise

Referenzlektüre

- Georg Vogeler und Patrick Sahle: „XML“, in: Digital Humanities: Eine Einführung, hg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, 2017, 128-146.

Weitere Empfehlungen

- Vonhoegen, Helmut. *Handbuch: Einstieg in XML. Grundlagen, Praxis, Referenz*. Bonn, 2015.
- Allen Renear: „Text Encoding“. In: Susan Schreibman et. al., eds.: *Companion to Digital Humanities*. 2006.
- Christof Schöch: „Ein digitales Textformat für die Literaturwissenschaft. Die Richtlinien der Text Encoding Initiative und ihr Nutzen für Textedition und Textanalyse“, *Romanische Studien* 4 (2016), S. 325–364, URL: http://www.romanischestudien.de/index.php/rst/article/view/58.
- Melissa Terras, Edward Vanhoutte, Ron Van den Branden, hg.: *TEI by Example*. URL: http://teibyexample.org/ (interaktives Tutorial)