

Semantische Datenintegration

Integration von Dokumentformaten

Jakob Voß

2017-04-01

Dokumente

Was ist ein Dokument?

- ▶ Kernbegriff der Dokumentationswissenschaft
- ▶ Nicht einfach zu beantworten (Buckland 1997)
- ▶ Definition von Briet (1951)

A document is evidence in support of a fact [...] any physical or symbolic sign, preserved or recorded, intended to represent, to reconstruct, or to demonstrate a physical or conceptual phenomenon

Beispiele für Dokumente?

- ▶ ...
- ▶ ...
- ▶ ...
- ▶ ...

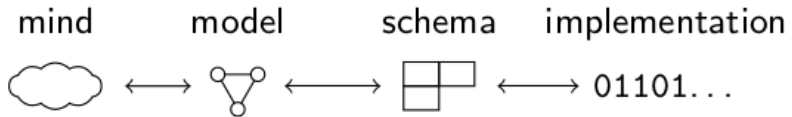
Beispiele für Dokumente

- ▶ Archäologische Objekte
- ▶ Spuren
- ▶ Publikationen
- ▶ Kunstwerke (sic)

Daten als digitale Dokumente

- ▶ Auffassungen von Daten
 - ▶ Daten als Fakten (\Rightarrow Fachgebiete)
 - ▶ Daten als Beobachtungen (\Rightarrow Statistik)
 - ▶ Daten als Dokumente (\Rightarrow Informationsmanagement)
- ▶ Semiotischer Ansatz: Daten/Dokumente als Zeichen für etwas
- ▶ Dokumente haben
 - ▶ Ursprung, Zwecke, Adressaten...
 - ▶ Form (*hier Thema*)

Dokumentformate



Wo finden wir Daten-/Dokumentenformate?

Arten von Datensprachen

- ▶ Modellierungs-Sprachen (UML, ERM...)
- ▶ **Schema-Sprachen** (RDF Schema, XML Schema, RegExp...)
 - ▶ Abfragesprachen (SQL, XPath...)
- ▶ **Datenstrukturierungssprachen** (CSV, XML, JSON...)
 - ▶ Auszeichnungssprachen (HTML, TEI, Markdown...)
- ▶ Kodierungen (Unicode, ASCII, Binärcode...)

Beschreibung von Datenformaten

- ▶ Spezifikationen/Standards
- ▶ Schemata
 - ▶ z.B. ein XML-Schema für ein XML-Format
- ▶ Implementierungen (!)
- ▶ Datenstrukturierungssprache als Grundlage
 - ▶ z.B. XML für ein XML-Format

Dateiformate

- ▶ Beliebige Datenformate für Daten in Dateien
- ▶ Oft Containerformate (MPEG, PDF...)
- ▶ Oft weitere Dateien/Datenbanken als Grundlage (z.B. ODT)
- ▶ *Ist "Datei" nicht zunehmend eine Methapher?*

Exkurs: MIME-Type

- ▶ 1996: Multipurpose Internet Mail Extensions (MIME, 1996)
- ▶ content type / media type
- ▶ Verwendet u.A. in HTTP-Nachrichten
- ▶ Identifier für Dateiformate

⇒ *IANA list of official media types*

Übung: Anatomie einer OpenDocument-Datei



Medienformate

aka *Content-Formate*

- ▶ Formate und Kodierungen für “Inhalte”
- ▶ Audio-, Video-, Bild-Formate
 - ▶ Eigenes Thema, mehr Mathematik
- ▶ Prinzipiell alle Arten digitaler Dokumente

Dokumentformate im engeren Sinne

- ▶ Seitenbeschreibungssprachen
 - ▶ PDF, PostScript, DVI...
- ▶ Dokumentformate für (Text-)Dokumente
 - ▶ OpenDocument Format
 - ▶ Plain Text
 - ▶ HTML
 - ▶ LaTeX
 - ▶ TEI
 - ▶ DocBook
 - ▶ EPUB
 - ▶ ...
- ▶ Basieren meist auf Auszeichnungssprachen

Auszeichnungssprachen

Was sind Auszeichnungssprachen?

- ▶ Modellierungs-Sprachen (UML, ERM...)
- ▶ Schema-Sprachen (RDF Schema, XML Schema, RegExp...)
 - ▶ Abfragesprachen (SQL, XPath...)
- ▶ Datenstrukturierungssprachen (CSV, XML, JSON...)
 - ▶ **Auszeichnungssprachen** (HTML, TEI, Markdown...)
- ▶ Kodierungen (Unicode, ASCII, Binärcode...)

Beispiel: HTML

- ▶ Basiert (theoretisch) auf SGML/XML
- ▶ Spezifikation durch W3C
- ▶ Deskriptive Elemente (<title>, <h1>, , <code>...)
- ▶ Präsentative Elemente (<i>, <tt>...)
- ▶ Legt zusammen mit CSS allgemeines Erscheinungsbild einer Webseite (oder EPUB, UI...) fest ("HTML-Design")

Beispiel: TEI

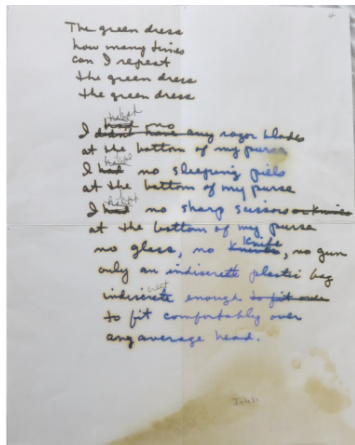
- ▶ Text-Encoding Initiative (TEI) seit 1987 (!)
- ▶ Basierte auf SGML, inzwischen XML
- ▶ Verbreitet für Texteditionen in den Digital Humanities
- ▶ Umfangreiches Regelwerk der TEI
 - ▶ Mehr als 500 Elemente
 - ▶ Mehrere Module und Schemata
- ▶ Beschreibt welche Inhalte in einem Dokument vorkommen (deskriptiv)

Beispiel für TEI

Beispiel: <surface> und <zone>

Welche Oberfläche, welche Zonen/Bereiche könnte man hier unterscheiden?

- Surface: die gesamte Seite
- Zone 1: Nummer oben rechts
- Zone 2: Textblock oben
- Zone 3: Textblock unten
- Zone 4: Titel unten rechts
- evtl. Zone 5: Verfärbung (Polygon)



CC-BY DARIA-DE "Digitale Textedition mit TEI"

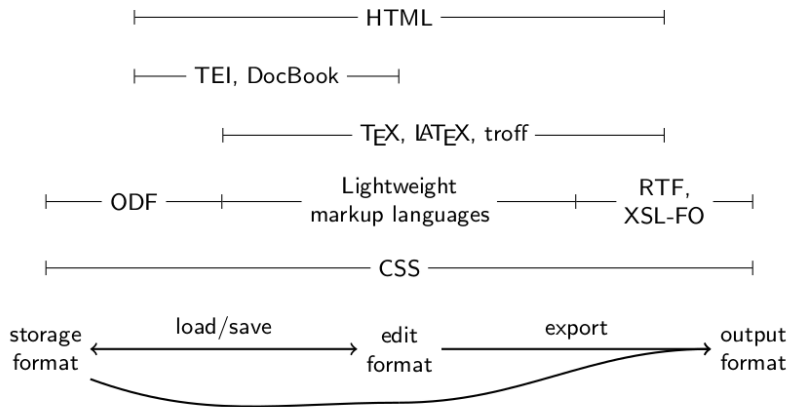
Beispiel: Markdown

- ▶ Einfache Bearbeitung von Texten
- ▶ Beschränkung auf typische Elemente
 - ▶ Überschriften
 - ▶ Absätze, Listen, Tabellen, Bilder,
 - ▶ Fett, Kursiv
 - ▶ ...

Übung: Markdown-Bearbeitung mit HackMD

`https://hackmd.io/`

Anwendungsschwerpunkte von Auszeichnungssprachen



Vergleich von Auszeichnungssprachen

language	bold face	italic face	monospace	sub-/superscript
DocBook	<code><emphasis role='strong'></code>	<code><emphasis></code> , <code><firstterm></code>	<code><code></code> , <code><varname></code>	<code><subscript></code> , <code><superscript></code>
HTML	<code></code>	<code><i></code>	<code><tt></code> , <code><code></code>	<code><sub></code> , <code><sup></code>
TEI	<code><hi rend="bold"></code>	<code>rend="italics"</code>	<code>rend="typewriter"</code>	<code>rend="subscript"</code> , <code>rend="superscript"</code>
T _E X	<code>\textbf {text}</code>	<code>\textit {text}</code> , <code>\emph {text}</code>	<code>\texttt {text}</code> , <code>\verb #text#</code>	<code>~{text}</code> , <code>_ {text}</code>
RTF	<code>{\b text}</code>	<code>{\i text}</code>	<code>{\fmodern text}</code>	<code>{\sub text}</code> , <code>{\sup text}</code>
MediaWiki	<code>'''text'''</code>	<code>''text''</code>	<code><tt></code> , <code><code></code>	<code><sub></code> , <code><sup></code>
Markdown	<code>**text**</code> , <code>__text__</code>	<code>*text*</code> , <code>_text_</code>	<code>‘text’</code>	<code><sub></code> , <code><sup></code>
Textile	<code>*text*</code>	<code>_text_</code>	<code>@text@</code>	<code>~text~</code> , <code>^text^</code>
reStructuredText	<code>**text**</code>	<code>*text*</code>	<code>‘text’</code>	<code>:sub: ‘text’</code> , <code>:sup: ‘text’</code>
POD	<code>B<text></code>	<code>I<text></code>	<code>C<text></code>	–
GNU troff (man)	<code>.fam B text .fam,</code> <code>\fBtext \fP</code>	<code>.fam I text .fam,</code> <code>\fItext \fP</code>	<code>.fam C text .fam,</code> <code>\fCtext \fP</code>	<code>~text~</code> , <code>^text^</code>

Konvertierung zwischen Dokumentformaten

Eingangsformat → **Dokumentmodell** → Ausgangsformat

- ▶ Datenmodell von Dokumentstrukturen
- ▶ Tool: Pandoc
 - ▶ <https://pandoc.org/>
 - ▶ <https://cloudconvert.com/>
 - ▶ <https://foliovision.com/seo-tools/pandoc-online>
 - ▶ ...

Pandoc-Filter mit Pandoc::Elements (Perl)

```
use Pandoc::Filter qw(pandoc_filter);
use Pandoc::Elements qw(BulletList Para Strong Str);

pandoc_filter DefinitionList => sub {
    BulletList [ map { to_bullet($_) } @{$_->items} ]
};

sub to_bullet {
    my $item = shift;
    [ Para [ Strong $item->term ], map { @$_ } @{$item->definitions} ]
}
```

Pandoc-Filter mit panflute (Python)

```
from panflute import toJSONFilter, DefinitionList, BulletList, ListItem

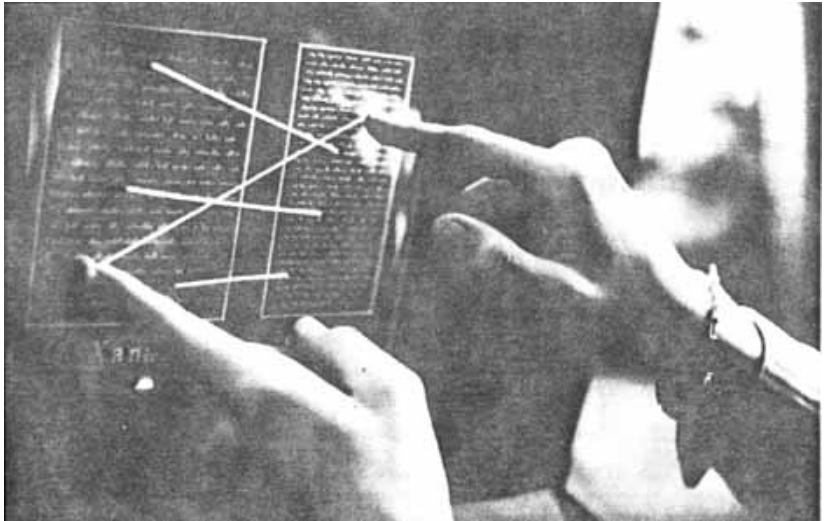
def deflists(elem, doc):
    if type(elem) == DefinitionList:
        bullets = [tobullet(item) for item in elem.content]
        return BulletList(*bullets)

def tobullet(item):
    ans = [Para(Strong(*item.term))]
    for definition in item.definitions:
        for block in definition.content:
            ans.append(block)
    return ListItem(*ans)

if __name__ == "__main__":
    toJSONFilter(deflists)
```

Zusammenfassung und Ausblick

Hypertext



Ted Nelson 1972

Beyond the PDF

- ▶ Versionierung
- ▶ Annotationen
- ▶ Dynamische Dokumente
- ▶ Eingebettete Daten
- ▶ ...

Beispiele:

- ▶ Jupyter/iPython Notebook
- ▶ Hypothes.is

Zusammenfassung

- ▶ Digitale Dokumente sind Datenobjekte, die als Dokument wahrgenommen werden
- ▶ Datenformat
 - ▶ Dateiformat
 - ▶ Medienformat
 - ▶ **Dokumentformat**
- ▶ Auszeichnungssprachen
 - ▶ gibt es viele
 - ▶ basieren auf einem Dokumentmodell
 - ▶ haben kleinsten gemeinsamen Nenner

Briet, Suzanne. 1951. *Qu'est-ce que la documentation?* Paris: Éditions documentaires, industrielles et techniques.

Buckland, Michael. 1997. „What is a ‚document‘?“ *Journal of the American Society of Information Science (JASIST)* 48 (9): 804–9.

Nelson, Theodor Holm. 1999. „Xanalogical Structure, Needed Now More Than Ever: Parallel Documents, Deep Links to Content, Deep Versioning, and Deep Re-Use“. *ACM Computing Surveys* 31 (4es). doi:10.1145/345966.346033.