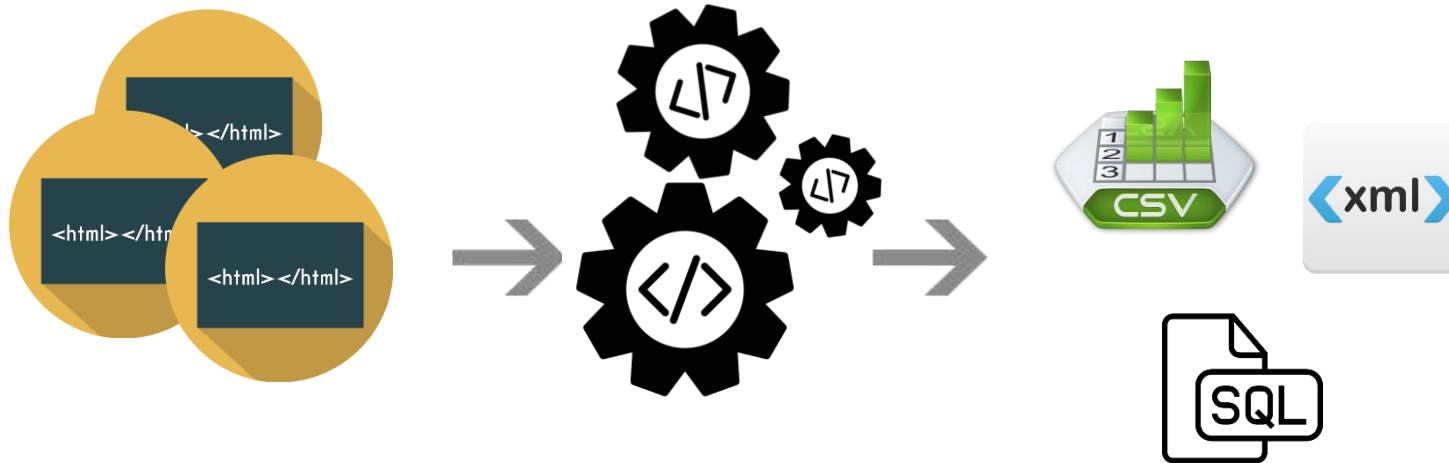


Einführung in Web Scraping mit dem Chrome-Plugin *Scraper*



Agnes Brauer
a.brauer@ub.uni-frankfurt.de

Vorbereitung

- Melden Sie sich für den moodle-Kurs [Praxislabor Digitale Geisteswissenschaften](#) an und schreiben Sie sich ein:

Praxislabor Digitale Geisteswissenschaften: Einführungsworkshops zu Methoden und Werkzeugen der Digital Humanities

Praxislabor Digitale Geisteswissenschaften:
Einführungsworkshops zu Methoden und Werkzeugen der Digital Humanities

Zum ersten Kennenlernen von Methoden und Werkzeugen der Digital Humanities bietet die Universitätsbibliothek JCS (im Bibliothekszentrum Geisteswissenschaften) Studierenden und Mitarbeiterinnen der Goethe-Uni im kommenden Wintersemester Workshops an. In niederschweligen Einführungen werden anhand von überschaubaren, konkreten Beispielen aus der Praxis Methoden, Tools oder Themen der digitalen Geisteswissenschaften vorgestellt und geübt und so ein erster Einblick in die Möglichkeiten gegeben, wie klassische Methoden der Geisteswissenschaften mithilfe digitaler Verfahren der Textanalyse sowie der Text- und Datenaufbereitung sinnvoll ergänzt werden können.

Die Workshopreihe besteht jeweils aus inhaltlich zusammenhängenden Zweierblöcken, in denen auf eine Präsentation eine Sitzung zur Vertiefung und Übung folgt.

Die Workshops richten sich an interessierte Einsteiger; besondere Kenntnisse werden nicht vorausgesetzt. Nähere Informationen sowie die Möglichkeit zur Anmeldung finden Sie unter: <http://www.ub.uni-frankfurt.de/digitalhumanities/workshops.html>.

Praxislabor Digitale Geisteswissenschaften: Einführungsworkshops zu Methoden und Werkzeugen der Digital Humanities

Startseite / Kurse / Verschiedenes / Praxislabor Digitale Geisteswissenschaften

Allgemeines

In dieses kollaborative Dokument können Themenvorschläge für die Hands-on-Sessions eingetragen werden. Informationen zur Anmeldung und Kurszeiten unter: <http://www.ub.uni-frankfurt.de/digitalhumanities>

Einführung in TEI / XML

Dozentin: Agnes Brauer

Der Workshop führt in die Grundlagen der Textauszeichnung mit TEI ein, einer XML-basierten und sich mittlerweile als De-facto-Standard etablierten Auszeichnungssprache speziell für die Zwecke der Geisteswissenschaften. Nach einer knappen allgemeinen Einführung werden die Teilnehmer anhand einer kleinen Übung die Praxis der Textauszeichnung mit TEI kennenlernen und sich einen ersten Überblick über die Bedeutung und die verschiedenen Module dieser Sprache verschaffen.

Link: <http://www.tei-c.org/>

Hands-on Übung zur TEI/XML-Einführung

Dozentin: Agnes Brauer

Einführungsworkshops zu Methoden und Werkzeugen der Digital Humanities

Praxislabor Digitale Geisteswissenschaften:
Einführungsworkshops zu Methoden und Werkzeugen der Digital Humanities

Zum ersten Kennenlernen von Methoden und Werkzeugen der Digital Humanities bietet die Universitätsbibliothek JCS (im Bibliothekszentrum Geisteswissenschaften) Studierenden und Mitarbeiterinnen der Goethe-Uni im kommenden Wintersemester Workshops an. In niederschweligen Einführungen werden anhand von überschaubaren, konkreten Beispielen aus der Praxis Methoden, Tools oder Themen der digitalen Geisteswissenschaften vorgestellt und geübt und so ein erster Einblick in die Möglichkeiten gegeben, wie klassische Methoden der Geisteswissenschaften mithilfe digitaler Verfahren der Textanalyse sowie der Text- und Datenaufbereitung sinnvoll ergänzt werden können.

Die Workshopreihe besteht jeweils aus inhaltlich zusammenhängenden Zweierblöcken, in denen auf eine Präsentation eine Sitzung zur Vertiefung und Übung folgt.

Die Workshops richten sich an interessierte Einsteiger; besondere Kenntnisse werden nicht vorausgesetzt. Nähere Informationen sowie die Möglichkeit zur Anmeldung finden Sie unter: <http://www.ub.uni-frankfurt.de/digitalhumanities/workshops.html>

Trainerin: Agnes Brauer
Trainerin: Jakob Frohmann

Selbsteinschreibung (Teilnehmer/in)

Kein Einschreibeschlüssel notwendig

EINSCHREIBEN

Hinweis

- Tragen Sie bitte bei Bedarf / Interesse Themenvorschläge für die Hands-on Übung in das [kollaborative Dokument](#) ein

Webscraping mithilfe von XPath

- *XPath* ist eine Abfragesprache und dient der **Suche und Navigation** innerhalb von XML-Dokumenten
- *XPath*-Ausdrücke lokalisieren Teile eines XML-Dokuments und lesen ihre Eigenschaften aus
- XPath kann auch für XML-ähnliche Strukturen wie **HTML** verwendet werden
- diese Eigenschaft macht sich der Chrome Scraper zu Nutze

Aufbau eines XML-Dokuments

- XML-Deklaration: `<?xml version="1.0" encoding="UTF-8"?>`
- Wurzelement
- Elemente: `<head>` Dies ist eine Überschrift `</head>`
- Attribute: `<hi rend=„italic“>` Dies ist eine kursive Hervorhebung `</hi>`
- `<!-- Kommentare -->`

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
```

```
<fileDesc>
  <titleStmt>
    <title>TEI-Minimal-Beispiel</title>
  </titleStmt>
  <publicationStmt>
    <p>Frei verfügbar</p>
  </publicationStmt>
  <sourceDesc>
    <p>Dieser Text ist digital born.</p>
  </sourceDesc>
</fileDesc>
```

```
<!-- Ein XML Kommentar -->
```

```
<head>Minimalbeispiel</head>
```

```
<hi rend="italic">Textauszeichnung mit
TEI</hi>
```

```
</body>
</text>
</TEI>
```

Beispiel einer HTML-Datei

```
<!DOCTYPE html>
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <meta charset="utf-8" />
    <title>HTML-Minimal-Beispiel</title>
  </head>
  <body>
    <p>Ein Beispieltext von <b>Agnes Brauer</b>
      <br/>für die Übung <i>Textauszeichnung mit TEI</i>.</p>
  </body>
</html>
```

XPath

- Ein XPath besteht aus einem oder mehreren **Pfadabschnitten** (location steps)
- die Pfadabschnitte bestehen aus einem Schrägstrich (/) und einem **Knotentest** (node test)
- dem Knotentest kann eine **Achse** (axis) vorangestellt werden
- die Ergebnismenge eines Pfadabschnitts kann durch **Bedingungen** (predicates) eingeschränkt werden
- der letzte angegebene Knotentest im XPath bestimmt den **Typ** des Ergebnisses

XPath

Knotentypen:

- **Element:** geprüft über **Knotennamen** oder ***** (als Abkürzung für ein beliebiges Element)
- **Attribut:** geprüft über **@Knotennamen** oder **@***
- **Text:** geprüft durch **text()**
- **Kommentar:** geprüft durch **comment()**

XPath

Beispiel für einen XPath:

/ html / body / h1

Knotentest
(node test)

/ Pfadabschnitt
(location step)

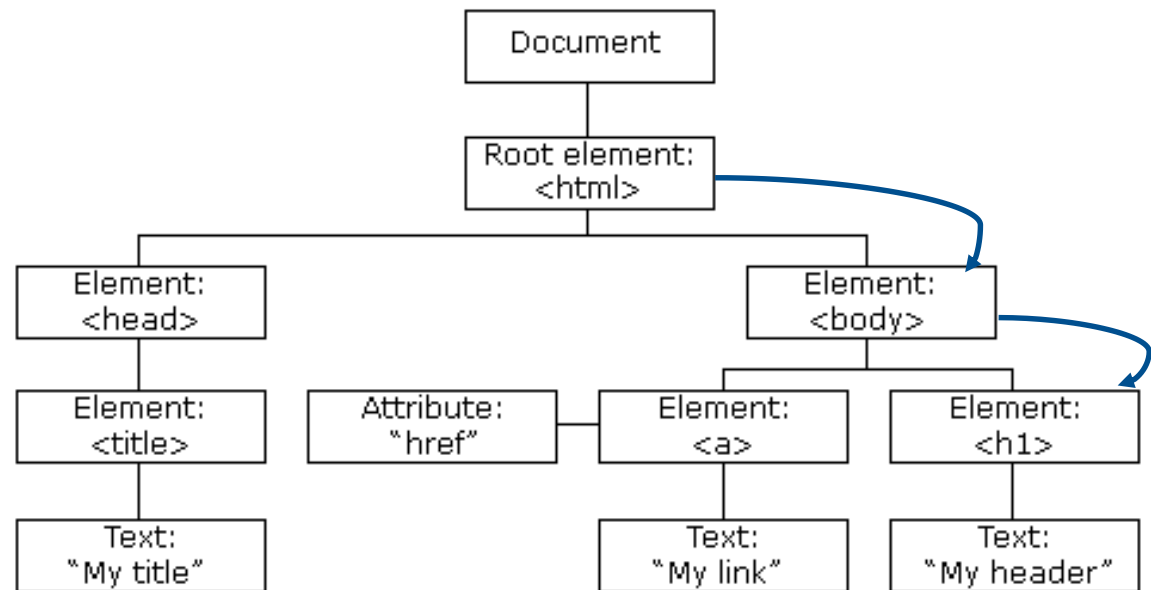


Abb.: <https://librarycarpentry.org/lc-webscraping/02-xpath/index.html>

Wichtige XPath-Achsen

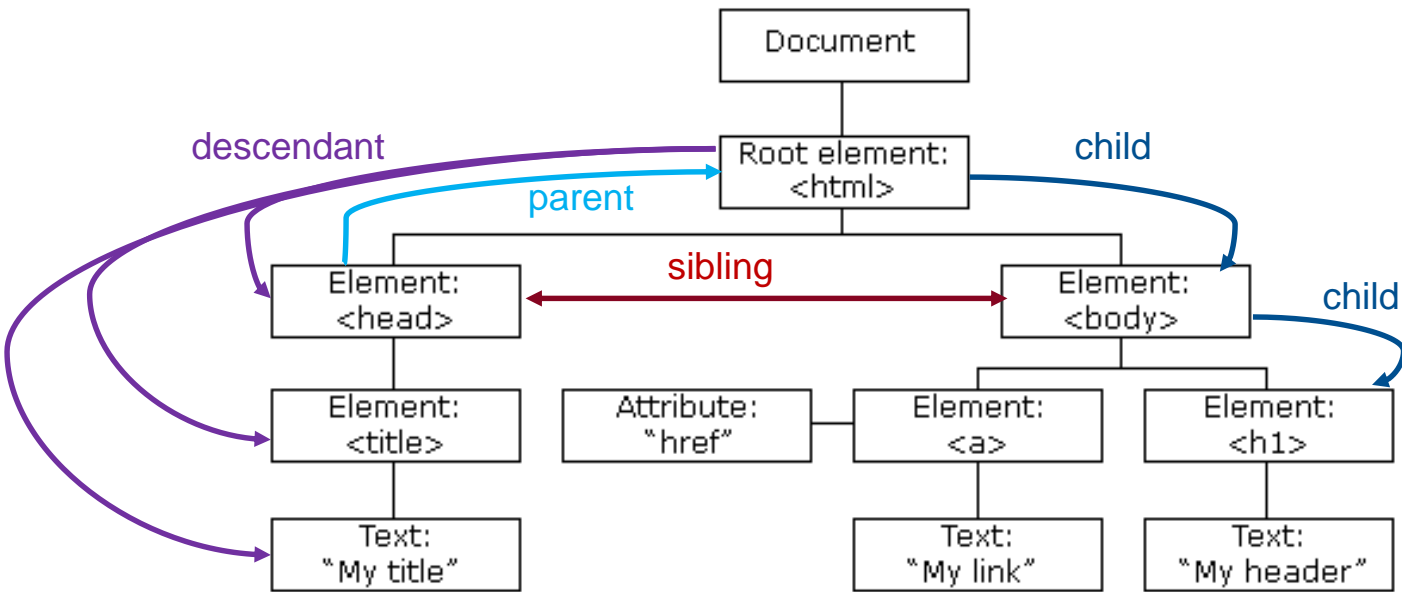


Abb.: <https://librarycarpentry.org/lc-webscraping/02-xpath/index.html>

XPath

Wichtige Achsen:

self:: der aktuelle Kontextknoten (.)

child:: direkte Kindelemente ()

parent:: direkter Elternknoten (..)

ancestor:: alle Vorfahren

descendant:: alle Nachkommen (//)

preceding:: alle Knoten vorher

following:: alle Knoten nachher

following-sibling:: alle Geschwisterknoten nachher (gemeinsamer Elternknoten)

attribute:: alle Attribute (@)

XPath

Ergebnisse einschränken:

- um eine bestimmte Einschränkung des Abfrageergebnisses zu erlangen, können sogenannte **Prädikate** verwendet werden
- hierbei handelt es sich um zusätzliche Bedingungen, die an den Knotentest geknüpft werden

Beispiel:

```
/ TEI / text / body / p / hi [@rend='italic']
```

Einschränkung
predicate

XPath

Beispiele für die Verwendung von XPath-Funktionen:

```
/ TEI / text / body // head [starts-with(text(), 'Herz') ]
```

```
/ TEI / text / body // head / string-length()
```

```
/TEI/text/body//p[last()]
```

```
//p[ exists(./term)]
```

```
/ TEI / text / body //p [ not( exists(./ q ) ) ]
```

```
//p[1]/term/substring(text(), 1,6)
```

```
exists( //head [ . / parent::body ] )
```

```
/ TEI / text / body // div [ not( exists(./ head ) ) ]
```

Links

<https://www.w3.org/TR/xpath-31/>

https://www.w3schools.com/xml/xsl_functions.asp

<trailer>Vielen Dank für Ihre Aufmerksamkeit!</trailer>