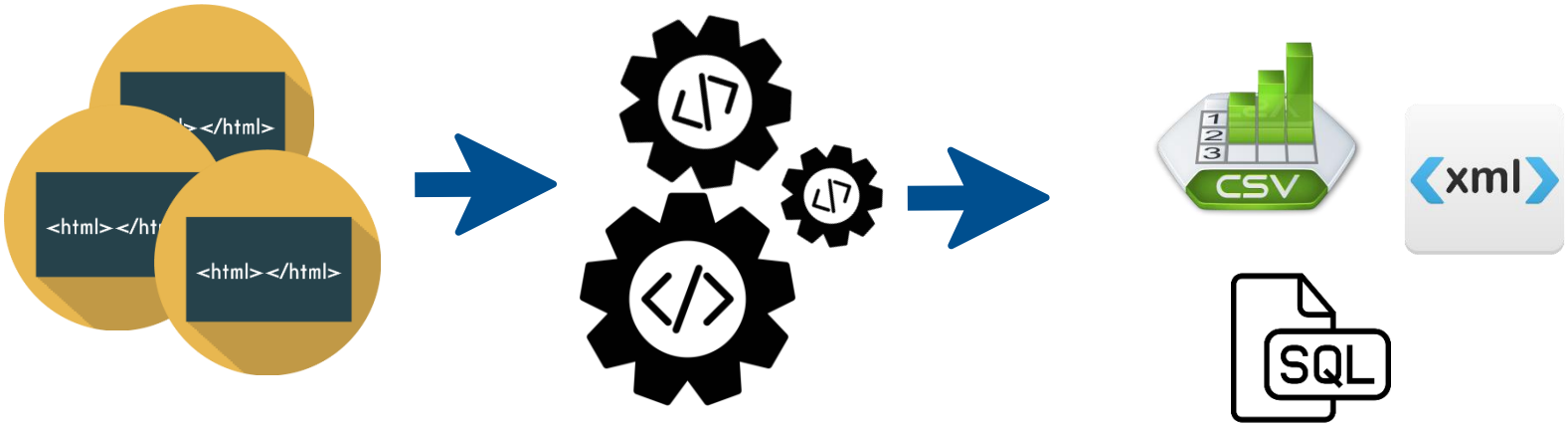
UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

Einführung in manuelles Webscraping mit dem Browser-Plugin *Scraper*



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

Was ist Webscraping?

- Technik zur Extraktion von spezifischen Informationen aus Webseiten
- Überführung von unstrukturierten Daten in ein nutzbares, strukturiertes Format
- Anwendungsfälle:
  - Marktstudien (kommerziell)
  - Aggregation von Datensets für wissenschaftliche Fragestellungen / Auswertungen
  - Für Zwecke der Archivierung
  - …

Voraussetzung für viele Scraping-Techniken: XPath

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

- *XPath* ist eine Abfragesprache und dient der **Suche und Navigation** innerhalb von XML-Dokumenten
- XPath gehört zu den sogenannten X-Technologien und ist Bestandteil vieler Anwendungen
- *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

XML- Basics

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

- XML-Dokumente haben eine Baumstruktur, die durch Knoten strukturiert sind:
  - Elemente
  - Attribute
  - Textknoten

- XML-Elemente haben öffnende und schließende Tags. Z.B.:
  <head>Dies ist eine Überschrift</head>

- XML-Tags sind case-sensitive, d.h.:
  <head> ist nicht gleich <Head>

- XML-Elemente müssen ordentlich geschachtelt sein, z.B.:

```
<beispiele>
  <beispiel>Erstes Beispiel</beispiel>
  <beispiel>Zweites Bespiel</beispiel>
</beispiele>
```

Aufbau eines XML-Dokuments

- XML-Deklaration: <?xml version="1.0" encoding="UTF-8"?>
- Wurzelelement
- 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">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>TEI-Minimal-Beispiel</title>
      </titleStmt>
      <publicationStmt>
        <p>Frei verfügbar</p>
      </publicationStmt>
      <sourceDesc>
        <p>Dieser Text ist digital born.</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <!-- 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>
        <h1>Beispieltext</h1>
        <p>Ein Beispieltext von <b>Agnes Brauer</b>
          <br/>für die Übung <i>Textauszeichnung mit TEI</i>.</p>
      </body>
    </html>
```

XPath

Beispiel für einen XPath:

/ html / body / h1

Knotentest
(node test)

/ Pfadabschnitt
(location step)

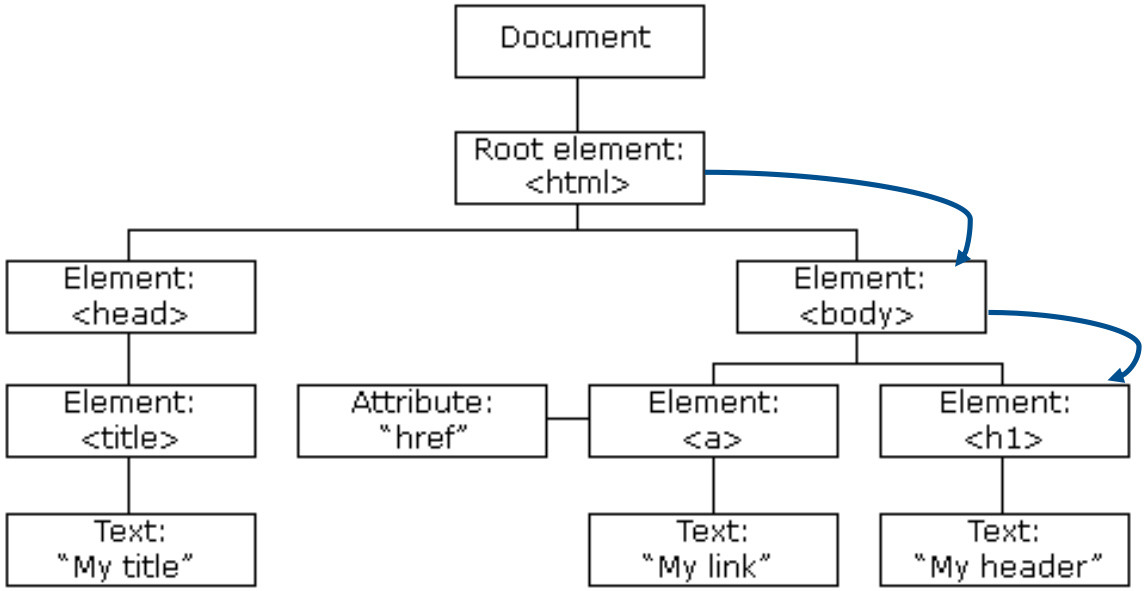


Abb.: https://librarycarpentry.org/lc-webscraping/02-xpath/index.html

XPath - Grundlagen

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

- 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()**

Wichtige XPath-Achsen

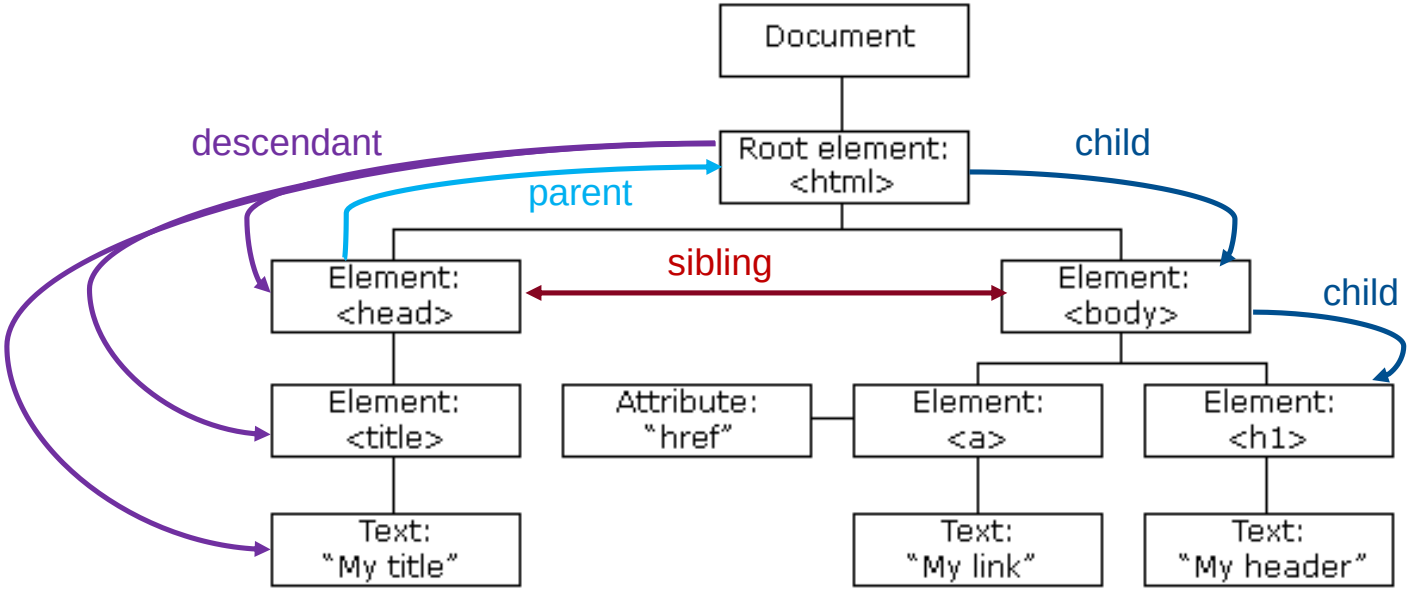


Abb.: https://librarycarpentry.org/lc-webscraping/02-xpath/index.html

Xpath-Achsen

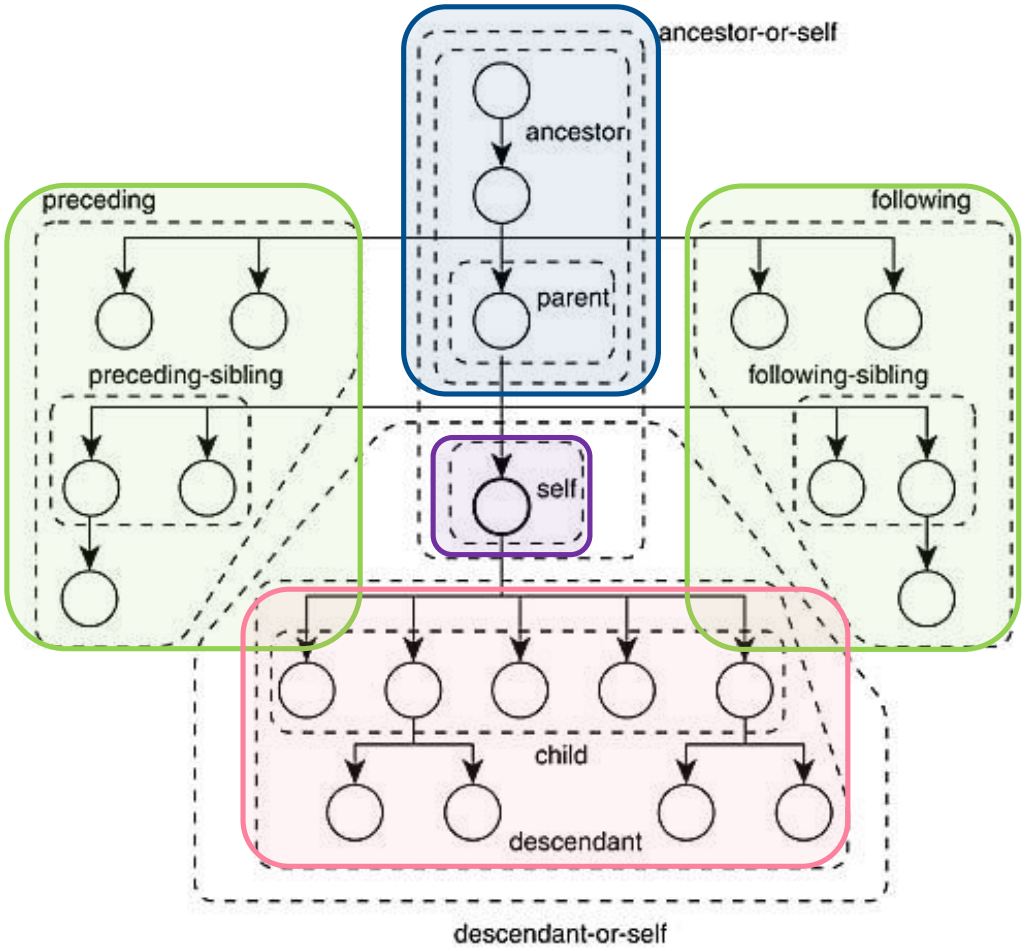


Abb.: https://librarycarpentry.org/lc-webscraping/02-xpath/index.html

XPath

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

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: Wichtige Selektionen

• nodename	selektiert Knotenpunkte des gleichen Namens
• /	selektiert Kindknoten
• //	Selektiert alle direkten und indirekten Childnodes (d.h. alle Nachfahren) → erlaubt ein (Über-)Springen
• .	meint aktuell selektierten/s Element/Knoten (=Kontextknoten)
• ..	selektiert das Elternelement des aktuell gesuchten Elements
• @	selektiert Attribute des Kontextknoten
• [@attribute='value']	selektiert Elemente mit Attributen, die einen bestimmten Wert haben
• text()	selektiert den Text eines Elements

Quelle: https://librarycarpentry.org/lc-webscraping/02-xpath/index.html

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 - Funktionen

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

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

Chrome Scraper

Übung im Pad

Chrome Scraper

Übung

Extraktion von Informationen der Seite: https://www.geschichtsquellen.de

Schritt 1:

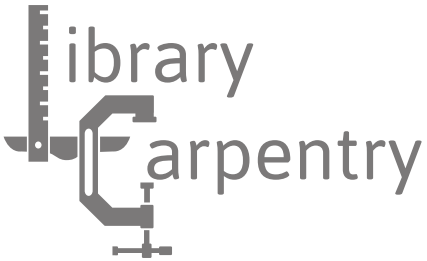
- Eingrenzen der Treffermenge:

  Suche > Erweiterte Suche > Filter Gattung = Chronologie → 30 Treffer

Schritt 2:

- Scrapen der Informationen:
  - Titel
  - Autor
  - Link zur ausführlichen Beschreibung

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



Workshop konzipiert in Anlehnung an Library Carpentry: Introduction to web scraping.