ÖAW AUSTRIAN ACADEMY OF SCIENCES

Summer School “Digitale Methoden der Zeitungsanalyse, 2. Auflage”
4. September 2025

Metadatenanalyse am Beispiel von digiPress & Zeitschriftendatenbank

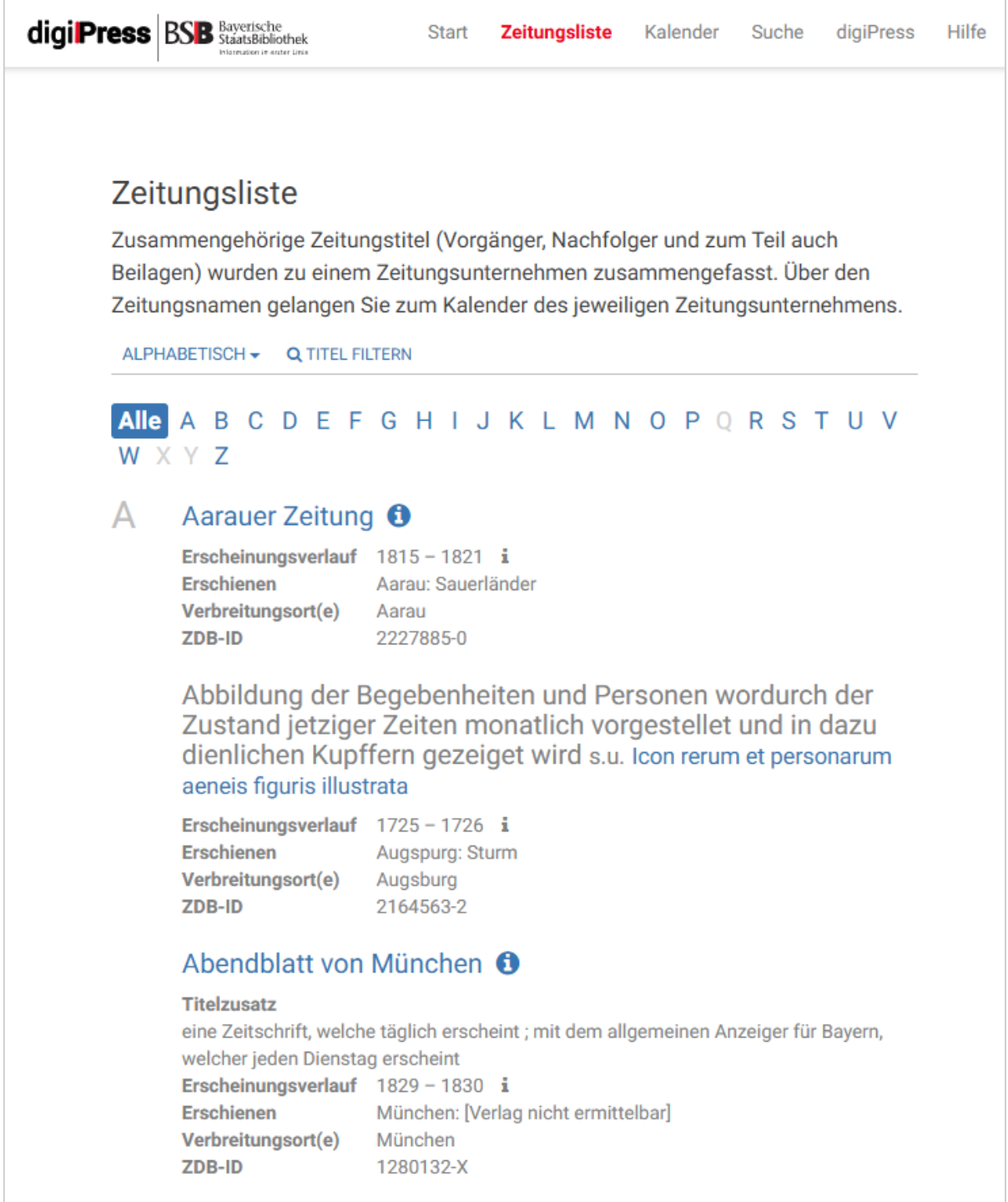
Nina C. Rastinger
(Austrian Centre for Digital Humanities, Österreichische Akademie der Wissenschaften)

ÖAW AUSTRIAN ACADEMY OF SCIENCES

Summer School “Digitale Methoden der Zeitungsanalyse, 2. Auflage”
4. September 2025

Metadatenanalyse am Beispiel von digiPress & ~~Zeitschriftendatenbank~~

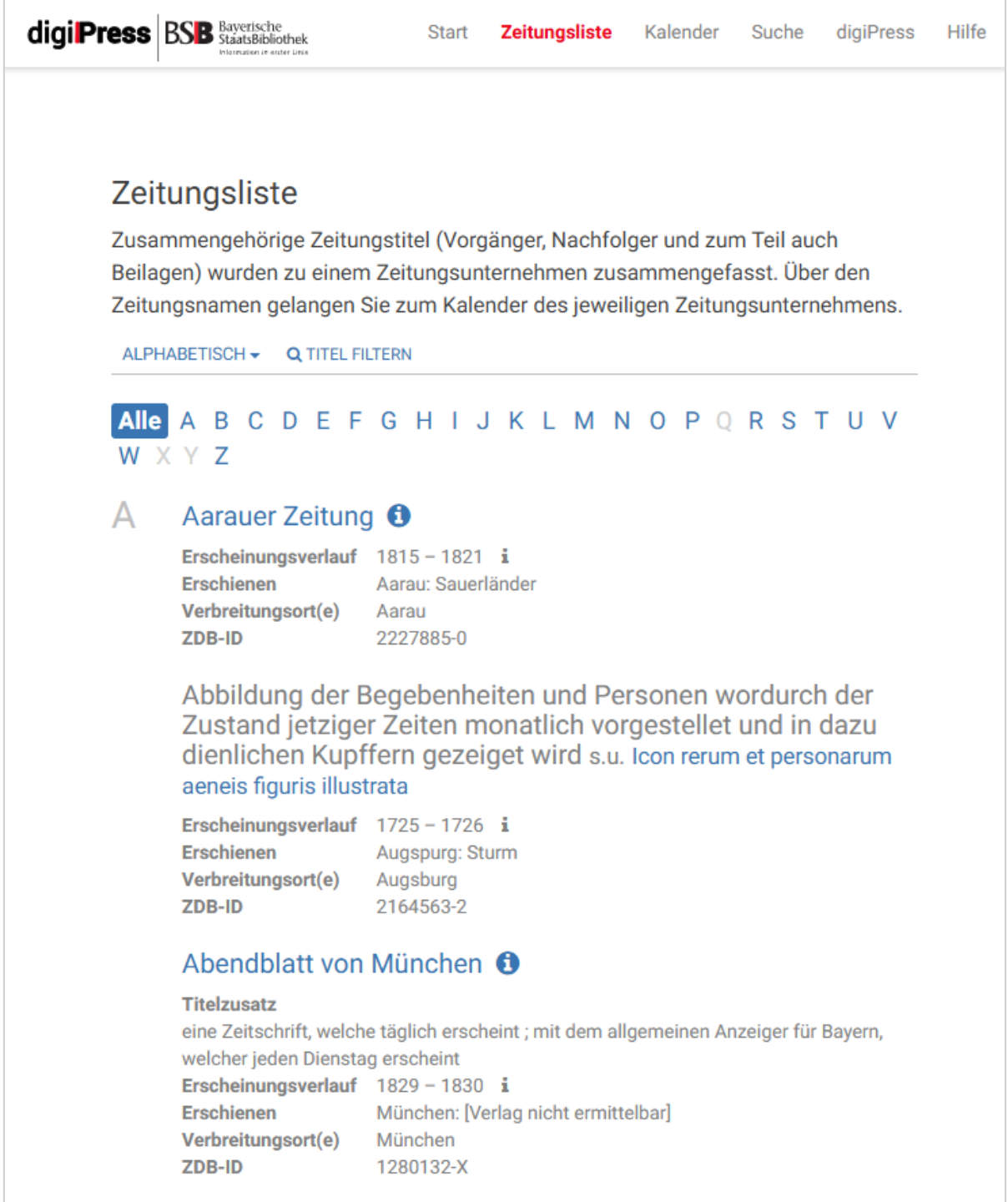
Nina C. Rastinger
(Austrian Centre for Digital Humanities, Österreichische Akademie der Wissenschaften)



Zeitungsliste von digiPress

https://digipress.digitale-sammlungen.de/titles

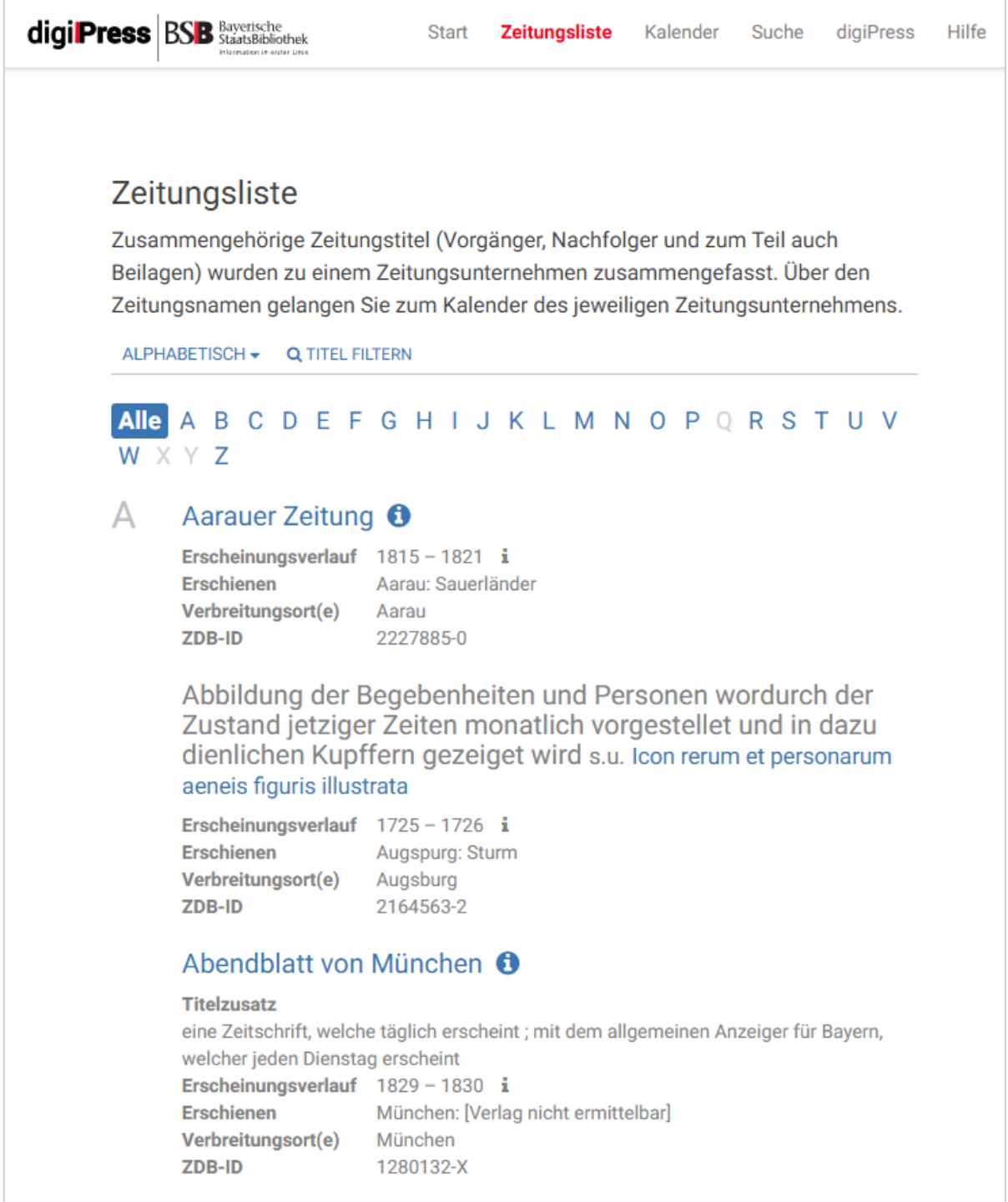
enthält alle Titel + Link zu Ausgaben + Metadaten (Erscheinungszeitraum, Publikationsort, Verlag, Verbreitungsort(e), ZDB-ID, …)



Zeitungsliste von digiPress

Integrierte Möglichkeiten:

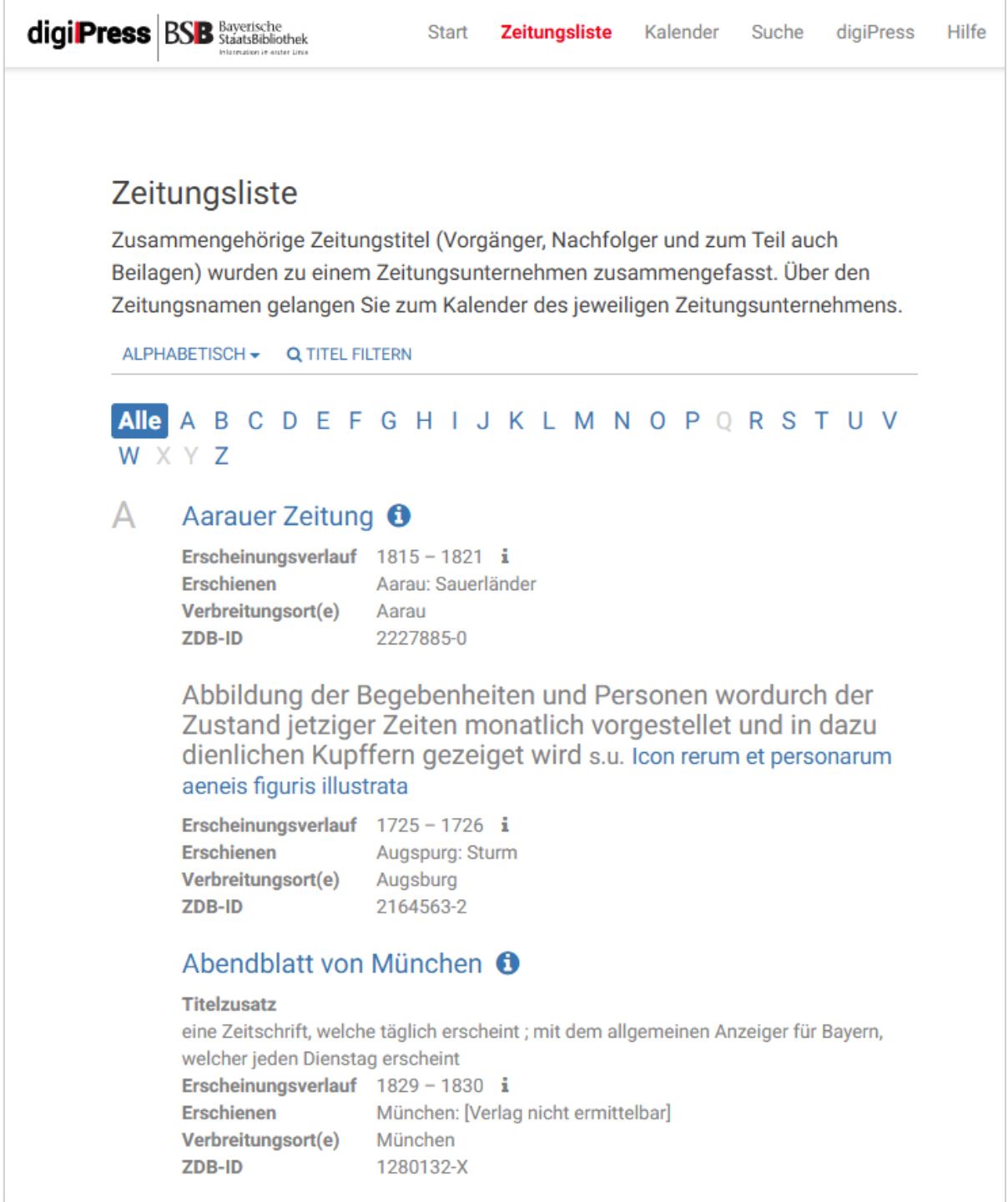
- Filterung nach Titel (= Volltextsuche in Angabe von Titel bzw. Unternehmen)
- Alphabetische Ordnung nach Titel
- Alphabetische Ordnung nach Verbreitungsort



Zeitungsliste von digiPress

Wofür ist dieser Zugang gut geeignet, wofür weniger?

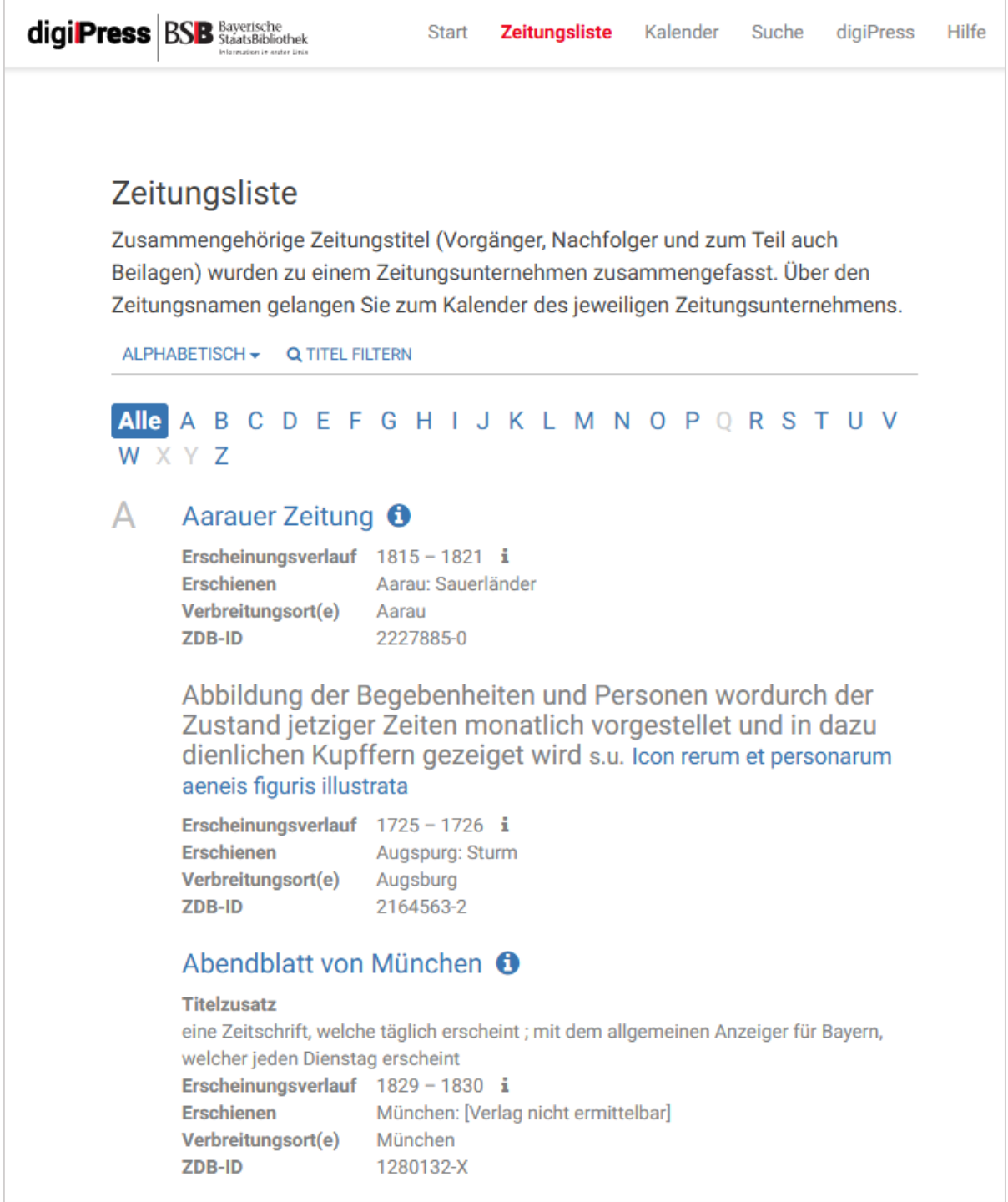
Welche Fragen über den Bestand des Portals lassen sich über die Zeitungsliste effizient beantworten, welche benötigen mehr (Zeit-)Aufwand?



Zeitungsliste von digiPress

Sehr gut geeignet für:

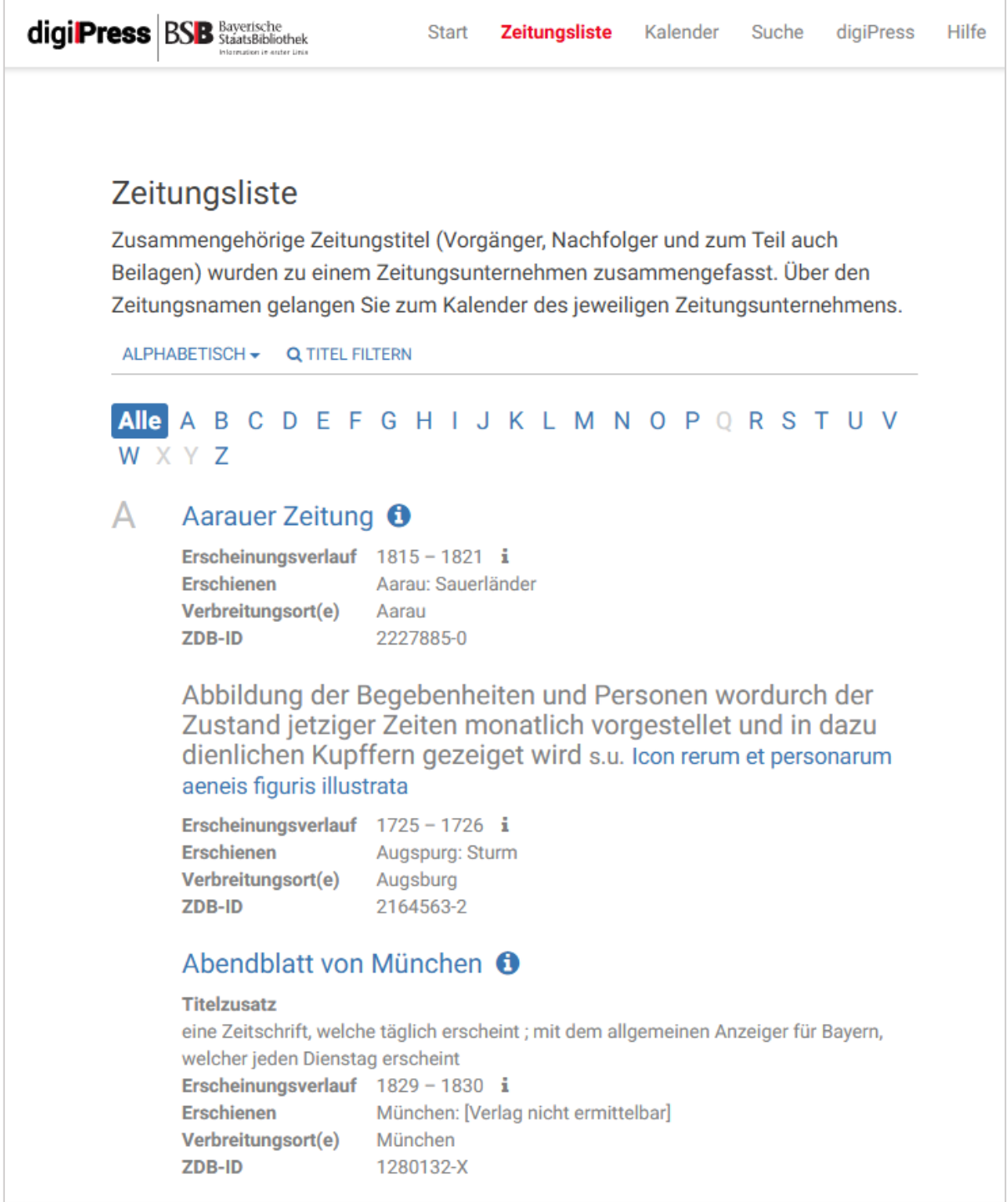
- Punktuelle Einstiege
  - Interesse an konkreten Titeln oder Unternehmen: Ist Zeitung XY digital verfügbar?
  - Interesse an der Presselandschaft bestimmter Orte: Welche Periodika sind für Ort XY vorhanden?
- Erster kursorischer Überblick über angebotene Digitalisate



Zeitungsliste von digiPress

Durch große Menge an Titeln (>1.300) weniger gut geeignet für Gesamtüberblick:

- Wie viele Titel sind insgesamt verfügbar?
- Wie viele Unternehmen sind insgesamt verfügbar?
- Aus welchen Jahrhunderten stammen die in digiPress verfügbaren Periodika am häufigsten? *kursorisch auch über Kalenderansicht einsehbar
- Wie verteilen sie sich räumlich, von welchen Publikationsorten/ Ländern kommen sie?



Zeitungsliste von digiPress

Zusätzlich könnten – je nach Projekt – weitere Filter von Interesse sein, z.B.:

- Filter nach Zeit (z.B. nur von 1800-1820 erschienene Periodika)
- Filter nach Verlag
- Filter nach bestimmtem Land
- …

→ Einzelne Zeitungsportale bieten erweiterte Abfragemöglichkeiten über **API (= Programmierschnitt-stelle)** – in digiPress (noch) nicht integriert

ÖAW AUSTRIAN ACADEMY OF SCIENCES

Alternative: Einsatz von Screen Scraping

Umwandlung der Zeitungsliste als Website in strukturierten, weiterverarbeitbaren Datensatz

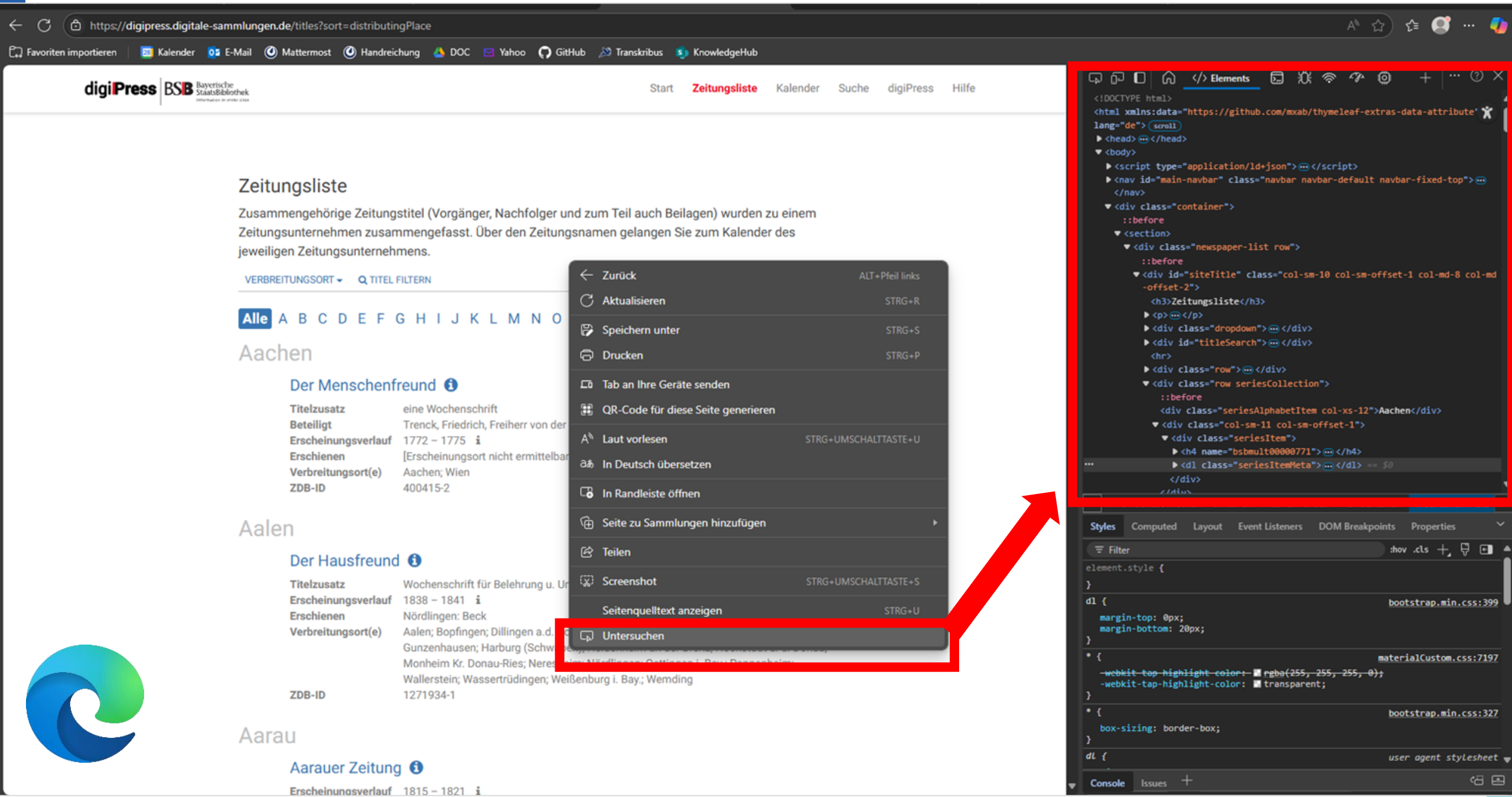
→ Nutzen des bereits strukturierten **HTMLs**, das – wie bei jeder Website – im Hintergrund liegt

Wer hat schon mal “hinter” die “Kulissen” des Internets geschaut und das HTML einer Website betrachtet?

ÖAW AUSTRIAN ACADEMY OF SCIENCES

HTML in 100 Sekunden

https://youtu.be/ok-plXXHlWw?si=dxvV0YSBIF0WloBn

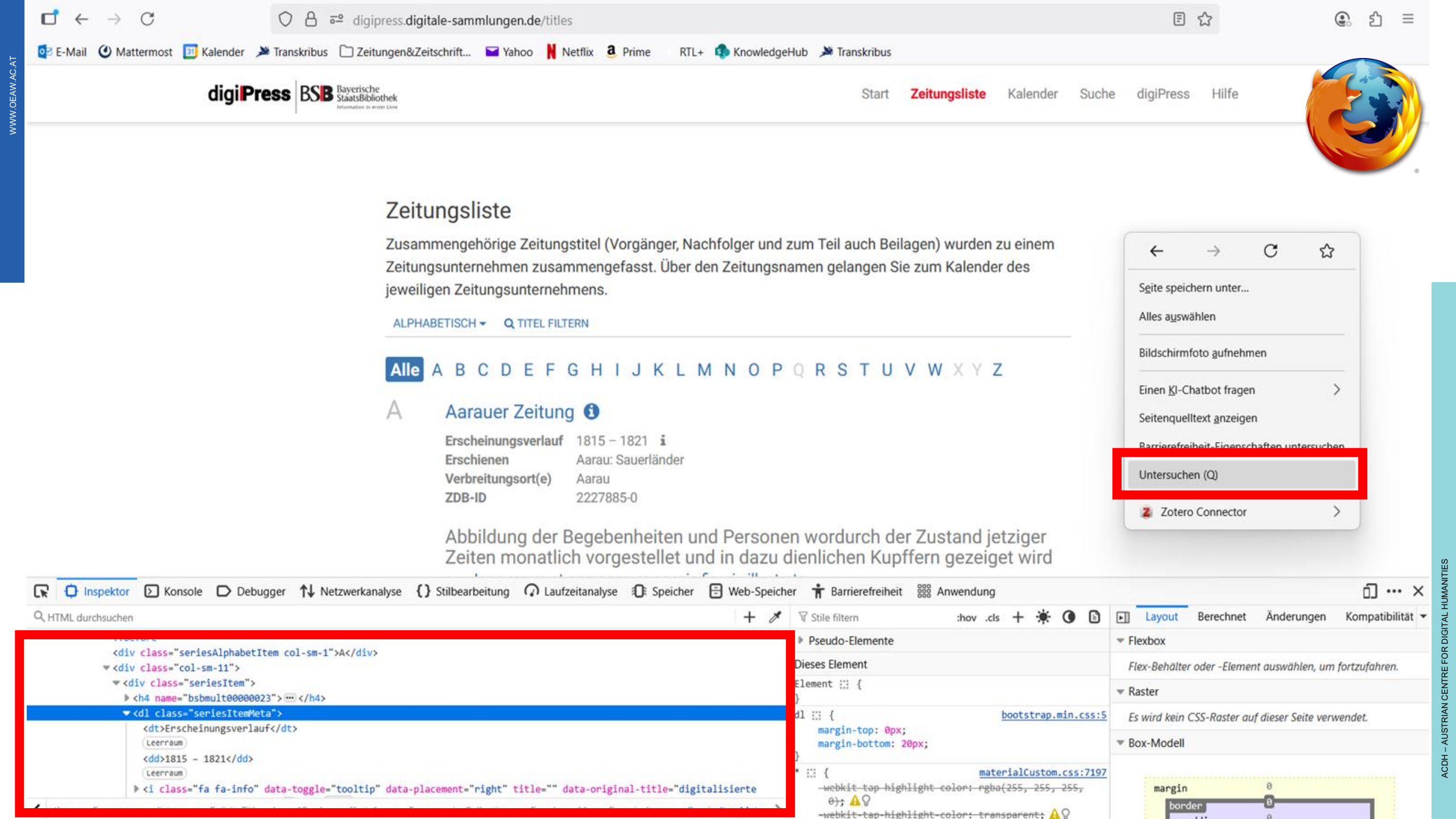

https://digipress.digitale-sammlungen.de/titles?sort=distributingPlace
Favoriten importieren
Kalender
E-Mail
Mattermost
Handreichung
DOC
Yahoo
GitHub
Transkribus
KnowledgeHub
digiPress
BSB Bayerische StaatsBibliothek
Start
Zeitungsliste
Kalender
Suche
digiPress
Hilfe
Zeitungsliste
Zusammengehörige Zeitungstitel (Vorgänger, Nachfolger und zum Teil auch Beilagen) wurden zu einem Zeitungsunternehmen zusammengefasst. Über den Zeitungsnamen gelangen Sie zum Kalender des jeweiligen Zeitungsunternehmens.
VERBREITUNGSORT
TITEL FILTERN
Alle A B C D E F G H I J K L M N O
Aachen
Der Menschenfreund
Titelzusatz eine Wochenschrift
Beteiligt Trenck, Friedrich, Freiherr von der
Erscheinungsverlauf 1772 – 1775
Erschienen [Erscheinungsort nicht ermittelbar
Verbreitungsort(e) Aachen; Wien
ZDB-ID 400415-2
Aalen
Der Hausfreund
Titelzusatz Wochenschrift für Belehrung u. Ur
Erscheinungsverlauf 1838 – 1841
Erschienen Nördlingen: Beck
Verbreitungsort(e) Aalen; Bopfingen; Dillingen a.d.
Gunzenhausen; Harburg (Schw
Monheim Kr. Donau-Ries; Neres
Wallerstein; Wassertrüdingen; Weißenburg i. Bay.; Wemding
ZDB-ID 1271934-1
Aarau
Aarauer Zeitung
Erscheinungsverlauf 1815 – 1821
Zurück ALT+Pfeil links
Aktualisieren STRG+R
Speichern unter STRG+S
Drucken STRG+P
Tab an Ihre Geräte senden
QR-Code für diese Seite generieren
Laut vorlesen STRG+UMSCHALTTASTE+U
In Deutsch übersetzen
In Randleiste öffnen
Seite zu Sammlungen hinzufügen
Teilen
Screenshot STRG+UMSCHALTTASTE+S
Seitenquelltext anzeigen STRG+U
Untersuchen
Elements
<!DOCTYPE html>
<html xmlns:data="https://github.com/mxab/thymeleaf-extras-data-attribute" lang="de">
<head>…</head>
<body>
<script type="application/ld+json">…</script>
<nav id="main-navbar" class="navbar navbar-default navbar-fixed-top">…</nav>

<section>

<h3>Zeitungsliste</h3>
…
…
…
<hr>
…

Aachen

<h4 name="bsbmult00000771">…</h4>
<dl class="seriesItemMeta">…</dl> == $0
Styles
Computed
Layout
Event Listeners
DOM Breakpoints
Properties
Filter
:hov .cls
element.style {
}
dl { bootstrap.min.css:399
margin-top: 0px;
margin-bottom: 20px;
}
* { materialCustom.css:7197
-webkit-tap-highlight-color: rgba(255, 255, 255, 0);
-webkit-tap-highlight-color: transparent;
}
* { bootstrap.min.css:327
box-sizing: border-box;
}
dl { user agent stylesheet
Console
Issues



ÖAW AUSTRIAN ACADEMY OF SCIENCES

Screen Scraping oder Web Scraping

„Der Begriff **Screen Scraping** (engl., etwa: „am Bildschirm schürfen“) umfasst generell alle Verfahren zum Auslesen von Texten aus Computerbildschirmen. Gegenwärtig wird der Ausdruck jedoch beinahe ausschließlich in Bezug auf Webseiten verwendet (daher auch **Web Scraping** oder **Web Harvesting**).“ („Screen Scraping“, Wikipedia)

Bild generiert mit DALL-E

ÖAW AUSTRIAN ACADEMY OF SCIENCES

Exkurs: **Webscraper** vs. **Webcrawler**

- **Webscraping** = Extrahieren von Daten von einer oder mehreren Websites
- **Webcrawling** = Identifizieren von (relevanten) URLs oder Links auf einer Website
- Begriffe oft synonym verwendet
- Gehen meist auch Hand in Hand, z.B. Links diverser Websites identifizieren, um von diesen wiederum Daten zu sammeln

Bild generiert mit DALL-E

ÖAW AUSTRIAN ACADEMY OF SCIENCES

Aufgabe: Blick “hinter” die Zeitungsliste

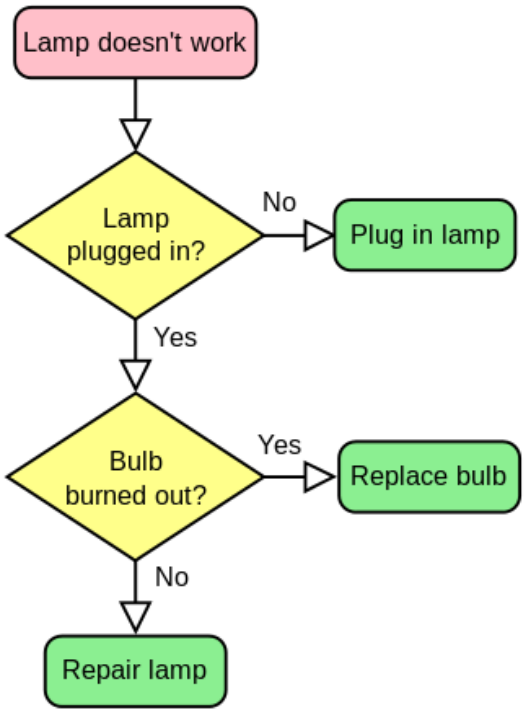
(ca. 10-15 Min, Kleingruppen zu 2-3 Personen, dann Besprechen der Lösungen im Plenum)

Untersucht das HTML der digiPress-Zeitungsliste und beantwortet folgende Fragen:

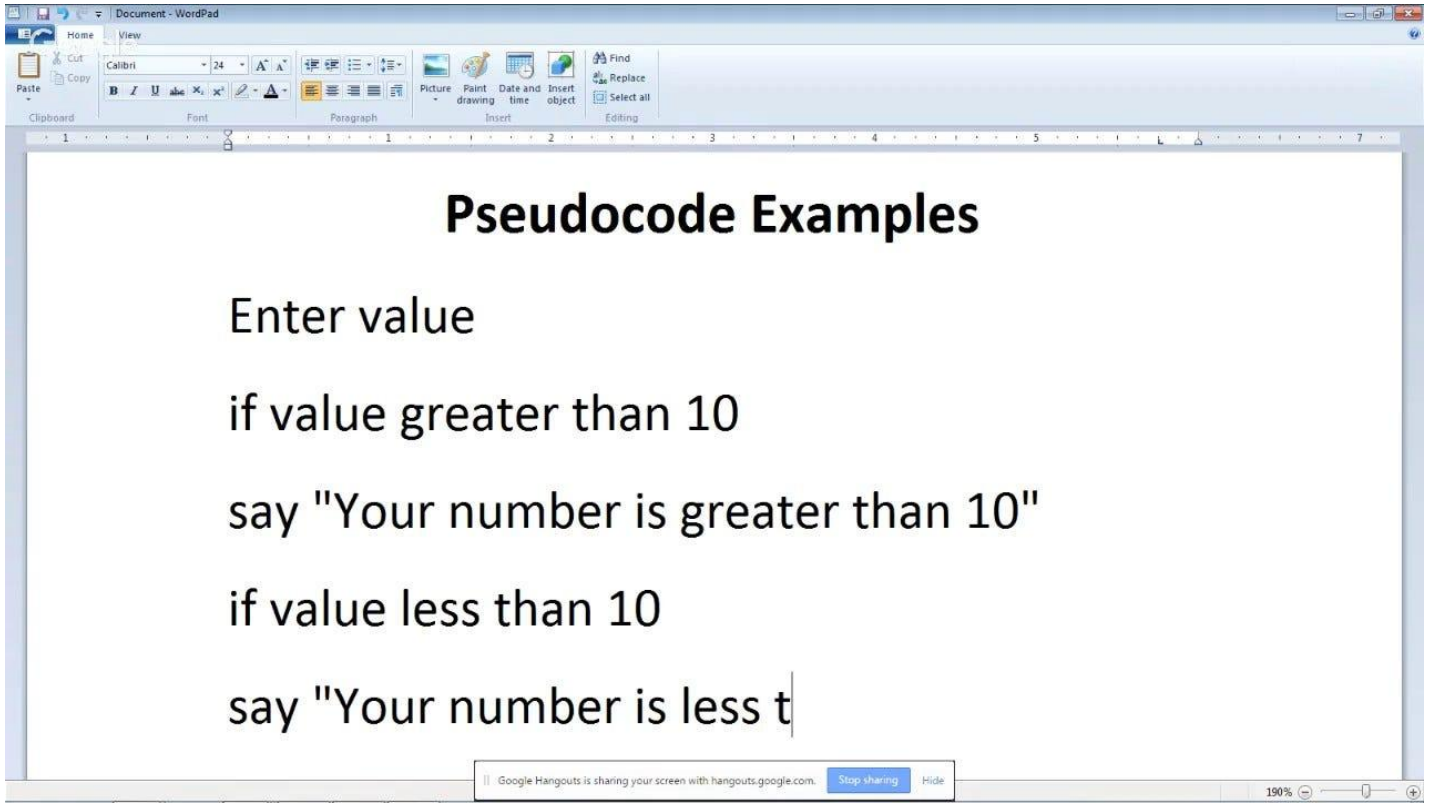
- Welcher Struktur folgt die Zeitungsliste?
- Mehr Infos zu HTML-Elementen finden sich hier: https://developer.mozilla.org/en-US/docs/Web/HTML/Reference/Elements
- Wo befinden sich für uns interessante Informationen (z.B. Zeitungstitel, Erscheinungsverlauf, digiPress-ID, …)?
- Wonach müsste ein Webscraper suchen, um diese Inhalte zu extrahieren?
- Wenn Zeit: Versucht, die Suchanweisungen in Form eines “Pseudocodes” zu verschriftlichen (s. Beispiele)

ÖAW AUSTRIAN ACADEMY OF SCIENCES

Pseudocode



A simple flowchart representing a process for dealing with a non-functioning lamp.



https://en.wikipedia.org/wiki/Flowchart / https://www.youtube.com/watch?v=11YjjL4Ib2Q

ÖAW AUSTRIAN ACADEMY OF SCIENCES

Zeitungsliste = Liste von Containern oder Sektionen (div) mit Klasse “seriesItem”

Alle A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A

Aarauer Zeitung

Erscheinungsverlauf	1815 – 1821
Erschienen	Aarau: Sauerländer
Verbreitungsort(e)	Aarau
ZDB-ID	2227885-0

```
<div class="seriesItem"> ··· </div>
<div class="seriesItem"> ··· </div>
<div class="seriesItem"> ··· </div>
<div class="seriesItem"> ··· </div>
```

Abbildung der Begebenheiten und Personen wordurch der Zustand jetziger Zeiten monatlich vorgestellet und in dazu dienlichen Kupffern gezeiget wird

s.u. Icon rerum et personarum aeneis figuris illustrata

Erscheinungsverlauf	1725 – 1726
Erschienen	Augspurg: Sturm
Verbreitungsort(e)	Augsburg

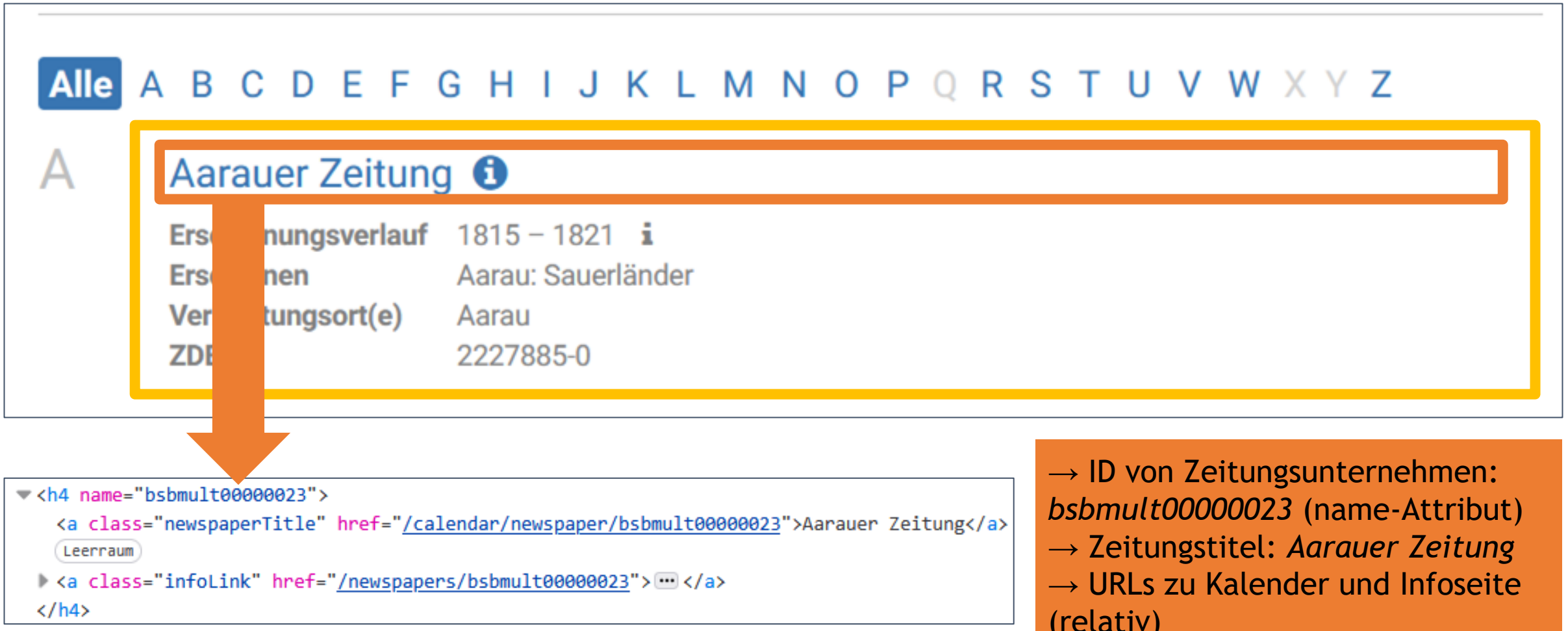
ÖAW AUSTRIAN ACADEMY OF SCIENCES

Inhalt jedes einzelnen Containers:
Überschrift (h4) + Beschreibungsliste (dl)


```
<div class="seriesItem">
  <h4 name="bsbmult00000023"> ... </h4>
  <dl class="seriesItemMeta"> ... </dl>
</div>
```

ÖAW AUSTRIAN ACADEMY OF SCIENCES

Welche Informationen können wir aus dem h4-Element sammeln?




```
<h4 name="bsbmult00000023">
  <a class="newspaperTitle" href="/calendar/newspaper/bsbmult00000023">Aarauer Zeitung</a>
  Leerraum
  <a class="infoLink" href="/newspapers/bsbmult00000023">…</a>
</h4>
```

→ ID von Zeitungsunternehmen: *bsbmult00000023* (name-Attribut)
→ Zeitungstitel: *Aarauer Zeitung*
→ URLs zu Kalender und Infoseite (relativ)

ÖAW AUSTRIAN ACADEMY OF SCIENCES

Welche Informationen können wir aus dem h4-Element sammeln?


```
<h4 name="bsbmult00000023">
  <a class="newspaperTitle" href="/calendar/newspaper/bsbmult00000023">Aarauer Zeitung</a>
  Leerraum
  <a class="infoLink" href="/newspapers/bsbmult00000023">…</a>
</h4>
```

→ gegeben in Form von zwei <a>-Elementen (= Links)
→ für uns von Interesse: <a> mit class = “newspaperTitle”

ÖAW AUSTRIAN ACADEMY OF SCIENCES

Vorsicht vor Abweichungen: stellenweise auch <h4>-Elemente mit Klasse “noLink” und anderer Struktur

Abbildung der Begebenheiten und Personen wordurch der Zustand jetziger Zeiten monatlich vorgestellet und in dazu dienlichen Kupffern gezeiget wird

s.u. Icon rerum et personarum aeneis figuris illustrata

Erscheinungsverlauf	1725 – 1726
Erschienen	Augspurg: Sturm
Verbreitungsort(e)	Augsburg
ZDB-ID	2164563-2

```
<h4 class="noLink">
  <span>
    Abbildung der Begebenheiten und Personen wordurch der Zustand jetziger Zeiten monatlich vorgestellet und in dazu
    dienlichen Kupffern gezeiget wird
  </span>
  <span class="smaller">s.u.</span>
  Leerraum
  <a class="smaller linkToPrimaryTitle" href="#bsbmult00000658">Icon rerum et personarum aeneis figuris illustrata</a>
  event
</h4>
```

→ Titel, die Teil von anderem Zeitungs-unternehmen sind

→ bei Crawl ebenfalls zu berücksichtigen

ÖAW AUSTRIAN ACADEMY OF SCIENCES

Welche Informationen können wir aus dem dl-Element sammeln?


```
<dl class="seriesItemMeta">
  <dt>Erscheinungsverlauf</dt>
  Leerraum
  <dd>1815 - 1821</dd>
  Leerraum
  <i class="fa fa-info" data-toggle="tooltip" data-placement="right"
  title="" data-original-title="digitalisierte Ausgaben siehe Kalender">
   event
    ::before
  </i>
  <br>
  <dt>Erschienen</dt>
  Leerraum
  <dd class="publication-info">
    <span>Aarau: Sauerländer</span>
  </dd>
  <br>
  <dt>Verbreitungsort(e)</dt>
  Leerraum
  <dd>Aarau</dd>
  <br>
  <dt>ZDB-ID</dt>
  Leerraum
  <dd>2227885-0</dd>
</dl>
```

class = “seriesItemMeta” → Element enthält angeführte Metadaten pro Titel

ÖAW AUSTRIAN ACADEMY OF SCIENCES

Welche Informationen können wir aus dem dl-Element sammeln?

```
<dl class="seriesItemMeta">
  <dt>Erscheinungsverlauf</dt>
  Leerraum
  <dd>1815 – 1821</dd>
  Leerraum
  <i class="fa fa-info" data-toggle="tooltip" data-placement="right"
  title="" data-original-title="digitalisierte Ausgaben siehe Kalender">
   event
    ::before
  </i>
  <br>
  <dt>Erschienen</dt>
  Leerraum
  <dd class="publication-info">
    <span>Aarau: Sauerländer</span>
  </dd>
  <br>
  <dt>Verbreitungsort(e)</dt>
  Leerraum
  <dd>Aarau</dd>
  <br>
  <dt>ZDB-ID</dt>
  Leerraum
  <dd>2227885-0</dd>
</dl>
```

Informationen innerhalb der Metadatenliste folgen einheitlicher Struktur:

- Paare aus <dt> und <dd> (+ diverse Zusätze)
- <dt> → Art der Metainformation (z.B. *Erscheinungsverlauf*)
- <dd> → Information selbst (z.B. *1815 - 1821*)

→ diese und die vorherigen Muster können zur automatischen Extraktion der Daten genutzt werden

ÖAW AUSTRIAN ACADEMY OF SCIENCES

Jupyter Notebook: Webscraper für Zeitungsliste von digiPress

Herunterladen oder Klonen von Github-Repository:
https://github.com/nrastinger/metadatenanalyse-digipress

Öffnen von Notebook „1_Webscraper_Zeitungsliste _digiPress.ipynb“ in Google Colab: https://colab.google/