

Historische Forschung digital: Ein Workshop zum Deutschen Zeitungsportal

„Digitale Methoden der
Zeitungsanalyse“

Zentralbibliothek Zürich,
12. September 2024

Stephanie Nietsche,
Franziska Fuchs,
Michael Büchner,
und Lisa Landes



Einführung



„Historische Forschung digital: ein Workshop zum Deutschen Zeitungsportal“

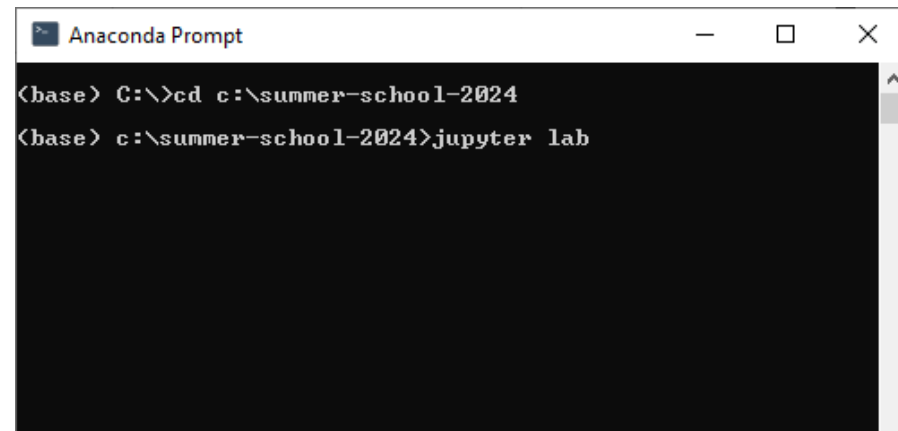
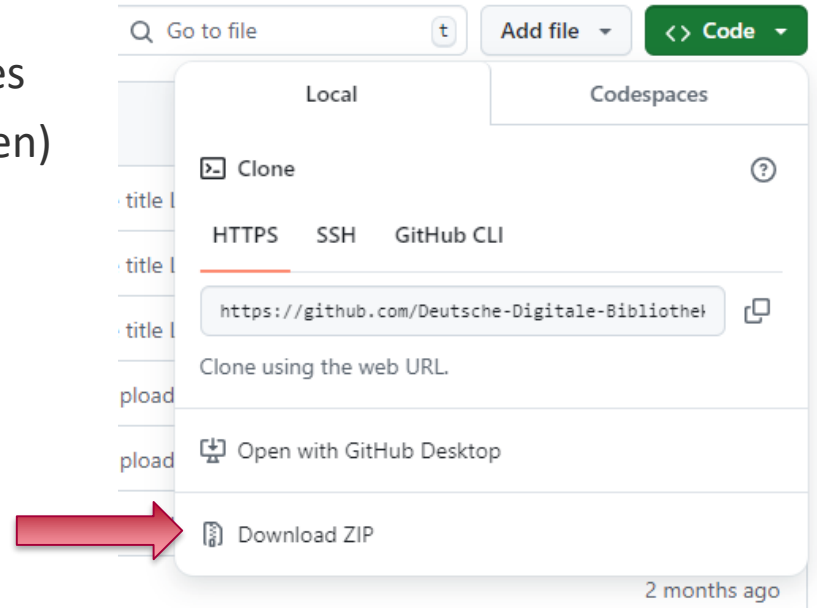
- Das **Deutsche Zeitungsportal** bietet über vier Millionen digitalisierte Zeitungsausgaben aus mehr als 1.800 Titeln mit Volltextsuche an.
- Der **Workshop** vermittelt, wie das Zeitungsportal für historische Forschung und die Digital Humanities genutzt werden kann.
- Teilnehmende lernen den Umgang mit der **API der Deutschen Digitalen Bibliothek**, um Daten aus dem Zeitungsportal herunterzuladen und zu analysieren.
- Der Workshop beinhaltet praktische **Datenanalysen mit Jupyter Notebooks und Python**, begleitet vom DNBLab-Team.

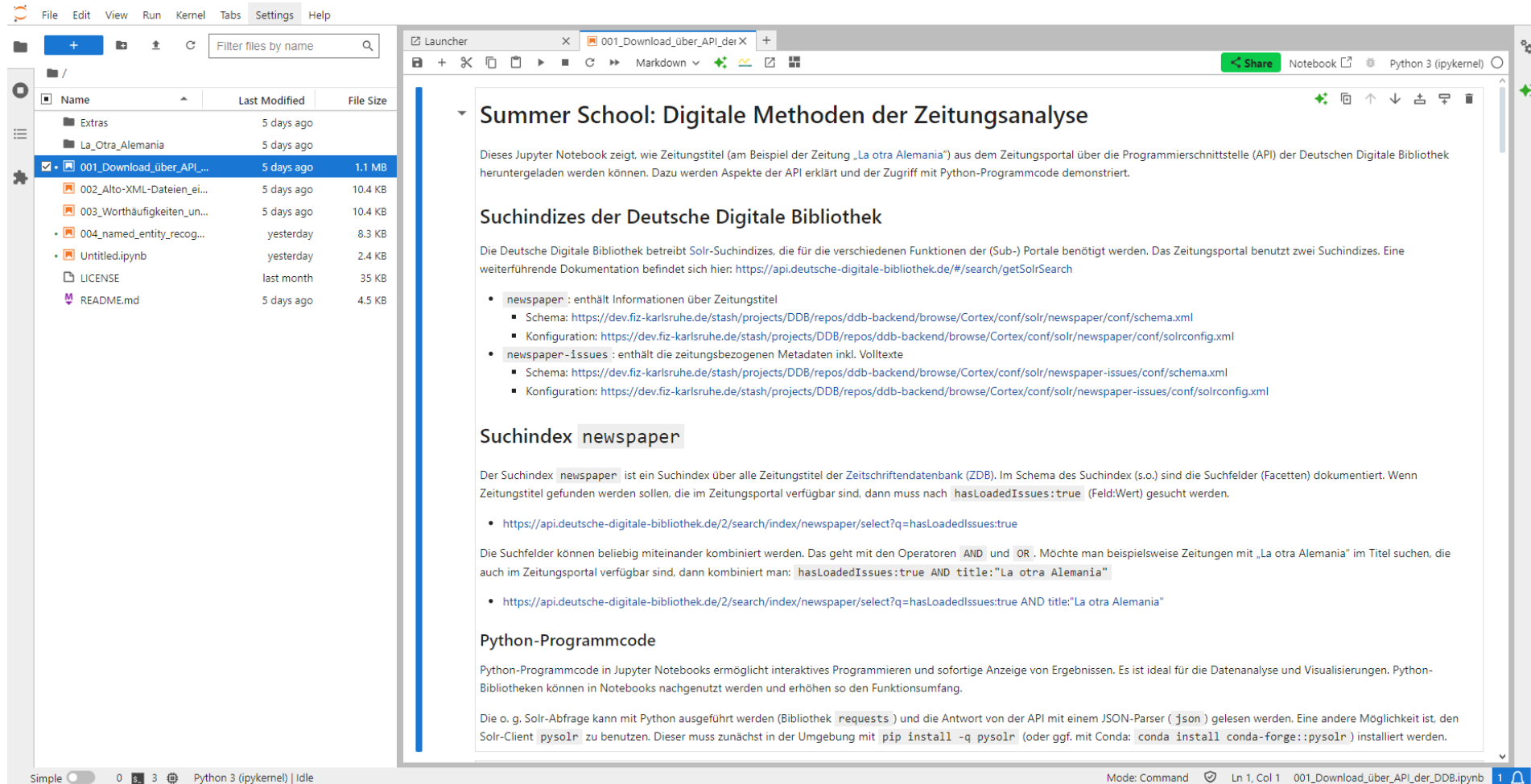
Übersicht

Uhrzeit	Programmpunkt	Notebook	Dozierende
13:30	Einführung ins „Deutsche Zeitungsportal“		Michael Büchner
13:45	Download der Zeitungsportaldaten über die API der Deutschen Digitalen Bibliothek	001_Download_über_API_der_DDB	Michael Büchner
14:45	Pause		
15:00	Einführung ins DNBLab		Franziska Fuchs
15:15	Datenanalyse <ul style="list-style-type: none">• Text aus ALTO-XML extrahieren• Worthäufigkeiten analysieren und visualisieren• Optional: Kurzer Einblick in Named Entity Recognition	002_Alto-XML-Dateien_einlesen_und_Texte_extrahieren 003_Worthäufigkeiten_und_Analyse 004_named_entity_recognition	Franziska Fuchs und Stephanie Nitsche
16:45	Fragen und Feedback		
17:00			

Vorbereitungen

1. Download der Dateien des GitHub-Repositories
 - Download über GitHub (oder Repro klonen)
 - Entpacken in lokales Dateisystem
z.B. in „C:\summer-school-2024\"
2. Starten von Jupyter Lab (unter Windows)
 - Start → „Anaconda Prompt“ starten
 - Eingabeaufforderung eingeben:
 - > cd c:\summer-school-2024
 - > jupyter lab
 - Alternative: Binder oder Google Colab
 - siehe README.md





The screenshot shows the Jupyter Lab interface. On the left is a file browser with a table of files:

Name	Last Modified	File Size
Extras	5 days ago	
La_Otra_Alemania	5 days ago	
001_Download_über_API...	5 days ago	1.1 MB
002_Alto-XML-Dateien_ei...	5 days ago	10.4 KB
003_Worthäufigkeiten_un...	5 days ago	10.4 KB
004_named_entity_recog...	yesterday	8.3 KB
Untitled.ipynb	yesterday	2.4 KB
LICENSE	last month	35 KB
README.md	5 days ago	4.5 KB

The main area displays a Jupyter Notebook titled "001_Download_über_API_der_X". The notebook content is as follows:

Summer School: Digitale Methoden der Zeitungsanalyse

Dieses Jupyter Notebook zeigt, wie Zeitungstitel (am Beispiel der Zeitung „La otra Alemania“) aus dem Zeitungsportal über die Programmierschnittstelle (API) der Deutschen Digitale Bibliothek heruntergeladen werden können. Dazu werden Aspekte der API erklärt und der Zugriff mit Python-Programmcode demonstriert.

Suchindizes der Deutsche Digitale Bibliothek

Die Deutsche Digitale Bibliothek betreibt Solr-Suchindizes, die für die verschiedenen Funktionen der (Sub-) Portale benötigt werden. Das Zeitungsportal benutzt zwei Suchindizes. Eine weiterführende Dokumentation befindet sich hier: <https://api.deutsche-digitale-bibliothek.de/#/search/getSolrSearch>

- newspaper**: enthält Informationen über Zeitungstitel
 - Schema: <https://dev.fiz-karlsruhe.de/stash/projects/DB/repos/ddb-backend/browse/Cortex/conf/solr/newspaper/conf/schema.xml>
 - Konfiguration: <https://dev.fiz-karlsruhe.de/stash/projects/DB/repos/ddb-backend/browse/Cortex/conf/solr/newspaper/conf/solrconfig.xml>
- newspaper-issues**: enthält die zeitungsbezogenen Metadaten inkl. Volltexte
 - Schema: <https://dev.fiz-karlsruhe.de/stash/projects/DB/repos/ddb-backend/browse/Cortex/conf/solr/newspaper-issues/conf/schema.xml>
 - Konfiguration: <https://dev.fiz-karlsruhe.de/stash/projects/DB/repos/ddb-backend/browse/Cortex/conf/solr/newspaper-issues/conf/solrconfig.xml>

Suchindex newspaper

Der Suchindex **newspaper** ist ein Suchindex über alle Zeitungstitel der **Zeitschriftendatenbank (ZDB)**. Im Schema des Suchindex (s.o.) sind die Suchfelder (Facetten) dokumentiert. Wenn Zeitungstitel gefunden werden sollen, die im Zeitungsportal verfügbar sind, dann muss nach `hasLoadedIssues:true` (Feld:Wert) gesucht werden.

- <https://api.deutsche-digitale-bibliothek.de/2/search/index/newspaper/select?q=hasLoadedIssues:true>

Die Suchfelder können beliebig miteinander kombiniert werden. Das geht mit den Operatoren **AND** und **OR**. Möchte man beispielsweise Zeitungen mit „La otra Alemania“ im Titel suchen, die auch im Zeitungsportal verfügbar sind, dann kombiniert man: `hasLoadedIssues:true AND title:"La otra Alemania"`

- [https://api.deutsche-digitale-bibliothek.de/2/search/index/newspaper/select?q=hasLoadedIssues:true AND title:"La otra Alemania"](https://api.deutsche-digitale-bibliothek.de/2/search/index/newspaper/select?q=hasLoadedIssues:true AND title:)

Python-Programmcode

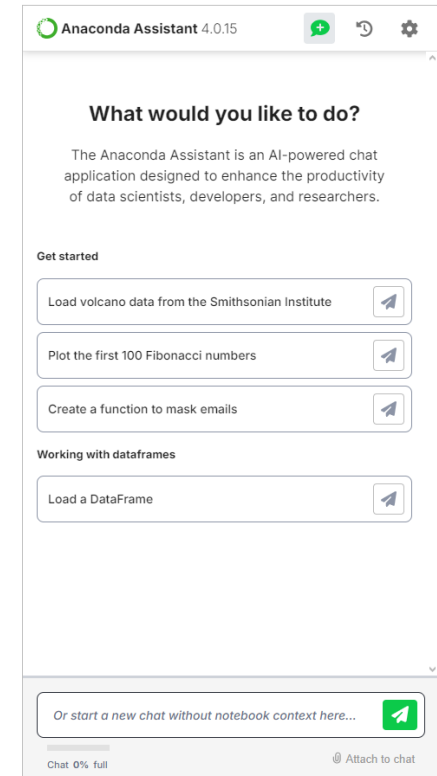
Python-Programmcode in Jupyter Notebooks ermöglicht interaktives Programmieren und sofortige Anzeige von Ergebnissen. Es ist ideal für die Datenanalyse und Visualisierungen. Python-Bibliotheken können in Notebooks nachgenutzt werden und erhöhen so den Funktionsumfang.

Die o. g. Solr-Abfrage kann mit Python ausgeführt werden (Bibliothek `requests`) und die Antwort von der API mit einem JSON-Parser (`json`) gelesen werden. Eine andere Möglichkeit ist, den Solr-Client `pyso1r` zu benutzen. Dieser muss zunächst in der Umgebung mit `pip install -q pyso1r` (oder ggf. mit Conda: `conda install conda-forge::pyso1r`) installiert werden.

At the bottom of the Jupyter Lab interface, the status bar shows: Simple, 0, 3, Python 3 (ipykernel) | Idle, Mode: Command, Ln 1, Col 1, 001_Download_über_API_der_DDB.ipynb, 1.

Verwendung von KI

- Download der Zeitungsportaldaten: wir benutzen Large Language Models (**KI**), um Python-Code zu erzeugen
 - experimentell, aber funktioniert erstaunlich gut! :-)
- ChatGPT: <https://chatgpt.com/>
- Perplexity AI: <https://www.perplexity.ai/>
- Anaconda Assistant in JupyterLab
- Jupyter-AI: <https://github.com/jupyterlab/jupyter-ai>
- uvm.



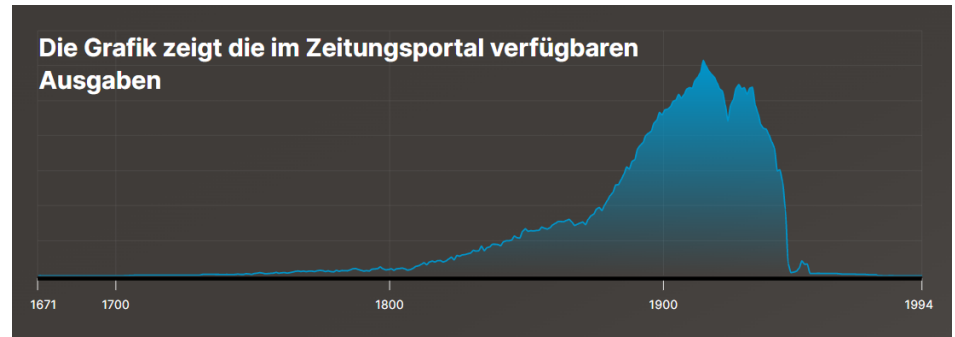
Einführung ins „Deutsche Zeitungsportal“



Inhalte des Deutschen Zeitungsportals

Inhalte September 2024

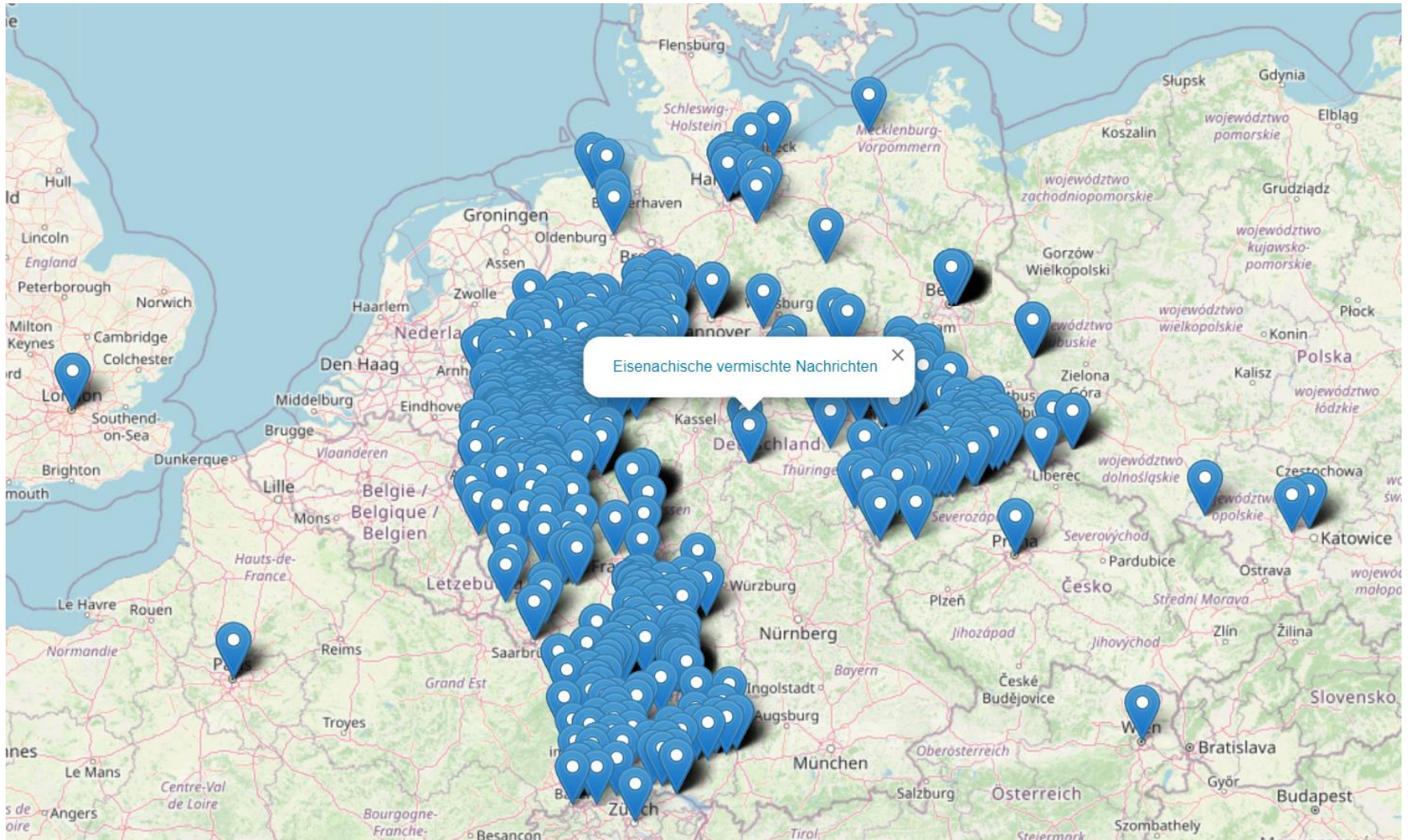
- 4,2 Mio. Zeitungsausgaben
- > 26 Mio. Seiten
- Davon ca. 90% mit Volltexten
- >1.800 Zeitungen (ZDB-IDs)
- Ausgaben aus den Jahren 1671-1994



Aktuell 22 Datenpartner (nach Größe sortiert)

- | | |
|-------------------------------------|---|
| – ULB Bonn (zeit.punktNRW) | – MARCHIVUM Mannheim |
| – ULB Münster (zeit.punktNRW) | – UB Hannover |
| – SLUB Dresden | – Bibliothek der Friedrich-Ebert-Stiftung |
| – Badische Landesbibliothek | – UB Heidelberg |
| – Württembergische Landesbibliothek | – LB Oldenburg |
| – SUB Hamburg | – UB Mannheim |
| – ULB Halle | – Stadtarchiv Viernheim |
| – SBB Berlin | – UB der TU Chemnitz |
| – Kreisarchiv Biberach | – Deutsches Exilarchiv |
| – UB Gießen | – Stiftung Haus Oberschlesien |
| – UB Düsseldorf (zeit.punktNRW) | – Schifffahrtsmuseum Rostock |

Zeitungen im Deutschen Zeitungsportal nach Verbreitungsort



Stand: 18.07.2024



Geschichte aus erster Hand

Entdecken Sie historische Zeitungen aus den Jahren 1671 bis 1994

Suche in historischen Zeitungen



Zeitung über Titel
auswählen



Zeitung über Ort auswählen



Zeitung über Jahr
auswählen





von 31 8 1872
bis 10 2 1951

Anwenden

Zeitung

+ Börsenblatt für den deutschen Buchhandel, bbb ; Fachzeitschr. für Verlagswesen u. Buchhandel 33

+ Kölnische Zeitung, mit Wirtschafts- und Handelsblatt 14

+ Schwäbischer Merkur, mit Schwäbischer Kronik und Handelszeitung : Süddeutsche Zeitung 6

+ Dresdner Nachrichten 6

Alle anzeigen

Verbreitungsort

+ Leipzig 42

+ Köln 16

+ Stuttgart 16


+ Jülich 14

Alle anzeigen

Datengeber

+ Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden 56

164 Ergebnisse

 Suche speichern

SORTIEREN NACH

Relevanz

ERGEBNISSE PRO SEITE

20

< 1 2 3 4 5 >



Börsenblatt für den deutschen Buchhandel : bbb ; Fachzeitschr. für Verlagswesen u. Buchhandel

Dienstag, 30.08.1910

Ausstellung weiterer seltener Original-Drucke und Handzeichnungen Klingers angliedert. Zentralbibliothek in Zürich. — Der »Anzeiger für den Schweizerischen Buchhandel« teilt folgendes mit: Unter Vorbehalt der Genehmigung

> 4 Treffer in dieser Ausgabe



Dorstener Volkszeitung und Wochenblatt. 1910-1919

Dienstag, 01.06.1915

begonnen . Hindernisse der Italiener in Friaul . * Zürich , 31 . Mai . Wie man hier erfährt , werden die Operationen der Italiener im Gebiet von Friaul durch Hochwasser der Flüsse sehr stark behindert . Die schon

> 6 Treffer in dieser Ausgabe



Kölnische Zeitung. 1803-1945

Samstag, 29.10.1927

. (Zürich . Jüngst hatte ich in der Zentralbibliothek etwas z erledigen und geriet daher auch in den Raum , der dem Andenken Gottfried Kellers geweiht ist . Als bald wurde dort mein Blick durch ein Gemälde des Meisters

> 4 Treffer in dieser Ausgabe



Berliner Tageblatt und Handels-Zeitung, Morgen-Ausgabe

Sonntag, 06.05.1917

Zentralbibliothek. Unser Korrespondent schreibt uns: DaS geistige Leben Zürichs hat mitten im Krieg« einen starken Schwung empfangen, denn Anfang Mai wird die Zentralbibliothek der allgemeinen Benutzung übergeben. Man hat zum

> 2 Treffer in dieser Ausgabe



Kölnische Zeitung. 1803-1945

Samstag, 25.09.1915

Ausbildung unter die Fahne gerufen würde . WTB Zürich , 25 . Sept . (Telegr .) Dem Vernehmen nach ist zum Präsidenten der sin Nr . 977 ausführlich gekennzeichnet neten] schweizerischen



Volltextanzeige Als Pop-up öffnen

werde da? Militärkabinett als der Schrecken des Ossizierkops angesehen. Ter Kriegsm insier sei vielfach mir Prügelknabe des preußischen Militärkabinett 8. Eine reinliche Abgrenzung der Zuständigkeit sei nötig. DaS Militärkabinett müsse im Interesse deS Heeres unter dem Krieg-minister

Die Züricher Zentralbibliothek. Unser Korrespondent schreibt uns: Das geistige Leben Zürichs hat mitten im Kriege einen starken Schwung empfangen, denn Anfang Mai wird die Zentralbibliothek der allgemeinen Benutzung übergeben. Man hat zum Beginn der schweren Kriegszeit den Bau dieses Bücherhauses begonnen, und man hat ihn heute mit großer Energie, Ueberlegtheit und wohlangebrachter Rücksicht auf alle modernen Interessen vollendet. So wurden alle bisher über die Stadt zerstreuten Bücherschätze unter ein Dach gebracht. Die Sammlungen der naturwissenschaftlichen und juristischen Gesellschaften, die Stadt- und Universitätsbibliothek und andere kleinere Bestände wurden vereinigt. Der Zentralkatalog, eine vielbewunderte Schöpfung der Züricher Bibliographen, ist nun wirklich der Führer durch ein einziges Haus geworden. Bei dem Umzug werden natürlich die Erinnerungen an das verlassene Heim der Stadtbibliothek aufgefrischt. Sie hat drei, hundert Jahre lang in der sogenannten Wasserflurkirche gestanden. Sie ist ein Märtyrerheiligtum mit wunderbaren Quellen

Berliner Tageblatt und Handels-Zeitung, Morgen-Ausgabe



Sonntag, 06.05.1917

Seite 3/48

Zurück zur Ergebnisübersicht




mir Prügelknabe des preußischen Militärkabinett 8. Eine reinliche Abgrenzung der Zuständigkeit sei nötig. Das Militärkabinett müsse im Interesse des Heeres unter dem Kriegsminister

Die Züricher Zentralbibliothek. Unser Korrespondent schreibt uns: Das geistige Leben Zürichs hat mitten im Kriege einen starken Schwung empfangen, denn Anfang Mai wird die Zentralbibliothek der allgemeinen Benutzung übergeben. Man hat zum Beginn der schweren Kriegszeit den Bau dieses Bücherhauses begonnen, und man hat ihn heute mit großer Energie, Ueberlegtheit und wohlangebrachter Rücksicht auf alle modernen Interessen vollendet. So wurden alle bisher über die Stadt zerstreuten Bücherschätze unter ein Dach gebracht. Die Sammlungen der naturwissenschaftlichen und juristischen Gesellschaften, die Stadt- und Universitätsbibliothek und andere kleinere Bestände wurden vereinigt. Der Zentralkatalog, eine vielbewunderte Schöpfung der Züricher Bibliographen, ist nun wirklich der Führer durch ein einziges Haus geworden. Bei dem Umzug werden natürlich die Erinnerungen an das verlassene Heim der Stadtbibliothek aufgefrischt. Sie hat drei, hundert Jahre lang in der sogenannten Wasserflurkirche gestanden. Sie ist ein Märtyrerheiligtum mit wunderbaren Quellen

Berliner Tageblatt und Handels-Zeitung

Großmünster Zürich


ZEITSCHRIFTEN-DATENBANK

Verbreitungsort

Berlin

Erscheinungsfrequenz

täglich

Erscheinungsverlauf

1872, Nr. 1 (1. Januar 1872)-61. Jahrgang, Nr. 52 (31. Januar 1932)

Sprache

Deutsch


ZDB-ID

2764651-8

Weiterführende Informationen zu dieser Zeitung finden Sie in der Zeitschriftenbibliothek:

[Gedruckte Exemplare der Zeitung in deutschen Bibliotheken](#)

[Vorläufer und Nachfolger der Zeitung](#)


Voltext

Verfügbare Ausgaben

19. Jahrhundert										20. Jahrhundert									
1800	1801	1802	1803	1804	1805	1806	1807	1808	1809	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909
1810	1811	1812	1813	1814	1815	1816	1817	1818	1819	1910	1911	1912	1913	1914	1915	1916	1917	1918	1919
1820	1821	1822	1823	1824	1825	1826	1827	1828	1829	1920	1921	1922	1923	1924	1925	1926	1927	1928	1929
1830	1831	1832	1833	1834	1835	1836	1837	1838	1839	1930	1931	1932	1933	1934	1935	1936	1937	1938	1939
1840	1841	1842	1843	1844	1845	1846	1847	1848	1849	1940	1941	1942	1943	1944	1945	1946	1947	1948	1949
1850	1851	1852	1853	1854	1855	1856	1857	1858	1859	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959
1860	1861	1862	1863	1864	1865	1866	1867	1868	1869	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
1870	1871	1872	1873	1874	1875	1876	1877	1878	1879	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
1880	1881	1882	1883	1884	1885	1886	1887	1888	1889	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
1890	1891	1892	1893	1894	1895	1896	1897	1898	1899	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999

Vorgeschlagene Ausgaben dieser Zeitung (12)



AUSGABE VOM
26. Oktober 1892



AUSGABE VOM
03. Januar 1884



AUSGABE VOM
19. März 1905



AUSGABE VOM
30. Dezember 1920

API der Deutschen Digitalen Bibliothek



Einführung „DDB-API“



- API ist **öffentlich** und (bald) vollständig ohne API-Key nutzbar
 - **keine Download-Beschränkungen**
 - Metadaten sind **CC0-lizenziert** (Beschreibungstexte manchmal nicht)
 - digitale Objekte haben eine individuelle Lizenz: siehe Lizenzkorb
<http://www.deutsche-digitale-bibliothek.de/content/ueber-uns/lizenzen-und-rechtehinweise-der-lizenzkorb-der-deutschen-digitalen-bibliothek/>
- API ist **OpenAPI**-dokumentiert
<https://api.deutsche-digitale-bibliothek.de/OpenAPI>
 - Solr-Suchindizes
 - Methoden, um auf Daten zuzugreifen
 - Bilddaten in der DDB: IIIF Image API 2.0
- Achtung: **dezentrale Datenhaltung**, d.h. digitale Objekte (z.B. Zeitungsseiten) liegen i. d. R. auf den Servern unserer Datenpartner

<https://api.deutsche-digitale-bibliothek.de>

Einführung „Zeitungsportal“



- **Zeitungstitel (title)**
 - jede Zeitung hat einen Titel und eine eindeutige ID
 - Erfassung in der Zeitschriftendatenbank (ZDB): <https://zdb-katalog.de/>
- **Ausgaben (issue)**
 - ein Zeitungstitel hat ein oder mehrere Ausgaben (auch mehrere Ausgaben an einem Tag)
- **Seiten (page)**
 - eine Ausgabe hat eine oder mehrere Seiten
 - Seiten haben i.d.R. OCR-Volltext (und ein Digitalisat)
- Datenpartner liefern **METS/MODS**-Metadaten
 - eine METS/MODS-Datei beschreibt i.d.R. eine Ausgabe eines Zeitungstitel
 - eine METS/MODS-Datei ist ein **DDB-Objekt** und hat eine DDB-ID
 - (für Zeitungstitel kann es aber auch eine METS/MODS-Datei geben)

Suchindex „newspaper“



- Suchindex über alle **Zeitungstitel** der Zeitschriftendatenbank (ZDB)
- Technologie: **Apache Solr**
 - siehe https://solr.apache.org/guide/8_8/searching.html
 - **Facetten**: Im Schema des Suchindex sind die verschiedenen Suchfelder dokumentiert, die für spezifische Abfragen genutzt werden können.
- **Verfügbarkeit**: Um Zeitungen zu finden, die im Deutschen Zeitungsportal verfügbar sind, muss nach dem Wert „hasLoadedIssues:true“ gesucht werden.
- **kombinierte Abfragen**: Mehrere Suchkriterien können mit den Operatoren AND und OR kombiniert werden, z. B. um eine spezifische Zeitung wie „La otra Alemania“ zu finden.

Probieren wir's aus... (1)

- siehe Abschnitt „Suchindex newspaper“ im Jupyter Notebook
 - Link auf **Schema** und Konfiguration
- Suche nach „La otra Alemania“
 - <https://api.deutsche-digitale-bibliothek.de/2/search/index/newspaper/select?q=hasLoadedIssues:true AND title:"La otra Alemania">
- Einlesen mit Pythen und Ergebnis ausgeben...

KI-Prompt: Ich möchte in Python mit dem Solr-Client pysolr auf den Endpunkt <https://api.deutsche-digitale-bibliothek.de/2/search/index/newspaper> zugreifen. Kannst Du mir einen Python-Code erstellen, der im Feld location nach „Buenos Aires“ sucht und auch hasLoadedIssues auf „wahr“ setzt. Gibt bitte id, title, location, frequency und progress für jeden Suchtreffer aus.

Suchindex „newspaper-issues“



- Suchindex über alle **Ausgaben** (type:issue) und alle **Seiten** (type:page)
- Suche nur in einem Zeitungstitel mit „zdb_id:ZDB-ID “
(z.B. „zdb_id:2149754-0“)

Probieren wir's aus... (2)

- siehe Abschnitt „Suchindex newspaper-issues“ im Jupyter Notebook
- Suche nach Ausgaben (type:issue) von „La otra Alemania“ (zdb_id:2149754-0)
 - https://api.deutsche-digitale-bibliothek.de/2/search/index/newspaper-issues/select?q=zdb_id:2149754-0 AND type:issue

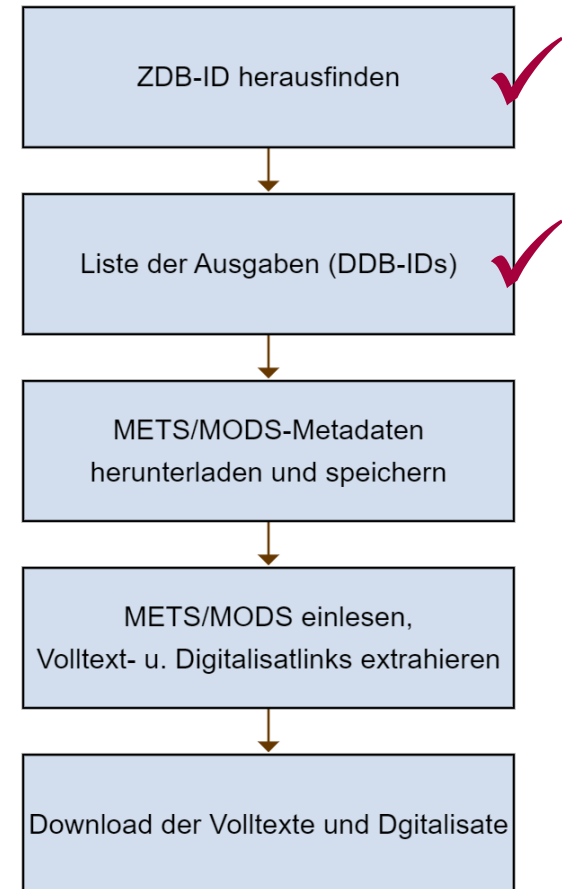
KI-Prompt: Schreibe ein Python-Code, der mithilfe der pysolr-Bibliothek eine Suche in einem Solr-Index durchführt und die Ergebnisse in ein Pandas DataFrame überführt. Der Solr-Index ist über die URL <https://api.deutsche-digitale-bibliothek.de/2/search/index/newspaper-issues> erreichbar. Die Suchabfrage soll nach Dokumenten mit der zdb_id „2149754-0“ und dem type „issue“ suchen und bis zu 1000 Ergebnisse zurückgeben. Anschließend sollen die Ergebnisse in ein Pandas DataFrame überführt und angezeigt werden.

- **Pandas:** Bibliothek in Python, die Datenstrukturen und Funktionen bereitstellt, um Daten zu analysieren, zu manipulieren und zu verarbeiten, insbesondere in Form von DataFrames (tabellarische Daten).

Download eines Zeitungstitels: Ablauf

Wie lädt man denn nun einen gesamten Zeitungstitel herunter?

1. **ZDB-ID** des gewünschten Zeitungstitels sollte bekannt sein
 - Suchindex „newspaper“ → „2149754-0“ für „La otra Alemania“
2. Erstellung einer **Liste der Ausgaben** (mit DDB-IDs)
 - Suchindex „newspaper-issues“
→ Suche nach „zdb_id:2149754-0 AND type:issue“
3. Download der **METS/MODS-Metadaten** aller Ausgaben
4. Extraktion der Links zu **Volltexten und Digitalisaten**
5. **Download** und Abspeichern der Volltexte und Digitalisate



Probieren wir's aus... (3)

- siehe Abschnitt ab „Download der METS/MODS-Daten “ im Jupyter Notebook

KI-Prompt: Erstelle mit Python ein Verzeichnis La_Otra_Alemania, in dem heruntergeladene XML-Dateien gespeichert werden können. Iteriere durch jede Zeile des bestehenden DataFrames df, der die Spalten id und publication_date enthält. Für jede Zeile:

1. Extrahiere den Wert der Spalte id.
2. Formatiere den ISO-DateTime-Wert der Spalte publication_date im Format YYYY-MM-DD.
3. Generiere eine URL <https://api.deutsche-digitale-bibliothek.de/2/items/{id}/source/record> zur API-Abfrage
4. Setze die HTTP-Header so, dass die Antwort im XML-Format akzeptiert wird.
5. Erstelle einen Dateipfad für die XML-Datei im Format {publication_date}_{id}.xml und speichere sie im erstellten Verzeichnis.

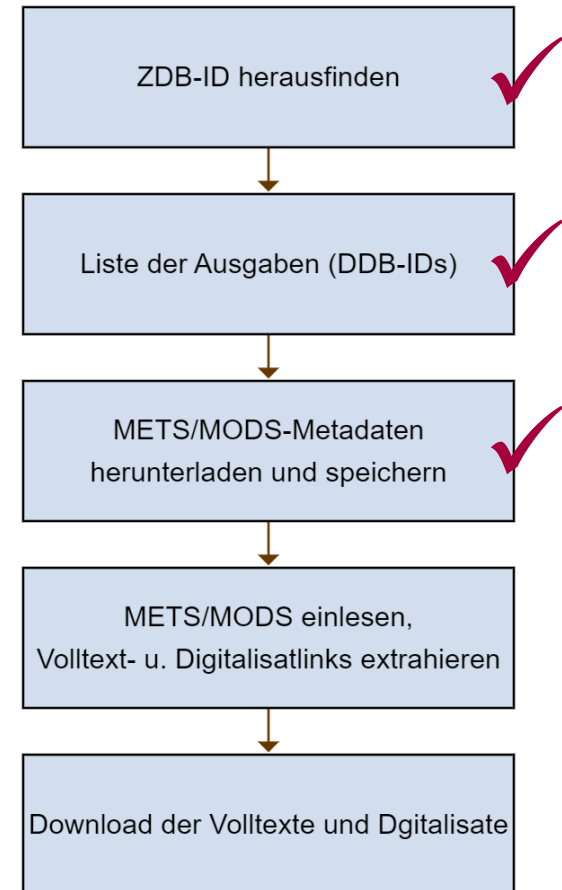
Extraktion der Links

- **METS/MODS**

- **METS:** Ein XML-Standard zur Strukturierung und Übermittlung digitaler Objekte in Archiven und Bibliotheken
- **MODS:** Ein XML-Format zur Beschreibung bibliografischer Daten, oft als Ergänzung zu METS genutzt

- **Extraktion** von Daten aus XML

- **XPath:** Eine Abfragesprache, die zum Navigieren und Auswählen von Elementen in XML-Dokumenten verwendet wird
- **Pfad-Ausdrücke:** XPath ermöglicht es, Teile eines XML-Dokuments mithilfe von Pfaden präzise zu lokalisieren und zu extrahieren
- siehe <https://en.wikipedia.org/wiki/XPath>




```
<mets:mets xmlns:mets="http://www.loc.gov/METS/" xmlns:mods="http://www.loc.gov/mods/v3">
```

```
<!-- Header -->
```

```
<mets:metsHdr CREATEDATE="2024-07-30T12:00:00">
```

```
<!-- Agent Information -->
```

```
</mets:metsHdr>
```

```
<!-- Descriptive Metadata -->
```

```
<mets:dmdSec ID="dmd001">
```

```
<mets:mdWrap MDTYPE="MODS">
```

```
<mets:xmlData>
```

```
<mods:mods>
```

```
<!-- MODS Metadata -->
```

```
</mods:mods>
```

```
</mets:xmlData>
```

```
</mets:mdWrap>
```

```
</mets:dmdSec>
```

```
<!-- File Section -->
```

```
<mets:fileSec>
```

```
<mets:fileGrp USE="DEFAULT">
```

```
<mets:file ID="file0001" MIMETYPE="image/jpeg">
```

```
<mets:FLocat LOCTYPE="URL" xlink:href="http://example.com/image1.jpg"/>
```

```
</mets:file>
```

```
</mets:fileGrp>
```

```
<mets:fileGrp USE="DDB_FULLTEXT">
```

```
<mets:file ID="file0002" MIMETYPE="text/xml">
```

```
<mets:FLocat LOCTYPE="URL" xlink:href="http://example.com/fulltext1.xml"/>
```

```
</mets:file>
```

```
</mets:fileGrp>
```

```
</mets:fileSec>
```

//mets:mets/mets:fileSec/mets:fileGrp[@USE="DEFAULT"]/mets:file/mets:FLocat/@xlink:href

```
<mets:structMap TYPE="logical">
```

```
<mets:div TYPE="document" DMDID="dmd001">
```

```
<mets:div TYPE="page" ORDER="1">
```

```
<mets:fptr FILEID="file0001"/>
```

```
</mets:div>
```

```
<mets:div TYPE="page" ORDER="2">
```

```
<mets:fptr FILEID="file0002"/>
```

```
</mets:div>
```

```
</mets:div>
```

```
</mets:structMap>
```

```
</mets:mets>
```

Probieren wir's aus... (4)

- siehe Abschnitt ab „Download der Bild- und Volltextdaten“ im Jupyter Notebook

KI-Prompt: Erstelle ein möglichst einfaches Python-Skript, das alle METS/MODS-Dateien in dem Verzeichnis „La_Otra_Alemania“ einliest und URLs mittels des XPath-Ausdrucks `//mets:mets/mets:fileSec/mets:fileGrp[@USE="DDB_FULLTEXT"]/mets:file/mets:FLocat/@xlink:href` extrahiert. Die extrahierten URLs sollen heruntergeladen und in Unterverzeichnissen gespeichert werden. Die Unterverzeichnisse werden nach den Dateinamen der METS/MODS-Dateien benannt. Die heruntergeladenen XML-Dateien sollen mit 1 beginnend durchnummeriert werden.

Das Gleiche soll mit JPEG-Dateien und dem XPath-Ausdruck `//mets:mets/mets:fileSec/mets:fileGrp[@USE="DEFAULT"]/mets:file/mets:FLocat/@xlink:href` gemacht werden.