

# Leveraging Zurich Zentralbibliothek's Jupyter Notebooks for Metadata Retrieval and Analysis from Alma

**Samsinger, Linda**

[linda.samsinger@zb.uzh.ch](mailto:linda.samsinger@zb.uzh.ch)

Zentralbibliothek Zürich, Schweiz

ORCID: 0009-0004-6133-8790

In the realm of research, digital tools have become essential for accessing relevant literature metadata (Döring et al. 2022; Padilla 2016). Hence, Zurich Zentralbibliothek has funded the KoLibri project, following the example set by the DNB (Taube 2023), ÖNB (Kaiser 2023, 207), and other libraries (Ames and Havens 2021, 50; Yesmin and Zabed 2016) by offering a set of seven Jupyter Notebooks as digital tools designed for metadata retrieval and analysis using Python code. The zeroth Jupyter Notebook serves as the foundational Python codebase for the other six. The next two notebooks allow users to conduct both simple and advanced searches across the library catalog, while the third notebook provides statistical analysis of the search results from the previous two, generating scientific visualizations. The remaining three notebooks mirror the functionality of the first three, with the added feature of GND/Wikidata data field enrichment for the search results. Metadata retrieval and analysis can be daunting without user-friendly tools and formats for conversion, organization, and analysis. Therefore, Zurich Zentralbibliothek has developed these high-quality Jupyter Notebooks (Candela 2023, 1550) available in German. They are designed to be intuitive, transparent and customizable, integrating with powerful Python libraries for statistical analysis and data visualization.

The Jupyter Notebooks operate by executing code blocks that search and harvest library data from the Swiss Library Service Platform (SLSP) online catalog in MARC XML format, accessed through an SRU interface – a procedure similar to using the Alma API (Hyams and Pilko 2022, 1). The SLSP-Alma database is a comprehensive library catalog, aggregating data from a nationwide network of Swiss libraries, boasting over 25 million media items. This catalog is a treasure trove of resources for scholars, researchers, and enthusiasts alike, surpassing the main Swisscovery website, which includes over six million media items and is the default search platform for many library users. The intuitive nature of these Jupyter Notebooks (Boscoe et al. 2017, 1) facilitates tapping into a more extensive repository of catalog data, as the notebooks are renowned for their seamless integration of code alongside explanatory rich text. While users may require modest to advanced programming skills

to fully utilize these notebooks, they cater to a broad audience of tech-savvy researchers eager to explore the rich data offered by the SLSP-Alma catalog. By utilizing these Jupyter Notebooks, researchers gain improved capabilities in building topic-specific corpora (Oberbichler and Pfanzelter 2021, 4), visualizing data, conducting literature reviews and compiling bibliographies, as compared to traditional methods of searching via the Swisscovery website.

Based on the user's search prompt in the Jupyter Notebooks, media records of books, articles, and journals can be harvested and exported into versatile formats like Excel, JSON or CSV. This functionality empowers digital humanities researchers to efficiently manage metadata using familiar spreadsheet formats. The data is structured in a clear, cleaned and straightforward manner, facilitating easy filtering, sorting and calculation of search results within tables. The extracted metadata includes essential bibliographic fields like title, author, publisher, publication year, and location, paramount for reference generation. Additionally, derived fields, which were purposefully created in-house, such as epoch and resource type provide deeper context. Links to the table of contents allow for local bulk downloads of books' table of contents in PDF format. Optional enrichment with Wikidata and GND reconciliation enhances insights derived from the amassed data.

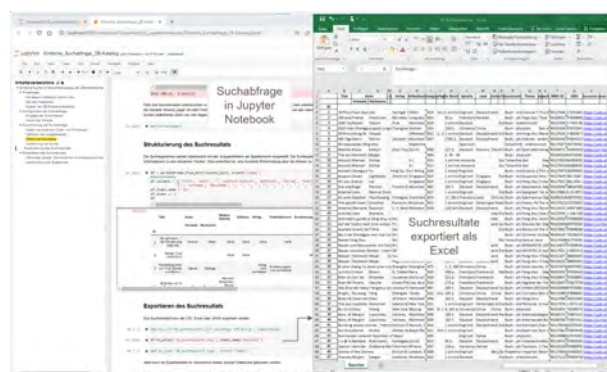


Figure 1: The initial Jupyter Notebook, titled 'Simple Search' (left), alongside its exported search results displayed in Excel (right).

Beyond mere retrieval, the notebooks offer robust capabilities for metadata analysis, leveraging natural language processing (NLP) techniques (Jentsch and Porada 2021, 89). NLP libraries enable in-depth analysis of textual metadata, including curated keywords of titles visualized as word clouds centered around the initial search term. Frequency and correlation across all bibliographic fields are available in bar, pie and line charts, aiding in trend and distribution interpretation. Moreover, the notebooks include features for similarity analysis, semantic visualization, and world map views, unveiling inherent insights and patterns within the dataset. Finally, tables and figures are exportable as PDFs, enhancing the accessibility and shareability of analytical findings for dissemination in reports, presentations, or scholarly articles. Reviewing these results allows users to discern thematic trends, popular authors,

temporal patterns and genre preferences within the library holdings. For instance, the spreadsheet search results for the query “Heidi Johanna Spyri” can be imported into a notebook for detailed data analysis. The analysis reveals that the most frequent publishing location is Zurich, with “Hemma” being the most common publisher. The majority of the media are physical books written in German, primarily published between 2010 and 2020. Interestingly, a small fraction (1.68%) of the media consists of physical and electronic audio recordings and computer files. The geographical focus of these works is predominantly on Switzerland and the Alps, with some references to Japan and the USA. Key terms associated with this search include “grandfather,” “orphan,” “alpine pasture,” “German” and “friendship,” which are illustrated in a word cloud. Such insights are valuable to various stakeholders, such as researchers, especially in computational linguistics and digital humanities, librarians involved in collection development and enthusiasts exploring literary landscapes (Herrmann et al. 2021, 2).

Furthermore, the flexibility and transparency of Jupyter Notebooks empower users to clone, customize, and fine-tune code snippets to meet specific functional objectives. With just a few lines of code, users can tailor visualizations to highlight trends, anomalies or correlations within the SLSP catalog metadata. As these notebooks are open-source, programming researchers and librarians have the freedom to fork and extend the source code, adding their own unique database APIs and features. The notebooks are freely accessible in both cloud (browser and server-based) and offline versions on GitLab at <https://data.zb.uzh.ch/map/books/data-map-der-zentralbibliothek-zurich/page/jupyter-notebooks-der-zentralbibliothek-zurich>.

In conclusion, these notebooks enable quantitative research and evaluation of a topic’s media metadata, available in the bounteous online catalog of Swiss libraries. Their functionalities are both innovative and user-friendly compared to alternatives. This solution not only promotes data-driven digital scholarship but also equips users with an adaptable toolkit to unlock the wealth of knowledge contained within the Swiss-wide library catalog.

## Bibliographie

- Ames, Sarah, and Lucy Havens.** 2021. "Exploring National Library of Scotland Datasets With Jupyter Notebooks." *IFLA Journal* 48 (1): 50-56.
- Boscoe, Gernadette M., Irene V. Pasquetto, Milena S. Golshan, and Christine L. Borgman.** 2017. "Using the Jupyter Notebook as a Tool for Open Science: An Empirical Study." *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 1-2.
- Candela, Gustavo.** 2023. "An Approach to Assess the Quality of Jupyter Projects Published by GLAM Institutions." *Journal of the Association for Information Science and Technology* 74 (13): 1550-1564.
- Döring, Karoline Dominika, Mareike König, Stefan Haas, and Jörg Wettlaufer.** 2022. *Digital History: Konzepte, Methoden und Kritiken Digitaler Geschichtswissenschaft*. Berlin: De Gruyter Oldenbourg.
- Herrmann, Berenike J., Giulia Grisot, Susanne Gubser, and Elias Kreyenbühl.** 2021. "Ein großer Berg Daten? Zur bibliothekswissenschaftlichen Dimension des korpusliteraturwissenschaftlichen Digital Humanities-Projekts ‘High Mountains – Deutschschweizer Erzählliteratur 1880–1930’." *027.7 Journal for Library Culture* 8 (1): 1-26.
- Hyams, Rebecca, and Tamara Pilko.** 2022. "'You could use the API!': A Crash Course in Working with the Alma APIs using Postman." *code4lib* 54: 1.
- Jentsch, Patrick, and Stephan Porada.** 2021. "From Text to Data: Digitization, Text Analysis and Corpus Linguistics." In *Digital Methods in the Humanities: Challenges, Ideas, Perspectives*, edited by Silke Schwandt, 89-128. Bielefeld: transcript Verlag.
- Kaiser, Max.** 2023. "Digitale Sammlungen als offene Daten für die Forschung: Strategische Zielsetzungen der Österreichischen Nationalbibliothek." *Bibliothek Forschung und Praxis* 47 (2): 200-212.
- Oberbichler, Sarah, and Eva Pfanzelter.** 2021. "Topic-specific Corpus Building: A Step Towards a Representative Newspaper Corpus on the Topic of Return Migration Using Text Mining Methods." *Journal of Digital History* (1): 74-98.
- Padilla, Thomas.** 2016. "Humanities Data in the Library: Integrity, Form, Access." *D-Lib Magazine* 22 (3/4).
- Taube, Anke.** 2023. *Open DNB Lab - Ein praktischer Einstieg in den Bezug und die Analyse der Daten und freien digitalen Objekte der DNB*. OPUS.
- Yesmin, Sabina, and Ahmed Zabed.** 2016. "Preference of Bangladesh University Students for Searching the Library Catalogue." *The Electronic Library* 34 (4): 683-695.