Swiss Google Books for Research

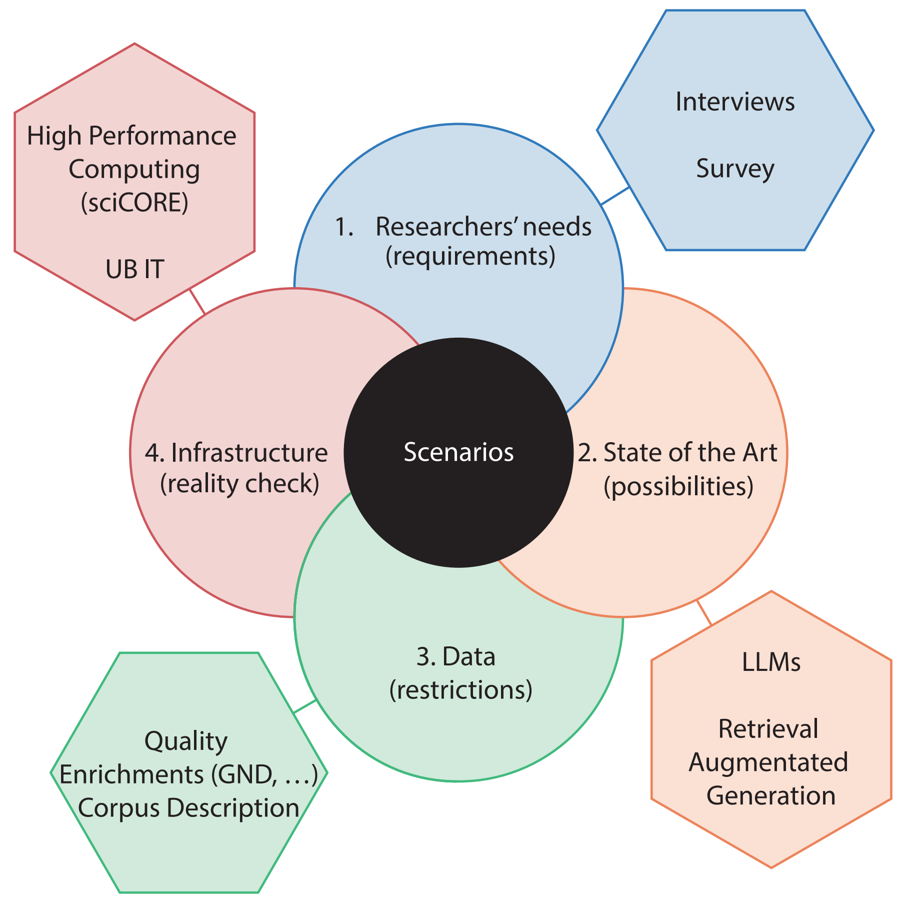
How to provide 90 million fulltext pages?

Martin Reisacher, Eric Dubey, Matteo Lorenzini
Universitätsbibliothek, Universität Basel

1 (Pre-)Project outline

The UB Bern, ZHB Luzern, ZB Zürich and UB Basel are digitizing with Google Books a big part of their **monographs from the 18th and 19th century**. With its nearly 300.000 volumes and over 90 million full text pages it is one of the biggest digitized corpus for Swiss researchers and offers great potential for data driven research and teaching.

Four working groups are examining how to distribute this dataset through an infrastructure that's not only versatile enough to cater to diverse research interests, but also easy to maintain. The aim is to provide different scenarios ranging from simple data dumps to a complex TDM environment and to decide on one of them and implement it in a follow-up project.



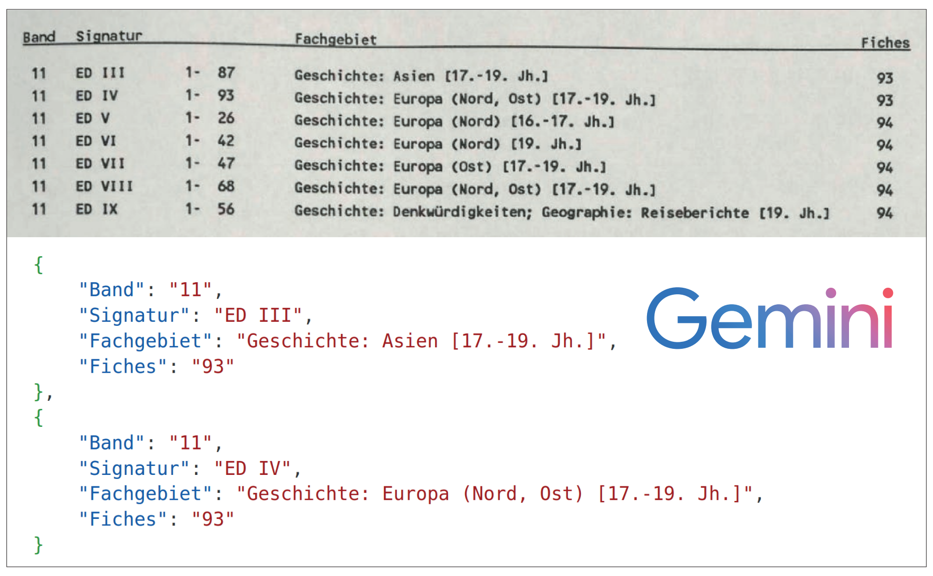
High Performance Computing (sciCORE)
UB IT
Interviews
Survey
1. Researchers' needs (requirements)
4. Infrastructure (reality check)
Scenarios
2. State of the Art (possibilities)
3. Data (restrictions)
Quality Enrichments (GND, …) Corpus Description
LLMs
Retrieval Augmented Generation

Figure 1: The four working groups with example tasks

2 Infrastructure / State of the Art

Two working packages concentrate on State of the Art solutions and existing infrastructures as this big dataset means a shift from libraries' focus on metadata to content and leads the path to new fields. Therefore we analyzed, for e.g., the existing TDM environments and the possibilities offered by Google Books itself.

This also means evaluating existing IT components of the libraries' infrastructures as well as building bridges with others (e.g. High Performance Computing) to either help researchers process large datasets or find solutions to enrich the content using large language models (LLMs).

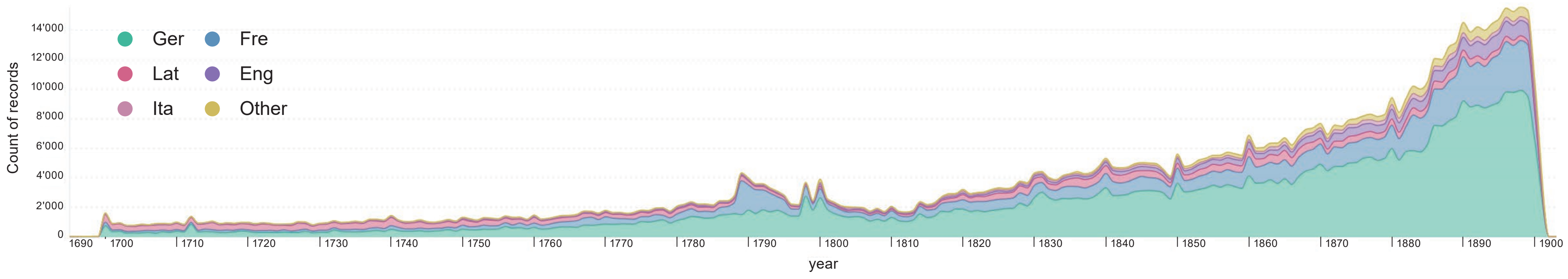


Band	Signatur		Fachgebiet	Fiches
11	ED III	1- 87	Geschichte: Asien [17.-19. Jh.]	93
11	ED IV	1- 93	Geschichte: Europa (Nord, Ost) [17.-19. Jh.]	93
11	ED V	1- 26	Geschichte: Europa (Nord) [16.-17. Jh.]	94
11	ED VI	1- 42	Geschichte: Europa (Nord) [19. Jh.]	94
11	ED VII	1- 47	Geschichte: Europa (Ost) [17.-19. Jh.]	94
11	ED VIII	1- 68	Geschichte: Europa (Nord, Ost) [17.-19. Jh.]	94
11	ED IX	1- 56	Geschichte: Denkwürdigkeiten; Geographie: Reiseberichte [19. Jh.]	94

```
{
    "Band": "11",
    "Signatur": "ED III",
    "Fachgebiet": "Geschichte: Asien [17.-19. Jh.]",
    "Fiches": "93"
},
{
    "Band": "11",
    "Signatur": "ED IV",
    "Fachgebiet": "Geschichte: Europa (Nord, Ost) [17.-19. Jh.]",
    "Fiches": "93"
}
```

Figure 2: Extracting topic enrichment from analog call numbers concordances using Gemini

DOI: 10.5281/zenodo.13908139



3 Data

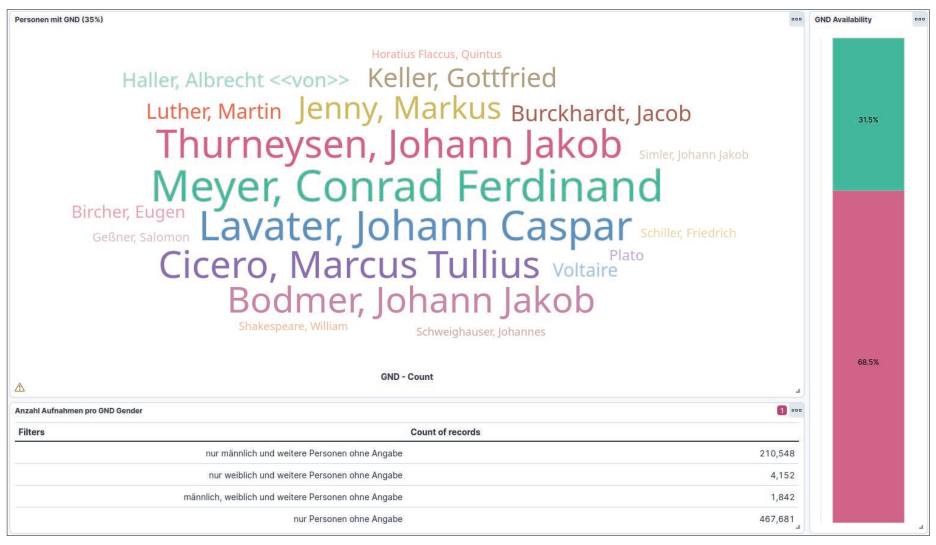
This is the connection to the data working package which tries to describe the limits and possibilities of the heterogeneous dataset to enhance its usability.

In addition to the metadata from the library catalog, the dataset comprises JPEG 2000 images (42 TB), hOCR, txt (300 GB), and METS files (180 GB) with structural information from Google regarding table of contents, indexes, images, and so forth (not always precise).

The libraries can regularly download new versions from Google that improve the OCR quality or even add missing pages. On the other hand, this means that the content is not static.

3.1 Metadata

Libraries' metadata records from this period are often quite limited and do not provide information about subject or genre, nor do they have many references to authoritative data. We are trying to make these gaps transparent and add enrichment where possible to provide better insight.



Personen mit GND (35%)

GND Availability

Anzahl Aufnahmen pro GND Gender

Filters	Count of records
nur männlich und weitere Personen ohne Angabe	210,548
nur weiblich und weitere Personen ohne Angabe	4,152
männlich, weiblich und weitere Personen ohne Angabe	1,842
nur Personen ohne Angabe	467,681

Figure 3: GND availability in records

3.2 Creating a Swiss corpus?

We also try to describe how representative this digital dataset is, in order to make the limitations of this Swiss corpus transparent and to think about possibilities to extend it with other digitally available resources.

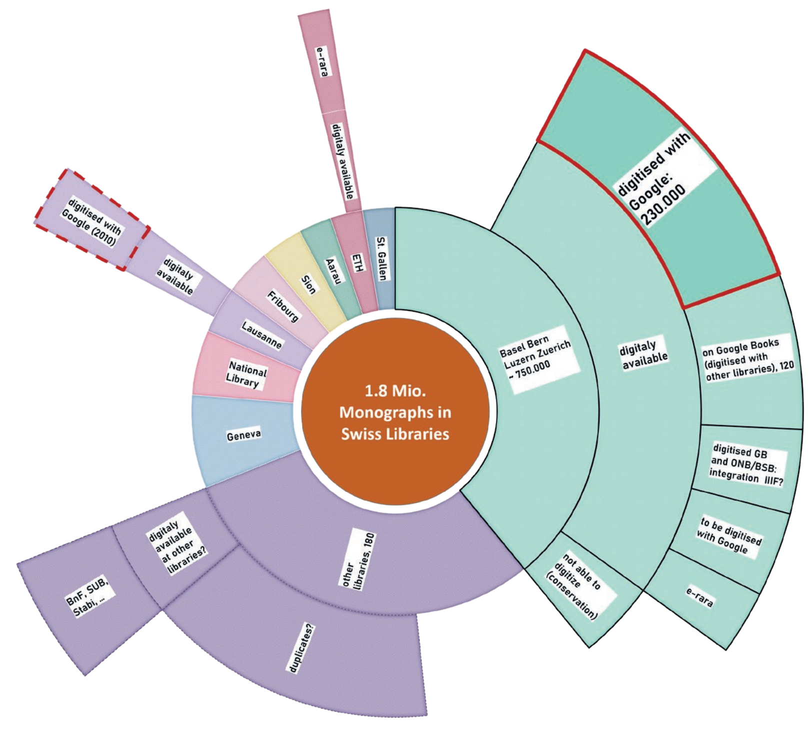


Figure 4: Proportion of the Swiss corpus digitally available

4 Researchers' needs

Our most crucial working package is geared towards understanding the different researchers' needs and extracting their requirements. We use the results of existing studies as a starting point.

We focus on getting in touch with a wide range of researchers from computer science to history in order to understand their **needs regarding this specific corpus.** We are using a variety of formats, from semi-structured interviews to roundtable discussions, conferences, and an online survey.

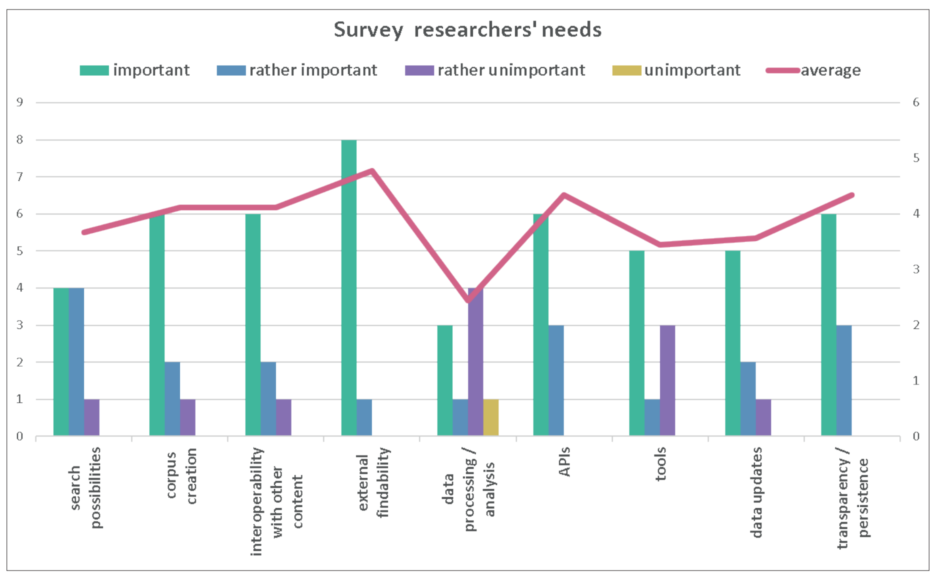


Figure 5: First evaluation of the quantitative part of the online survey using requirements clusters built beforehand

4.1 Usage rights

NON-COMMERCIAL USE ONLY

While the digitized content is within the public domain, an agreement was made with Google, who are financing the scanning of the books, and therefore no commercial use nor mass download is allowed. But exceptions are defined for researchers. Our **objective will be to sign agreements with all Swiss universities** so the data can be used freely by Swiss researchers without thinking about data agreements.

4.2 Findings

On a very abstract level we extracted that metadata is important to build relevant subsets, but that data quality is not the key issue for many. More transparency at all levels, especially about data production and its gaps, is crucial.

Libraries are also less expected to provide complete TDM environments but to provide good and easy to use APIs to access data and provide persistence.

Your input counts! Please share your needs and ideas with us or fill out our online survey!

Contact

Martin Reisacher	martin.reisacher@unibas.ch	iD 0009-0008-4529-5291
Eric Dubey	eric.dubey@unibas.ch	iD 0000-0002-9300-9762
Matteo Lorenzini	matteo.lorenzini@unibas.ch	iD 0009-0009-4159-5614