Multimodal Datasets of the Berlin State Library

David Zellhöfer
Berlin State Library/Staatsbibliothek zu Berlin, Germany
https://staatsbibliothek-berlin.de/
david.zellhoefer@sbb.spk-berlin.de

Abstract

To facilitate the handling of digital library content and its accompanying metadata, four multimodal and multilingual datasets are presented that are relying on the publicly available information systems of the Berlin State Library. They range from pre-processed extracts of the full main catalog of the library with ca. 9.8 million records, over various networks graphs modeling, e.g., relations between authors and languages, to more than half a million extracted illustrations detected by the day-to-day OCR process of ca. 22,000 historical media units such as historical books.

2012 ACM Subject Classification Applied computing → Digital libraries and archives

Keywords and phrases Digital Library, Bibliographic Metadata, Graph, Digitized Media, Dataset

Supplement Material https://github.com/elektrobohemian/StabiHacks

Funding Funded by the BMBF as part of the 'QURATOR - Curation Technologies' project.

1 Introduction

Most large research libraries such as the Berlin State Library are handling the core challenge of the digital transformation – the presentation of digitized content and retro-converted digitized catalog records – as a day-to-day routine, even at large scale. Media are digitized at a daily basis, extended with structural information, indexed, and treated with OCR engines, to be presented in web-based digitized collections[1] or to be distributed via typical metadata interchange interfaces such as OAI-PMH[2].

However, new tasks for research libraries are emerging, e.g., the digital curation of the owned collections and the provision of data for various research tasks from the wide field of digital humanities (DH). As these use cases fall outside the traditional bibliographic use case, i.e., the indexing and retrieval of different media, traditional bibliographic records do not not satisfy the requirements of both researchers in DH as well as digital curators. On the one hand, these records are often missing vital information such as named entities or other information that can be used to enable explorative information seeking strategies. On the other hand, these records contain very detailed information that is necessary for the bibliographic use case while being over-complex and cryptic for DH researchers. Furthermore, proprietary character encodings or system-specific annotations are putting an additional burden on the usage of the data outside the scope of common library tasks. Listing 1 illustrates this phenomenon very well with the help of the library management system's internal Pica+ format.

Because of the sheer amount of data available in large libraries, a manual conversion or augmentation of these records to fit the aforementioned needs would be very cost-intensive and hardly possible if it had to be carried out by library staff. Thus, a machine-assisted approach to transform traditional metadata records into datasets usable by digital curators or DH researchers is needed to cope with this problem. A recent proof of concept [2] shows

[1] https://digital.staatsbibliothek-berlin.de/
[2] https://www.openarchives.org/OAI/openarchivesprotocol.html

LDK 2019 - Posters Track.
Editors: Thierry Declerck and John P. McCrae

Listing 1 Excerpt of a bibliographic record in Pica+ format

```
011@ a1812
011B a2004-b2007
019@ aXD-US
021A aAn @oration pronounced at Dedham on the anniversary of
     American independence, July 4, 1812 hby Jabez Chickering
028A dJabez aChickering h1753-1812
033A pBoston nPrinted by Joshua Belcher
101@ a11
201B/01 014-03-17 t23:01:04.000
```

the feasibility of such an approach relying on methods from machine-based learning, data analysis, and traditional data management and batch processing.

The following section presents the core characteristics of four multimodal and multilingual datasets based on publicly available catalog and other metadata of the Berlin State Library that have been transformed by tools presented in [2]. For the sake of reproductiveness and transparency, all scripts are made available[3] with a permissive license.

2 Characteristics of the Datasets

All of the presented datasets are inter-linkable with the help of the so-called PPN (Pica production number). In most cases, the PPN can be seen as a unique identifier for analog or digitized media that is used in many systems of the Berlin State Library and the libraries of the GBV alliance[4], e.g., the central catalog. PPN can also be used to download image content via the IIIF[5] endpoint or metadata and OCR content via the OAI-PMH interface. For some sample scenarios, refer to [7].

2.1 Extract from the Library's Main Catalog

This dataset [3] is derived from the Pica+ serialization of the full library's main catalog from 2018 containing 9,850,467 records of analog, digitized, and digital-born material. The following fields have been extracted: title, author (incl. optional GND[6] ID), publisher, place of publication, country of publication, and year of publication. To facilitate further processing, the publications are split by language groups (ranging from ancient to modern languages).

The records are stored in a simple tabulator-separated field-based text format. Records are isolated by empty lines, whereas @ serves as a subfield indicator in case a GND ID or detailed location information is available. Table 1 presents a sample records, whose complete data can be referenced with the help of the given PPN[7]. Details on the different Pica+ field IDs and their contents are available under [3] accompanied by the creation script. A full list of available fields (in German) is also available[8].

[3] https://github.com/elektrobohemian/StabiHacks
[4] https://www.gbv.de/?set_language=en
[5] https://iiif.io/
[6] https://www.dnb.de/EN/Standardisierung/GND/gnd_node.html
[7] http://stabikat.de/DB=1/SET=1/TTL=1/PRS=PP%7F/PPN?PPN=0249445468
[8] https://www.gbv.de/bibliotheken/verbundbibliotheken/02Verbund/01Erschliessung/02Richtlinien/01KatRicht/inhalt.shtml

Table 1 Sample record of PPN 0249445468

PPN	Pica+ field ID	Content
0249445468	011@	1939
0249445468	019@	XD-US
0249445468	021A	The plays of William Shakespeare in thirty-seven volumes
0249445468	028A	Shakespeare, William@gnd/118613723
0249445468	033A	The Limited Editions Club@New York, NY

2.2 Metadata, Title Pages, and Network Graph of the Digitized Content of the Berlin State Library

The dataset has been downloaded via the OAI-PMH Dublin Core endpoint of the Berlin State Library's Digitized Collections[9] and has been converted into common tabular formats and graph representations in GML. It contains 146,000 records of digitized material older than 1920 in the format described in Table 2.

In addition to the bibliographic metadata, representative images of the works have been downloaded and resized to a 512 pixel maximum thumbnail JPEG image preserving the original aspect ratio. Title pages have been derived from structural metadata created by scan operators and librarians. If this information was not available, first pages of the media have been downloaded. In case of multi-volume media, title pages are not available. As a consequence, only 141,206 images title/first pages are present. Additionally, geo-spatial coordinates have been added to each record using the OpenStreetMap web service[10]. For details, refer to [5].

2.3 Title, Author, Publisher, Place of Publication, and Language-related Network Graphs of the Library Main Catalog

Three graphs (in GraphML, GML, and JSON) are made available in this dataset linking:

- authors, publishers, and places of publication (`author_publisher_location`);
- authors, publishers, places of publication, and titles (`author_publisher_location_title`);
- authors, publishers, and the language of publication (`languageLink`).

The languages of publication graphs spans all of the languages mentioned above and has 1,555,119 nodes and 1,659,596 edges (see Fig. 1 for an exemplary subgraph). Table 3 subsumes the core properties of each provided graph. For additional details, see [6].

2.4 Extracted Illustrations of the Berlin State Library's Digitized Collections

The largest dataset consists of ca. 22,142 digitized media units[11] that have been OCR-processed with the ABBYY FineReader Engine (at least version 11) and whose full-texts are made available in ALTO XML referenced in the METS/MODS XML file of each object[12]. Based on the results of the OCR, all found illustrations have been extracted and saved in

[9] https://digital.staatsbibliothek-berlin.de/oai
[10] https://www.openstreetmap.org/
[11] This number is subject to change as the OCR and image extraction process is ongoing.
[12] The data is available over the OAI-PMH METS/MODS endpoint of the digitized collections.

Table 2 Description of the tabular format of the extended metadata

Column Name	Description
title	The title of the medium
creator	Its creator (family name, first name)
subject	A collection's name as provided by the library
type	The type of medium
format	A MIME type for full metadata download
identifier	An additional identifier (most often the PPN)
language	A 3-letter language code of the medium
date	The date of creation/publication or a time span
relation	A relation to a project or collection a medium has been digitized for.
coverage	The location of publication or origin (ranging from cities to continents)
publisher	The publisher of the medium.
rights	Copyright information.
PPN	The unique identifier that can be used to find more information about the current medium in all information systems of Berlin State Library.
	The following fields contain data that is based on different processing steps.
spatialClean	In case of multiple entries in coverage, only the first place of origin has been extracted. Additionally, characters such as question marks, brackets, or the like have been removed. The entries have been normalized regarding whitespaces and writing variants with the help of regular expressions.
dateClean	As the original date may contain various format variants to indicate unclear creation dates (e.g., time spans or question marks), this field contains a mapping to a certain point in time.
spatialCluster	The cluster ID determined with the help of the Jaro-Winkler distance on the spatialClean string. This step is needed because the spatialClean fields still contain a huge amount of orthographic variants and latinizations of geographic names.
spatialClusterName	A verbal cluster name (controlled manually).
latitude	The latitude provided by OpenStreetMap of the spatialClusterName if the location could be found.
longitude	The longitude provided by OpenStreetMap of the spatialClusterName if the location could be found.
century	A century derived from the date.
textCluster	A text cluster ID on the basis of a k-means clustering relying on the title field with a vocabulary size of 125,000 using the tf*idf model and k=5,000.
creatorCluster	A text cluster ID based on the creator field with k=20,000.
titleImage	The path to the first/title page relative to the img/ subdirectory or None in case of a multi-volume work.

Table 3 Graph properties per language

Language	Graph Type	Nodes	Edges	Records
fry	author_publisher_location	298	264	360
fry	author_publisher_location_title	622	726	360
ice	author_publisher_location	505	448	1,200
ice	author_publisher_location_title	1,509	1,393	1,200
por	author_publisher_location	5,217	5,392	8,937
por	author_publisher_location_title	12,848	15,219	8,937
nor	author_publisher_location	4,948	6,049	12,016
nor	author_publisher_location_title	15,276	21,737	12,016
dan	author_publisher_location	9,127	11,711	20,089
dan	author_publisher_location_title	26,144	39,278	20,089
swe	author_publisher_location	15,350	18,367	30,628
swe	author_publisher_location_title	41,933	61,000	30,628
spa	author_publisher_location	24,404	27,477	42,540
spa	author_publisher_location_title	59,339	77,779	42,540
dut	author_publisher_location	36,503	42,128	67,000
dut	author_publisher_location_title	94,803	127,785	67,000
ita	author_publisher_location	71,151	95,054	158,851
ita	author_publisher_location_title	206,656	316,282	158,851
lat	author_publisher_location	91,584	148,224	230,588
lat	author_publisher_location_title	273,724	469,322	230,588
fre	author_publisher_location	174,650	204,299	380,569
fre	author_publisher_location_title	487,053	693,245	380,569
eng	author_publisher_location	606,112	880,989	1,309,172
eng	author_publisher_location_title	1,778,957	2,710,807	1,309,172
ger	author_publisher_location	705,468	1,104,502	2,316,600
ger	author_publisher_location_title	2,497,239	3,947,482	2,316,600
n/a	languageLink	1,555,119	1,659,596	4,578,537

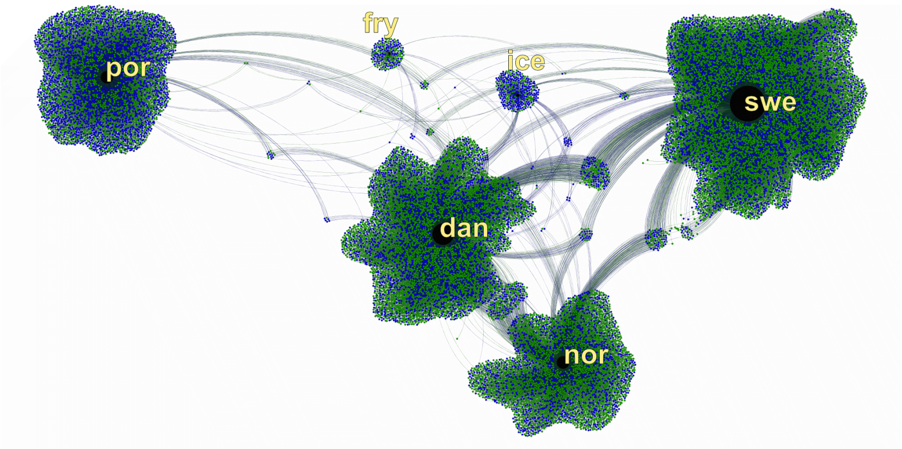

Figure 1 Network of Authors, Publishers, and Languages of Publication (Subgraph of the Languages: fry, ice, por, nor, dan, swe)

original size in JPEG format. In total, 531,484 illustrations have been extracted from 22,142 media units, i.e., an average of 24 extracted illustrations per unit [4].

In order to remove false positives from the corpus, e.g., stamps, hand-written signatures, or empty pages, pre-trained classifiers are provided in form of different Python scripts[1] based on a pre-trained VGGnet models implemented with Keras/TensorFlow.

References

1 Julia Berauer, Ralitsa Doncheva, Linh Nguyen, Luisa Rademacher, Carlos Tan, and Caglar Özel. Chasing Unicorns and Vampires in a Library. Technical report, HTW Berlin, 2018. URL: `https://github.com/elektrobohemian/imi-unicorns`.

2 David Zellhöfer. Exploring Large Digital Libraries by Multimodal Criteria. In Norbert Fuhr, László Kovács, Thomas Risse, and Wolfgang Nejdl, editors, *Research and Advanced Technology for Digital Libraries - 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016, Hannover, Germany, September 5-9, 2016, Proceedings*, volume 9819 of *Lecture Notes in Computer Science*, pages 307–319. Springer, 2016.

3 David Zellhöfer. Extract from the Library's Main Catalog, March 2019. URL: `https://doi.org/10.5281/zenodo.2590752`.

4 David Zellhöfer. Extracted Illustrations of the Berlin State Library's Digitized Collections, March 2019. URL: `http://doi.org/10.5281/zenodo.2602431`.

5 David Zellhöfer. Metadata, Title Pages, and Network Graph of the Digitized Content of the Berlin State Library (146,000 items), March 2019. URL: `https://doi.org/10.5281/zenodo.2582482`.

6 David Zellhöfer. Title, Author, Publisher, Place of Publication, and Language-related Network Graphs of the Library Main Catalog, March 2019. URL: `https://doi.org/10.5281/zenodo.2587801`.

7 David Zellhöfer. What is a PPN and Why is it Helpful?, May 2019. URL: `http://doi.org/10.5281/zenodo.2702544`.