From Domain Experts to Data Scientists

Connecting the Heritage of the Printed Book Database and its users


Andreas Walker

SUB Göttingen / Data Conversion Group
The HPB Database (previously called the Hand Press Book Database) is a steadily growing collection of files of catalogue records from major European and North American research libraries covering items of European printing of the hand-press period (c.1455-c.1830) integrated into one file. This makes it possible for information to be retrieved in one single search across all files. As the digitisation of collections in contributing libraries progresses, more and more catalogue records point to digital presentations of the early printed books.

The HPB Database is of interest to librarians and any one else with academic pursuits across many fields of study that use printed books as source material. It is especially valuable for research in intellectual history, social history, and transmission of thought as well as in the history of printing and the history of the book. It may be accessed for information retrieval and downloading by CERL member institutions, their staff and users.

Notice

- This database is presented in Unicode (UTF-8). Possible character coding problems might be related to locally missing character fonts (Points) in your operating system. The following links could be helpful: Display Problems (Unicode Consortium), UTF8 problems (Wikipedia).
- Usage of MSC (Mathematics Subject Classification) in line with Creative Commons Attribution Noncommercial Share Alike License.
CERL - Heritage of the Printed Book Database. The HPB Database (previously called the Hand Press Book Database) is a steadily growing collection of files of catalogue records from major European and North American research libraries covering items of European printing of the hand-press period (c.1455-c.1830) integrated into one file. This makes it possible for information to be retrieved in one single search across all files. As the digitisation of collections in contributing libraries progresses, more and more catalogue records point to digital presentations of the early printed books.
Project context

• My background: Librarianship / Linguistics / Media Studies

• 2017-2019: *Bibliotheksreferendariat* (Academic Librarian Training) at the SUB Göttingen, M.A. (LIS) at Humboldt University

• This project = final project + thesis

• **Aim:** To explore the potential of a linked open data representation of the bibliographic data in CERL‘s Heritage of the Printed Book Database (HPB), and the development of a web interface for this representation based on the needs of current HPB users.
About the HPB

- **Scope:** bibliographic records of early European prints (c. 1450-1830), aggregated from member libraries

- Roughly **eight million records** stored in a **Pica** database hosted in Göttingen and made accessible through a traditional OPAC interface at [http://hpb.cerl.org](http://hpb.cerl.org) (to humans) and through Z39.50 and SRU (to machines)

- Open to non-members since January 2018

- CERL's aim: increase the **visibility** and **usefulness** of this data to researchers and other interested users
The HPB Database (previously called the Hand Press Book Database) is a steadily growing collection of files of catalogue records from major European and North American research libraries covering items of European printing of the hand-press period (c.1455–c.1630) integrated into one file. This makes it possible for information to be retrieved in one single search across all files. As the digitisation of collections in contributing libraries progresses, more and more catalogue records point to digital presentations of the early printed books.

The HPB Database is of interest to librarians and any one else with academic pursuits across many fields of study that use printed books as source material. It is especially valuable for research in intellectual history, social history, and transmission of thought as well as in the history of printing and the history of the book. It may be accessed for information retrieval and downloading by CERL member institutions, their staff and users.

Notice

- This database is presented in Unicode (UTF-8). Possible character coding problems might be related to locally missing character fonts (Points) in your operating system. The following links could be helpful: Display Problems (Unicode Consortium), UTF-8 problems (Wikipedia).
- Usage of MSC (Mathematics Subject Classification) in line with Creative Commons Attribution Noncommercial Share Alike License.
1. Das Babtische Gesangbuch von 1545: Faksimiliedruck

2. Grammatica Amharica
   Ludolf, Hiob. - Nachdr. - Halle /Saale, 1866

3. Geistliche Lieder aufs new gebessert zu Wittenberg
   Luther, Martin. - Kassel: Bärenreiter, 1983

4. Marburger Kirchenordnung von 1527
   Hochhuth, Carl Wilhelm Hermann. - Kassel: [Drucker:] Scheel, 1878

5. Die Bibel oder die Heilige Schrift des Alten u. Neuen Testaments
   Luther, Martin. - 3. Aufl. - Leipzig: Brockhaus, [1875]

6. Zivot Martina Lutherova církevněhistorickým uvodem ku prospechu Krestana evangelickych. [Luther Mátó élete rövid egyháztörténeti bevezetéssel]
   Kuzmány Károly. - W Peszt: Traktner-Károlyi, 1842

7. Geschichte der deutschen Bibelübersetzung Luther's, zur Erinnerung an das dreihundertjährige Jubiläum dieselben im Jahre 1834
   Weidemann, Karl Albert. - Leipzig: Hinrichs, 1834

8. Luther als raadgever en verrooster voor onzen tijd: eene stem uit den tijd der Hervorming tot ons geslagen, bij het heerschen der cholera.
   Hey en zoon, Johannes van der Amsterdam. - Amsterdäm: Johannes van der Hey en zoon, 1832

9. Bibela, dat is: de geheele H. Schrift, behelzende al de boeken des Ouden en Nieuwen Testaments
   Luther, Martin. - Dortrecht : Blussé en Van Braam, 1832

    Luther, Martin. - Rinteln: Steuber, 1631

11. The history of the Church of Christ : from the days of the apostles, till the famous disputation between Luther an Militz, in 1520 ; with a further continuation to the present times, extracted from the Rev. T. Hewes' church history ; complete in one volume
    Milner, Joseph. - London [u.a.]: Tegg, (1830)

12. Das Alte Testament

13. Geschichte der vornehmsten Reformatoren und der Folgen ihrer Bemühungen von Jesus Christus an bis auf Martin Luther und den dreißigjährigen Krieg : ein belehrendes und unterhaltendes Volksgespräch zur Vorbereitung auf die Konfessionsteuer am 25. Juni 1830
    Oetwil, Eucharius Ferdinand Christian. - Nürnberg, 1830

14. Exegetica opera latina. 6 Continens Exarationes in Genesis, Cap. 24 - 26
    Luther, Martin. - [Erlanger Ausg.]. - Erlangen: Heyder, 1830

15. Exegetica opera latina. 5 Continens Exarationes in Genesis, Cap. 20 - 22
    Luther, Martin. - [Erlanger Ausg.]. - Erlangen: Heyder, 1830
**HPB Identifier:** NO-0030100.01.910360928

**Cataloguing Source:** National Library of Norway (Nasjonaldigitalebiblioteket, Oslo)

**Title:** Das Babstscbe Gesangbuch von 1545: Faksimiledruck / mit einem Geleitwort herausgegeben von Konrad Amelin

**Collaborator:** Amelin, Konrad; Babst, Valentin; Luther, Martin "1483-1546"

**Edition:** 3. Aufl.

**Published:** Kassel: Bärenreiter, 1988

**Language:** German

**Extent:** [554]. 16 s.: mus.

**Series:** Documenta musicologica. Reihe 1, Druckschriften-Faksimiles; 38

**Note:** Også med original titelside. Geystliche Lieder / mit einer newen vorhede Mart. Luth. Contains: Babst, Valentin: Geystliche Lieder

Contains: Luther, Martin: Geystliche Lieder

Related title: Geystliche Lieder

**ISBN:** 3-7618-0009-3

**Subject headings:** *Tyskland
*koraler
*salmodier

**Subject (DDC):** DDC/nor: 782.5250643

**Former Identifier:** NO-0030100.01.910360928

**Record in source system**

**Copies**

**Holding Institution:** Nasjonaldigitalebiblioteket

**Shelfmark:** TZ Doc:38

**Location:** Seksjon Biblioteket. Utlånet, NB/BRU Musik

**Country:** Norway
Our users

• The „typical“ HPB users is a **domain expert**, but not necessarily a **tech user** (distinctions due to Dadzie & Rowe 2011)

• New tools should integrate with researchers‘ **existing workflows**

• My methodology: **qualitative interviews** with five researchers to find out about their current use of the HPB and their desiderate for future version
What we learned

- Participants see the HPB as a **useful tool** but had encountered some problems working with it (e.g., incomplete data, difficulties in accessing the data, no clustering of copies by edition)

- Participants are interested in the possibility of manipulating and enriching data directly within a future **web interface** of the HPB, but also need their workflows based on standard office software to be supported

- Participants mostly mention **maps** and **timelines** as visualizations they would like to be able to generate from the data
Design decisions

- Develop an interface that is aimed at **domain experts** and does not require them to interact directly with technologies like SPARQL

- Let the data presentation be guided by researchers‘ workflows rather than by the internal data structure

- Maintain a **coherent visual identity** with the current HPB

- Make generous use of **map visualizations** to quickly generate added value to the new interface
The underlying technology

- Data conversion from Pica+ to RDF, using BIBFRAME 2.0 as the data model to increase interoperability, implemented in Python

- Enrichment of the data from external data sources, including the CERL Thesaurus and Wikidata

- Provision of a SPARQL endpoint through Apache Fuseki

- Web application built in Python / Flask, with the triple store serving as the database queried through the SPARQL endpoint
System architecture
CERL - Heritage of the Printed Book Database. The HPB Database (previously called the Hand Press Book Database) is a steadily growing collection of files of catalogue records from major European and North American research libraries covering items of European printing of the hand-press period (c.1455-c.1830) integrated into one file. This makes it possible for information to be retrieved in one single search across all files. As the digitisation of collections in contributing libraries progresses, more and more catalogue records point to digital presentations of the early printed books.
Advanced Search Options

SPARQL Interface
This version of the HPB is implemented as Linked Open Data in RDF. It allows complex queries using the powerful SPARQL language. Note that this requires an understanding of the underlying data model, BIBFRAME.

To the SPARQL interface →

Demonstrations
Combining multiple data sources through linked data allows for a number of interesting queries.

[1] Items held outside their country of origin (matches HPB country data to Wikidata’s country data for holding institutions)

Traditional interface

Title: 
Author: 
Place of Publication: 
Publisher: 
Country: Any
Language: Any

Search
Please note that the HPB data is stored in a named graph. You can query it by including the following statement in your SPARQL query:

FROM <http://141.5.107.58:3030/prot01/data/records>

```
SELECT *
FROM <http://141.5.107.58:3030/prot01/data/records>
WHERE {
}
LIMIT 10
```
Your results searching for items held outside their country of origin

11 records not displayed because they lack geographical data for their holdings. See list view for all records.
1. M. Lutheri Catechesis major  
   Luther, Martin (1483-1546)

2. Bildnis Margarethe Luther

3. Kurtze Christliche Fragstücke / Martin Luther  
   Luther, Martin (1483-1546)

4. Jesus Syrach Teutsch / Martin Luther [Vorr.]

5. Quae de genuino systemate Lutheranorum solicta est, et quindecim priores  
   §§ complectitur  
   Mörl, Gustav Philipp (1673-1750)

6. Vera fides, seu Romanæ ecclesiae evidens credibilitas et Lutheranae  
   aliarumque sectarum certa falsitas solide demonstrata  
   Pichler, Vitus (1670-1736)

7. Ein neuer Moses : Nemblich ein öffentlich Zeugniß, daß die Papisten mit  
   diesem ... Krieg nicht allein die genannten Calvinisten, sondern auch die  
   Lutheraner ... außzureutten gemeinet seind

8. D. Mort. Luthers liiden Catechismus

9. De egressu et ingressu Lutheri Wormatiam carmen  
   Philomusus, Decius

10. Decem Precepta witt-e|bergenst predicata || populo || per P. Mar||tinum  
    Luther || augustinia||num.||  
    Luther, Martin

11. Der ganze Psalter des Königes und Propheten Davids / verdeutscht von D.  
    Martin Luther. Mit jedes Psalms kurzen Summarien und nöthigsten  
    Parallelen.

12. Instructio Pro Vero Lutherano elegantci carmine scripta / ab Ernesto
Your results searching for luther

Place of publication: Regiomonti
Author: Krause, Johann Friedrich (1770-1820)
Title: Ulterior et quantum quum consilio et successu theologorum recitantes, qui omnia scripturae sacrae
interpretationem et rationem revocat. A Lutheri mente atque legibus, quas ille sequendas putavit.
decem. 1 /Joan. Frid. Krause/
See record

6 records not displayed because they lack geographical data. See list view for all records.
19. Reinchristlicher Religions-Unterricht nach Luthern : Ein Katechismus zum Gebrauch der Schullehrer / (C. L. Krüger)
Krüger, Carl Ludwig


21. Das Benedictus odder weyssagung des heyligen Zacharie, Luce. j. / durch D. Mart. Luther gepredigt und ausgelegt
Luther, Martin (1483-1546)

22. Looke beyond Luther : or an answere to that question proposed by our adversaries where our religion was before Luthers time
Bernard, Richard (1568-1641)

23. Eyn gute trostliche predig vo[n] der wirdigen bereytu[n]g zu de[m] hochwirdigen Sacrament / Doctor Martini Luther Augustiner tzu Wittenberg
Luther, Martin (1483-1546)

Luther, Martin

25. Grund und Karrn des Catechißmi, D. Martini Lutheri : Einfeltig u. kürzlich auß Gottes Wort, u. reiner Lehrer, sonderlich aber D. Lutheri Schriften ...

Luther, Martin (1483-1546)

27. Utrum et quantum quove consilio et successu theologi recentiores, qui omnem scripturae sacrae interpretationem ad rationem revocent, a Lutheri mente atque legibus, quas ille sequendas putavit, defecerint. 1 / [loan. Frid. Krause]
Krause, Johann Friedrich (1770-1820)
Schmalkaldische Artickel, So da hetten s ...
Evaluation of the first prototype

- Methodology: qualitative, semi-structured follow-up interviews with the previously interviewed researchers after providing them with a link to the prototype and giving them about a month for testing it.

- Questions asked were focused on:
  - Their overall visual impression of the interface
  - Their opinion on the presentation of search results
  - Their opinion on the presentation of single records
  - Any remaining thoughts on usability
Evaluation of the first prototype

- Researchers responded in writing. The following is a free summary of the points most salient to me; see Walker (2019) for a more complete list:

  - **Visual impression:**
    - Colours and layout mostly judged as attractive and well-organized
    - Search interface too small and cluttered

  - **Search results:**
    - Participants liked the map visualization and the options for filtering and manually editing result lists
    - Result counter not obvious enough
    - Result list should display different metadata
Evaluation of the first prototype

• Single records:
  • Participants liked the map visualization and the overall organization of the presentation
  • Wikidata and RDF links were judged as not useful
  • Interface displays less data than the original HPB
  • Other identifiers should be added (e.g., GdW, MEI, VD16)

• Usability:
  • Participants liked the additional integration of the HPB with other CERL resources (i.e., the CERL Thesaurus) and suggested adding more such links
  • The search functionality of the prototype was judged as inadequate
Advanced Search Options

This form gives you a traditional advanced search interface, and constructs queries in SPARQL, the query language for linked open data, in the background. You can see the SPARQL queries generated by the form in the Export pane of your search results.

If you want to write your own queries, click here →
Advanced Search Options

This form gives you a text interface, and constructs queries in SPARQL, the query language for linked open data, in the background. You can save searches generated by the form in the Export pane of your search results.

If you want to write your own SPARQL or SQL queries, you can do so here →
Advanced Search Options

Title
Luther

Country
France

Language
- Arabic
- Armenian
- Bokmål, Norwegian | Norwegian Bokmål
- Chinese
- Croatian
- Czech
- Danish
- Dutch | Flemish
- English
- Finnish
- French
- French, Middle (ca. 1400-1500)
- German
- Greek, Ancient (to 1453)
- Greek, Modern (1453-)
- Hebrew
- Hungarian
- Icelandic
- Italian
- Japanese

- Remove field
- Remove field
- Remove field

Begin search

This form gives you a traditional advanced search background. You can see the SPARQL queries generated by pressing 'Begin search' for linked open data, in the

If you want to write your own queries, click here.
Advanced Search Options

Title

Published near river

+ Add search field

Remove field

Remove field

This form gives you a traditional advanced search interface, and constructs queries in SPARQL, the query language for linked open data, in the background. You can see the SPARQL queries generated by the form in the Export pane of your search results.

If you want to write your own queries, click here →
Your 3 results searching for your SPARQL query

1. Exposition de l'histoire des dix lépreux [par Martin Luther]. Translatée de latin en français... [par Antoine Du Pinet] [Texte imprimé]
   Luther, Martin (1483-1546)

   Maimbourg, Louis

3. Histoire du Lutheranisme, par le P. Louis Maimbourg, de la Compagnie de Jesus.
   Maimbourg, Louis (1610-1686)
Your 3 results searching for your SPARQL query
<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>Maimbourg, Louis</td>
</tr>
<tr>
<td>Title</td>
<td>Histoire du Lutheranisme.</td>
</tr>
<tr>
<td>CID</td>
<td>GB-UkCyUK.CLC.CLC24878 (view in classic HPB)</td>
</tr>
<tr>
<td>Place of Publication</td>
<td>[Rotterdam?]</td>
</tr>
<tr>
<td>Publisher</td>
<td>Suivant la copie imprimée à Paris</td>
</tr>
<tr>
<td>Year</td>
<td>1681</td>
</tr>
<tr>
<td>Language</td>
<td>French</td>
</tr>
<tr>
<td>Country</td>
<td>Netherlands</td>
</tr>
<tr>
<td>Holdings</td>
<td>St Asaph Cathedral, Wales</td>
</tr>
<tr>
<td>Holdings</td>
<td>Salisbury Cathedral</td>
</tr>
<tr>
<td>Holdings</td>
<td>Worcester Cathedral</td>
</tr>
</tbody>
</table>
Histoire du Lutheranisme.

Author: Maimbourg, Louis
Title: Histoire du Lutheranisme.
CID: GB-UkCyUK.CLC.CLC24878 (view in classic HPB)
Place of Publication: [Rotterdam?]  
Publisher: Suivant la copie imprimée à Paris
Year: 1681
Language: French
Country: Netherlands
Holdings: St Asaph Cathedral
Holdings: Salisbury Cathedral
Holdings: Worcester Cathedral
Beyond prototypes

- **Scaling up:** A full database conversion currently runs at ~4 hours. We haven’t tested the web application with the full database yet, as it already runs slowly with ~15,000 records (mostly due to the effort of reconstructing records from the graph).
- **Improving search:** Full text search in SPARQL is not very efficient.

- **Potential solution:** Combine the triple store with a document-based database to build on their respective strengths.
Beyond prototypes

• Clustering records by editions: Partially already underway, independently of the conversion to RDF

• Visualization of results: Currently implemented through third party libraries that do not entirely match our requirements (e.g., permalinks, export)

• Enrichment of data: Identify useful data sets to link to, but also demonstrate the value of large data hubs like Wikidata to researchers
Beyond prototypes

- **Towards a research environment:** Allow users to create their own LOD data sets that build on HPB data and edit, search and access them through the same interface.

- **Expanding to new audiences:** Make the RDF data available to researchers who want to run large-scale analyses on our bibliographic data (cf. Lahti et al. 2019).
Thank you for your attention

I would also like to thank the DCG and CERL for hosting this project, as well as the five researchers who kindly provided their expertise.

Contact me at:
walker@sub.uni-goettingen.de

Bibliography

