Requirements for a Digital Library System: A Case Study in Digital Humanities

Hermann Kroll
krollh@acm.org
Institute for Information Systems,
TU Braunschweig
Braunschweig, Germany

Christin Katharina Kreutz
ckreutz@acm.org
TH Mittelhessen
Gießen, Germany
Herder Institute
Marburg, Germany

Mathias Jehn
Thomas Risse
{m.jehn,t.risse}@ub.uni-frankfurt.de
Goethe University Frankfurt,
University Library
Frankfurt, Germany

Abstract

Archives of libraries contain many materials, which have not yet been made available to the public. The prioritization of which content to provide and especially how to design effective access paths depend on potential users' needs. As a case study we interviewed researchers working on topics related to one German philosopher to map out their information interaction workflow. Additionally, we deeply analyze study participants' requirements for a digital library system. Moreover, we discuss how existing methods may meet their requirements and which implications these methods may have in a practical digital library setting.

CCS Concepts

• **Information systems** → *Digital libraries and archives*;

Keywords

Digital Humanities, Case Study, Discovery, Digital Libraries

ACM Reference Format:
Hermann Kroll, Christin Katharina Kreutz, Mathias Jehn, and Thomas Risse. 2024. Requirements for a Digital Library System: A Case Study in Digital Humanities. In *The 2024 ACM/IEEE Joint Conference on Digital Libraries (JCDL '24), December 16–20, 2024, Hong Kong, China.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3677389.3702502

1 Introduction

Archives of physical libraries are full of potentially interesting materials which have not yet been digitized [15]. The process of digitization is still laborious, therefore there needs to be a prioritization of what to make available to the public [9]. Digitization alone is not sufficient in enabling researchers to effectively access or explore the material – information access paths fitting the research interests or requirements should be constructed. In close cooperation with the University Library J. C. Senckenberg in Frankfurt am Main, we wanted to understand our users' requirements when implementing access paths to one of our library's materials, e.g., to two collections of the famous philosophers Horkheimer and Schopenhauer. While digitizing the content and processing it with object character recognition (OCR) is the first step of making the content available, the question arose about what kind of access paths could help users in their daily research lives. Briefly, is it enough to provide keyword-based access? Or are more advanced access paths desired by users? Which requirements need to be met?

We tackle the research question *(How) Can we support researchers from the philosophical and sociological domain?* in a case study with seven participants. Our technical report [8] extends this paper by discussing related work and concluded requirements in more detail.

2 Interviews with In-Domain Researchers

Seven senior researchers working on topics related to Schopenhauer participated in our study. We scheduled 30 minute online sessions with each participant. Sessions consisted of an introduction and consent (~ 5 min), a description of a researcher's workflow (~ 15 min), and semi-structured interview on liked, disliked and desired tools/parts in their workflow (~ 10 min). Participants described four components in their information interaction workflows:

1. Information Sources. Researchers usually rely on multiple information sources. They use physical libraries and archives, newspapers and if the research question requires it social media. General purpose digital libraries were mentioned. For catalogs participants mentioned OPAC and KVK. Other named general-purpose resources were Projekt Gutenberg-DE, the federal archive Germany, Arcinsys Hessen, and university or state libraries. Mentioned philosophy- or Schopenhauer-specific sources were PhilPapers, the Philosopher's Index, the Thesaurus Schopenhauerianus, the Schopenhauer archive and Schopenhauer yearbooks [3, 7].

2. Searching. Strategies range from simple keywords queries, keyword refinement, to combining keywords with Boolean operators. Some researchers use the linkage of search results to other archives to find more results or check the actual content of relevant texts for links to other documents. Filtering of search results and semantic search were regarded as helpful for some while it hindered others. Hand-crafted topical ordering of literature as found in libraries was mentioned as enabling serendipity finds.

3. Result Lists and Relevancy. In result lists of potentially relevant literature, all, many or only few items can be observed deeply. Sometimes lists are assessed in multiple passes. Relevancy decisions base on more than results' content: headlines or titles, keywords and/or semantics, publication date, the estate in which a material is stored, the content of a material's detail page, the document type, author, publication source and reviews from third parties. Observing material which one would consider irrelevant can also help sharpen and reaffirm one's notion of relevancy.

JCDL '24, December 16–20, 2024, Hong Kong, China

ACM ISBN 979-8-4007-1093-3/24/12
https://doi.org/10.1145/3677389.3702502

4. Working with Materials. Participants described either having a thought process ready before starting to search for literature, searching literature before coming up with a mental concept or refining the initial concept while working with literature. Literature is borrowed or archives and libraries are visited physically to read full texts. Some take manual notes, some annotate PDFs digitally. Typically, not all information is in the same language, so, an explicit translation step could take place. With the gained information articles can be written, biographic information on persons is checked, texts are manually classified or arranged as manual mind maps.

Liked, Disliked and Desired Components. Researchers *liked* their own workflows, the work of digital archives and the option of using LLMs for checking translations. In addition to some library-management related aspects in workflows, researchers *disliked* the inability to find relevant literature in different languages and searching for images in historical texts. Researchers *desired* the option to voice digitization requests online, obtaining citation information directly online, a bookmarking system with time stamps, semantic search, an overview of literature on popular topics, the option to search for complex interactions between actors, layperson versions of papers, a topical filtering or grouping of material and the ability to link extracted information from materials.

3 Requirements for a Digital Library System

With the help of our interviews, we collected a list of requirements, that a future digital library system should fulfill. We structure the requirements into the following categories: 1. Technical requirements which can be solved by established methods today, 2. requirements which are tackled in research, for which ready-to-use solutions, that work beyond *small demonstrations*, are yet missing. The full discussion is available in our technical report [8].

Technical Requirements. *Digitizing the Content.* First, the required content must be digitized with suitable OCR tools like tesseract or OCR-D. Content needs to be enriched with metadata as users wanted to filter by time, type, venue, author, and estate.

Keyword-based Search. Keyword-based search is a widely accepted paradigm for our users; see Apache Solr/Lucene for a practical implementation. It is easy to use and filters collections.

Federated Search. We saw that our users worked with many different systems, mainly because the content was distributed. Related information, for instance, to Schopenhauer, can be provided if other systems are either automatically queried or at least links to related systems/searches are shown in an overarching federated system.

Requirements and Possible Solutions. *Multilingual Retrieval.* Research in the area of Schopenhauer is published mainly in German, English, French, and Italian. Some of our users stated that they read articles in different languages but that the search is challenging because each query must be formulated for every language (and in different systems). The requirement here is thus a cross-lingual retrieval component. For instance, Europeana, Europe's largest platform for cultural heritage objects, faced a similar challenge, i.e., integrating and retrieving multilingual content [10].

Language Translation. Our users stated that they work with content in different languages. For instance, a person reads dissertations in Latin briefly to get an impression of what is contained and then uses ChatGPT to translate those dissertations into German. The person argued that Latin-to-German translation would be fine because the *person could read* Latin text. The person would, however, not use such a translation for Arabic as the *person*'s Arabic is *not good enough*. We argue that the quality of modern machine translation increased clearly and is *nearly* ready to be used for library content. However, the limitations of such translations must be clearly shown: When using translation to get an impression of what is told, systems for high-resource languages can be implemented today; see [14, 18] for evaluations of LLMs, [11] for a survey of neural machine translation, or platforms like DeepL. However, rechecking translations is still required and handling low-resource languages with NLP methods remains challenging [4, 5, 17].

Content Exploration. We observed that users wanted to explore a library's content, e.g., start by searching for *Schopenhauer AND Religion*, read some hits, and refine their search with terms related to or possibly replacing *Religion*. In brief, users manually derived *associated* terms to refine their searches. Existing approaches like finding associated terms through co-occurrences, word2vec [13] or generating keyword clouds with tools like YAKE [1, 2] could assist users in this process by displaying strongly related terms in the form of keyword clouds or some kind of query autocompletion.

More advanced access paths like question answering systems, e.g., Scopus AI, ChatGPT, and CORE-GPT [16], are becoming more present today. The main advantage from a user's perspective is that they can formulate their information needs as natural language questions. A system then replies with an answer formulated in natural language, sometimes accompanied by references/sources. For instance, CORE-GPT [16] takes a question, translates it into a keyword-based query, performs a search, and then replies with a list of possible related references. In contrast, LLMs like ChatGPT *just* generate an answer that could be based on a spectrum from something real to purely hallucinated. Distinguishing what is real and what is not is still the question of research [6]. In brief, we argue that LLMs could help in transforming natural language questions into internal query representations, e.g., keyword queries, SQL, or SPARQL, but the result of the process should, at this moment, still be content of the library and not a generated answer that might contain hallucinations. Even if a model comes with 99.9% factual consistency regarding the library's content, would the quality be sufficient in practice? As of today, we prefer methods like CORE-GPT [16] that *still* retrieves the actual library content.

Content Arrangement and Structuring. Content in a physical library is usually arranged by domain experts so that books are grouped based on similar topics. One person liked this way of exploring what else could be related to a specific article/book, e.g., by browsing through book/article titles with different terms or terminologies next to the initially found one in a physical library. Corresponding research and work for the digital world tries to mimic a physical library, e.g., Meghini et al. [12] demonstrated how narratives could be used to arrange content in Europeana.

4 Conclusion

This work summarized the key findings of our user study in digital humanities. We derived a list of requirements that need to be met when implementing a digital library system. We discussed how some requirements, but not all, can be met with existing methods.

References

[1] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Inf. Sci.* 509 (2020), 257–289. https://doi.org/10.1016/J.INS.2019.09.013

[2] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. YAKE! Collection-Independent Automatic Keyword Extractor. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings (Lecture Notes in Computer Science, Vol. 10772)*. Springer, 806–810. https://doi.org/10.1007/978-3-319-76941-7_80

[3] Schopenhauer Gesellschaft. 1912. *Schopenhauer-Jahrbuch*. Number Bd. 1. Verlag Waldemar Kramer. https://books.google.de/books?id=0-ZDAAAAIAAJ

[4] Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindrich Helcl, and Alexandra Birch. 2022. Survey of Low-Resource Machine Translation. *Comput. Linguistics* 48, 3 (2022), 673–732. https://doi.org/10.1162/COLI_A_00446

[5] Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. Association for Computational Linguistics, 2545–2568. https://doi.org/10.18653/V1/2021.NAACL-MAIN.201

[6] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12 (2023), 248:1–248:38. https://doi.org/10.1145/3571730

[7] M. Koßler, D. Birnbacher, and Verlag Königshausen & Neumann. 2023. *103. Schopenhauer Jahrbuch: für das Jahr 2022*. Königshausen & Neumann. https://books.google.de/books?id=WT5R0AEACAAJ

[8] Hermann Kroll, Christin K. Kreutz, Mathias Jehn, and Thomas Risse. 2024. Requirements for a Digital Library System: A Case Study in Digital Humanities (Technical Report). arXiv:2410.22358 [cs.DL] https://arxiv.org/abs/2410.22358

[9] Terry Kuny and Gary Cleveland. 1998. The Digital Library: Myths and Challenges. *IFLA Journal* 24, 2 (1998), 107–113. https://doi.org/10.1177/034003529802400205

[10] Mónica Marrero and Antoine Isaac. 2022. Implementation and Evaluation of a Multilingual Search Pilot in the Europeana Digital Library. In *Linking Theory and Practice of Digital Libraries - 26th International Conference on Theory and Practice of Digital Libraries, TPDL 2022, Padua, Italy, September 20-23, 2022, Proceedings (Lecture Notes in Computer Science, Vol. 13541)*. Springer, 93–106. https://doi.org/10.1007/978-3-031-16802-4_8

[11] Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2022. A Survey on Document-level Neural Machine Translation: Methods and Evaluation. *ACM Comput. Surv.* 54, 2 (2022), 45:1–45:36. https://doi.org/10.1145/3441691

[12] Carlo Meghini, Valentina Bartalesi, Daniele Metilli, and Filippo Benedetti. 2019. Introducing narratives in Europeana: A case study. *Int. J. Appl. Math. Comput. Sci.* 29, 1 (2019), 7–16. https://doi.org/10.2478/AMCS-2019-0001

[13] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. http://arxiv.org/abs/1301.3781

[14] Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive Machine Translation with Large Language Models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023*. European Association for Machine Translation, 227–237. https://aclanthology.org/2023.eamt-1.22

[15] Brian Ogilvie. 2016. Scientific Archives in the Age of Digitization. *Isis* 107 (03 2016), 77–85. https://doi.org/10.1086/686075

[16] David Pride, Matteo Cancellieri, and Petr Knoth. 2023. CORE-GPT: Combining Open Access Research and Large Language Models for Credible, Trustworthy Question Answering. In *Linking Theory and Practice of Digital Libraries: 27th International Conference on Theory and Practice of Digital Libraries, TPDL 2023, Zadar, Croatia, September 26-29, 2023, Proceedings (Lecture Notes in Computer Science, Vol. 14241)*. Springer, 146–159. https://doi.org/10.1007/978-3-031-43849-3_13

[17] Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural Machine Translation for Low-resource Languages: A Survey. *ACM Comput. Surv.* 55, 11 (2023), 229:1–229:37. https://doi.org/10.1145/3567592

[18] Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George F. Foster, and Gholamreza Haffari. 2024. Adapting Large Language Models for Document-Level Machine Translation. *CoRR* abs/2401.06468 (2024). https://doi.org/10.48550/ARXIV.2401.06468 arXiv:2401.06468