ORD-Xplore: Bridging Open Research Data Collections through Modality Abstractions

M. Sachdeva,[†1] M. Blum [1], Y. Stricker [2], T. Schreck [3], R. Mumenthaler [2,4], and J. Bernard[1,4]

[1]University of Zürich, Zürich, Switzerland; [2]University Library of Zürich, Zürich, Switzerland;
[3]Graz University of Technology, Graz, Austria; [4]Digital Society Initiative, Zürich, Switzerland

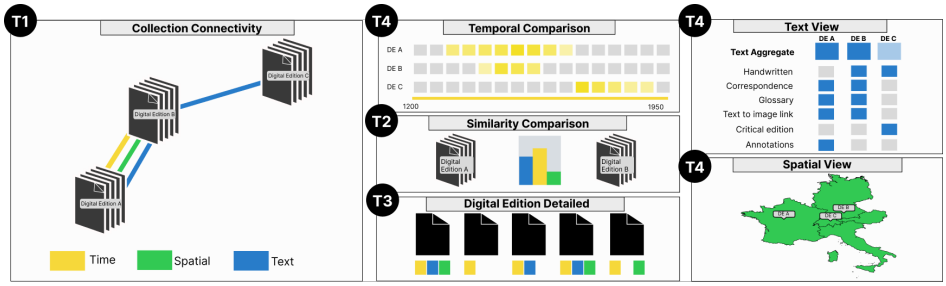

Figure 1: *The design prototype of ORD-Xplore shows how digital librarians will be enabled to explore connections between heterogeneous digital editions (collections of multiple digitized research artifacts). The four tasks include gaining an overview of connectivity across editions (T1), comparing connectivity between two editions (T2), analyzing details of an edition (T3), and analyzing modalities of the editions (T4).*

Abstract

We present ORD-Xplore, an approach to bridge gaps between digital editions, which represent valuable collections of multiple digitized research artifacts. However, digital editions often co-exist isolated, making it difficult for researchers to access, find, and re-use open research data from multiple digital editions. An ultimate goal is to unify library services across editions, even for editions with heterogeneity. In ORD-Xplore, we utilize abstraction methods from visualization research to help digital librarians identify unifying data modalities, as one important step towards standardization of heterogeneous digital editions.

CCS Concepts

• ***Human-centered computing*** → ***Information visualization; Visualization application domains;***

1. Introduction

Open Research Data (ORD) promotes the access and re-use of research data, providing transparency, knowledge sharing, and reproducibility of results [Sza12]. Today, a plethora of digital collections already exist, often used as platforms for research [HD22]. While aiming at meeting the FAIR principles (finding, accessing, interoperability, and reusing of research data), ORD practices of digital editions have different levels of ORD-readiness. Beyond stand-alone search support, not necessarily every edition adheres to (meta)data standards that would allow for the systematic *unification* through overarching digital library services [BBK*18]. Major challenges remain, in issues of interoperability, reuse, and long-term archiving [PT10]. To sum, the literature highlights the high degree of manual effort required from librarians and involved stakeholders, for data curation and standardization [WDA*16, DS20].

With ORD-Xplore, our goal is to enable digital librarians to visually identify structural connections between digital editions, to enable unified access, enhance FAIR principles, and allow long term usage. Abstractions [Mun09, SMM12] of edition domains, users, and edition data will reveal commonalities and differences across editions. This can provide a basis for deriving requirements, guidelines, and methods that better support the standardization and unification of digital editions. Our line of approach is to identify *modalities*, referring to abstract structural characteristics in edition data, recurring across multiple editions. We investigate, explore, and evaluate a comprehensive corpus of 14 digital editions, coordinated by the University of Zurich Center for Digital Editions (ZDE). Each digital edition holds a collection of multiple research artifacts, as core applications of digital humanities [JKR17]. The digital editions from ZDE vary in size ranging from a few to almost ten thousand documents, covering domains such as history, philology, law, and literature. In this poster, we share our abstraction process, including details and challenges of editions in the digital humanities, the primary stakeholder group, promising modalities for unification, an iteratively designed visualization prototype, and a discussion on workshop results conducted with 22 digital librarians and other domain experts.

† Corresponding Author

DOI: 10.2312/evp.20231065

2. Approach

2.1. Related Work and Problem Statement

The visualization community has seen pioneer works providing digital library support, e.g., in the context of digital humanities [SFC*15, EAGJ*16, BJP*19, BEAC*18], cultural heritage [ALC22,SZMV*15], multimedia databases [KKS*07], or scientific research data [SBS11, BRS*12a, BDF*15]. So far, most solutions supported the *within-edition* comparison [SFC*15, Mor05, CSV*16, BJP*19, ALC22], with a strong focus on text documents [LWC*19], e.g., for single text analysis, parallel text version analysis [BJP*19, JJW17, PMRM23], and within-corpus analysis [BGHE10, ALC20]. However, for the overarching challenge of unifying multiple editions of heterogeneous document types, fewer solutions exist [BBK*18], particularly from visualization research. Inspiration comes from approaches bridging gaps between two modalities, such as spatio-temporal analysis [AA06, AA13], time and metadata [BRS*12b, BRS*12a, ZS16], types of music documents [DFT*12], text and charts [LZK*22], or medical imaging and patient histories [BBJ*17]. In the context of digital editions, many corpora are curated according to the TEI-XML guidelines [FTM16]; a standard for the document annotation used in the digital humanities. ORD-Xplore will help in the unification process across editions, through abstractions of TEI-XML information.

2.2. User Characterization

The primary user group of ORD-Xplore is digital librarians who are tasked to organize, evaluate, and coordinate digital editions. Due to the high heterogeneity of editions, librarians often lack detailed knowledge that would help in the unification process.

2.3. Data Characterization

Modalities: We introduce a set of three modalities that re-occur in multiple editions, which can be derived from their TEI-XML data, as Figure 2 shows.

- **Text**: Refers to the (possible) textual nature of documents of an edition. We subdivide the text modality into a non-exhaustive list of types *Letter*, *Handwritten*, *Critical editions*, etc., to ease the identification of unifying aspects of documents across editions.
- **Temporal**: Describes the time-orientation of digital documents, if existing. Time will serve as an external primary key to align documents, even across editions
- **Spatial**: Describes edition documents according to their geographical alignment, if applicable. Enables the analysis of locations and spatial relationships across editions

2.4. Task Abstraction and Design Prototype

For the goal to support across-edition exploration, we draw four analysis tasks for digital librarians.

- (T1) **Connectivity Overview**: explore by which modalities digital editions have unifiable structures with little curation effort
- (T2) **Pairwise edition comparison**: compare multiple modalities and unification possibilities between two digital editions
- (T3) **Edition details**: analyze details of multiple modalities for a single digital edition in focus, at the granularity of documents
- (T4) **Modality Exploration**: explore a single modality across digital editions, for abstracted modalities (Text, Time, Spatial)

We present a design sketch of the prototype that will support all four tasks, shown in Figure 1. Collection Connectivity (T1)

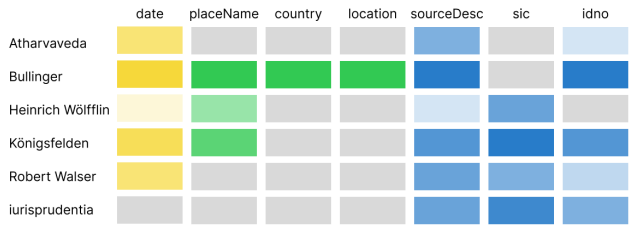

Figure 2: *Analysis of the heterogeneous use of tags and deriving their modalities from the TEI-XML data across six digital editions*

encodes digital editions as nodes and shared modalities as colored edges. Also, spatial distance of editions in the graph is a strong indicator for edition connectivity. In turn, this also enables the identification of outlying editions that lack connectivity through shared modalities. The Similarity Comparison View (T2) uses colored bar charts to indicate the coverage of shared modalities between two editions, for all entailed documents. The Edition Detailed View (T3) uses glyphs to encode the expression of modalities per document and digital edition. The Temporal Comparison View (T4) uses a heatmap with color saturation to indicate temporal alignments across editions. The Text View (T4) shows the presence of different text categories in a heatmap with color saturation. In addition, the text aggregate computes the across-edition compatibility, i.e., strong connections between editions. Finally, the Spatial View (T4) encodes the geo-referenced locations across-digital editions.

3. Preliminary Results, Discussion, and Conclusion

3.1. Workshop with Experts and Outcomes

We presented the modality abstractions and design prototype of ORD-Xplore to 22 digital librarians and other domain experts, using the 14 digital editions as the example case; predominantly with positive responses. From the workshop, we take four findings away: First, the majority of experts confirmed the need for edition unification, as well as non-trivial identified challenges, both indicating that our approach is taking the right track. Second, the experts strongly agreed on the importance of the three abstracted modalities to connect the dots between digital editions. Third, constructive discussions offered room for yet other modalities that can be considered, including *Persons* and *Named Entities* in general. Finally, some experts with more fine-grained data curation expertise reinforced that not only *across* editions, but also *within* editions, standardization is difficult. From a bottom-up perspective, editions often are heterogeneous and ambiguous in terms of tag usage, which calls for a more fine-grained analysis at the granularity of tags in TEI-XML.

3.2. Conclusion and Critical Remarks

We presented ORD-Xplore, an approach aimed at bridging gaps between heterogeneous digital editions through modality abstractions, to support digital librarians in the edition-unification process. A limitation is that our high-level top-down perspective does not consider the diverse types of data content within individual digital editions, but remains at a structural level. In addition to the promising top-down approach, we plan to also investigate bottom-up tag-based analysis, possibly augmented with content-based edition abstractions. Another possible extension of the scope could be a stronger focus on data curators, as a second user group. By doing so, we can help towards solving the gaps in metadata and ultimately benefit unification efforts.

References

[AA06] ANDRIENKO N. V., ANDRIENKO G. L.: *Exploratory analysis of spatial and temporal data - a systematic approach*. Springer, 2006. doi:10.1007/3-540-31190-4. 2

[AA13] ANDRIENKO N., ANDRIENKO G.: Visual analytics of movement: An overview of methods, tools and procedures. *Information visualization 12*, 1 (2013), 3–24. 2

[ALC20] ALHARBI M., LARAMEE R. S., CHEESMAN T.: Transvis: integrated distant and close reading of othello translations. *Transactions on Computer Graphics (TVCG) 28*, 2 (2020), 1397–1414. 2

[ALC22] ALHARBI M., LARAMEE R. S., CHEESMAN T.: Transvis: Integrated distant and close reading of othello translations. *Transactions on Computer Graphics (TVCG) 28*, 2 (2022), 1397–1414. doi:10.1109/TVCG.2020.3012778. 2

[BBJ*17] BANNACH A., BERNARD J., JUNG F., KOHLHAMMER J., MAY T., SCHECKENBACH K., WESARG S.: Visual analytics for radiomics: Combining medical imaging with patient data for clinical research. In *IEEE Workshop on Visual Analytics in Healthcare (VAHC)* (2017), IEEE, pp. 84–91. doi:10.1109/VAHC.2017.8387545. 2

[BBK*18] BLEIER R., BÜRGERMEISTER M., KLUG H. W., NEUBER F., SCHNEIDER G.: *Digital Scholary Editions as Interfaces*, vol. 12. BoD–Books on Demand, 2018. 1, 2

[BDF*15] BERNARD J., DABERKOW D., FELLNER D., FISCHER K., KOEPLER O., KOHLHAMMER J., RUNNWERTH M., RUPPERT T., SCHRECK T., SENS I.: Visinfo: a digital library system for time series research data based on exploratory search – a user-centered design approach. *International Journal on Digital Libraries (IJoDL) 16*, 1 (2015). doi:10.1007/s00799-014-0134-y. 2

[BEAC*18] BRADLEY A. J., EL-ASSADY M., COLES K., ALEXANDER E., CHEN M., COLLINS C., JÄNICKE S., WRISLEY D. J.: Visualization and the digital humanities. *IEEE computer graphics and applications 38*, 6 (2018), 26–38. 2

[BGHE10] BÜCHLER M., GESSNER A., HEYER G., ECKART T.: Detection of citations and textual reuse on ancient greek texts and its applications in the classical studies. In *DH* (2010), pp. 113–114. 2

[BJP*19] BAUMANN M., JOHN M., PFLÜGER H., HERBERICHS C., VIEHHAUSER G., KNOPKI W., ERTL T.: An interactive visualization for the analysis of annotated text variance in the legendary der heiligen leben, redaktion. In *Symp. on Visualization in Applications* (2019). 2

[BRS*12a] BERNARD J., RUPPERT T., SCHERER M., KOHLHAMMER J., SCHRECK T.: Content-based layouts for exploratory metadata search in scientific research data. In *ACM/IEEE-CS Joint Conference on Digital Libraries* (New York, NY, USA, 2012), ACM, pp. 139–148. doi:10.1145/2232817.2232844. 2

[BRS*12b] BERNARD J., RUPPERT T., SCHERER M., SCHRECK T., KOHLHAMMER J.: Guided discovery of interesting relationships between time series clusters and metadata properties. In *Knowledge Management and Knowledge Technologies* (2012), ACM, pp. 22:1–22:8. doi:10.1145/2362456.2362485. 2

[CSV*16] CASTERMANS T., SPECKMANN B., VERBEEK K., WESTENBERG M. A., BETTI A., VAN DEN BERG H., ET AL.: Glammap: geovisualization for e-humanities. In *Workshop on Visualization for the Digital Humanities (Vis4DH)* (2016). 2

[DFT*12] DAMM D., FREMEREY C., THOMAS V., CLAUSEN M., KURTH F., MÜLLER M.: A digital library framework for heterogeneous music collections: from document acquisition to cross-modal interaction. *International Journal on Digital Libraries 12* (2012), 53–71. 2

[DS20] DÄNGELI P., STUBER M.: Nachhaltigkeit in langjährigen erschliessungsprojekten. fair-data-kriterien bei editions-und forschungsplattformen zum 18. jahrhundert. *xviii. ch: Jahrbuch der Schweizerischen Gesellschaft zur Erforschung des 18. Jahrhunderts 11* (2020), 34–51. 1

[EAGJ*16] EL-ASSADY M., GOLD V., JOHN M., ERTL T., KEIM D. A.: Visual text analytics in context of digital humanities. In *IEEE VIS Workshop on Visualization for the Digital Humanities as part of the IEEE VIS* (2016). 2

[FTM16] FRANZINI G., TERRAS M., MAHONY S.: A catalogue of digital editions. *Digital scholarly editing: Theories and practices* (2016), 161–182. 2

[HD22] HANSSON K., DAHLGREN A. N.: Open research data repositories: Practices, norms, and metadata for sharing images. *J. Assoc. Inf. Sci. Technol. 73*, 2 (2022), 303–316. doi:10.1002/asi.24571. 1

[JJW17] JÄNICKE S., JOSEPH WRISLEY D.: Visualizing mouvance: Toward a visual analysis of variant medieval text traditions. *Digital Scholarship in the Humanities 32*, suppl_2 (2017), ii106–ii123. 2

[JKR17] JANNIDIS F., KOHLE H., REHBEIN M.: *Digital Humanities*. Springer, 2017. 1

[KKS*07] KROTTMAIER H., KURTH F., STEENWEG T., APPELRATH H.-J., FELLNER D.: Probado–a generic repository integration framework. In *European Conference on Digital Libraries* (2007), Springer, pp. 518–521. 2

[LWC*19] LIU S., WANG X., COLLINS C., DOU W., OUYANG F., EL-ASSADY M., JIANG L., KEIM D. A.: Bridging text visualization and mining: A task-driven survey. *Transactions on Computer Graphics (TVCG) 25*, 7 (2019), 2482–2504. doi:10.1109/TVCG.2018.2834341. 2

[LZK*22] LATIF S., ZHOU Z., KIM Y., BECK F., KIM N. W.: Kori: Interactive synthesis of text and charts in data documents. *Transactions on Computer Graphics (TVCG) 28*, 1 (2022), 184–194. doi:10.1109/TVCG.2021.3114802. 2

[Mor05] MORETTI F.: *Graphs, maps, trees: abstract models for a literary history*. Verso, 2005. 2

[Mun09] MUNZNER T.: A nested process model for visualization design and validation. *Transactions on Computer Graphics (TVCG) 15*, 6 (2009), 921–928. doi:10.1109/TVCG.2009.111. 1

[PMRM23] PÖCKELMANN M., MEDEK A., RITTER J., MOLITOR P.: Lera—an interactive platform for synoptical representations of multiple text witnesses. *Digital Scholarship in the Humanities 38*, 1 (2023), 330–346. 2

[PT10] PARK J.-R., TOSAKA Y.: Metadata creation practices in digital repositories and collections: Schemata, selection criteria, and interoperability. *Information Technology and Libraries 29*, 3 (2010), 104–116. 1

[SBS11] SCHERER M., BERNARD J., SCHRECK T.: Retrieval and exploratory search in multivariate research data repositories using regressional features. In *ACM/IEEE Joint Conference on Digital Libraries* (2011), ACM, pp. 363–372. doi:10.1145/1998076.1998144. 2

[SFC*15] STEFAN J., FRANZINI G., CHEEMA M. F., GERIK S., ET AL.: On close and distant reading in digital humanities: A survey and future challenges. In *Eurographics Conference on Visualization (EuroVis)* (2015), R. Borgo, F. Ganovelli, I. Viola, pp. N–A. 2

[SMM12] SEDLMAIR M., MEYER M. D., MUNZNER T.: Design study methodology: Reflections from the trenches and the stacks. *Transactions on Computer Graphics (TVCG) 18*, 12 (2012), 2431–2440. doi:10.1109/TVCG.2012.213. 1

[Sza12] SZALAY A.: Data-intensive discoveries in science: the fourth paradigm. In *Workshop on Data-Intensive Distributed Computing* (2012), ACM, pp. 1–2. doi:10.1145/2286996.2286998. 1

[SZMV*15] SALVADOR A., ZEPPELZAUER M., MANCHON-VIZUETE D., CALAFELL A., GIRO-I NIETO X.: Cultural event recognition with visual convnets and temporal models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2015), pp. 36–44. 2

[WDA*16] WILKINSON M. D., DUMONTIER M., AALBERSBERG I. J., APPLETON G., AXTON M., BAAK A., BLOMBERG N., BOITEN J.-W., DA SILVA SANTOS L. B., BOURNE P. E., ET AL.: The fair guiding principles for scientific data management and stewardship. *Scientific data 3*, 1 (2016), 1–9. 1

[ZS16] ZEPPELZAUER M., SCHOPFHAUSER D.: Multimodal classification of events in social media. *Image and Vision Computing 53* (2016), 45–56. 2