CORDI_ Conference on Research Data infrastructure

Postersession

https://doi.org/10.5281/zenodo.10074281



Published: 2023-11-06

# Lessons learned from joining forces across disparate disciplines in the NFDI

## The Conference on Research on Text Analytics

Ulrich Krieger[1][https://orcid.org/0000-0001-6705-7464] and Thorsten Trippel[2][https://orcid.org/0000-0002-7211-7393]

[1] BERD@NFDI, University of Mannheim, Germany

[2] Text+, Leibniz-Institut für Deutsche Sprache, Mannheim, Germany

[3]

**Abstract.** This contribution summarizes the lessons learned from the organization of a joint conference on text analytics research by the Business, Economic, and Related Data (BERD@NFDI) and Text+ consortia within the National Research Data Infrastructure (NFDI) in Germany. The collaboration aimed to identify common ground and foster interdisciplinary dialogue between scholars in the humanities and in the business domain. The lessons learned include the importance of presenting research questions using textual data to establish common ground, similarities in methodology for processing textual data between the consortia, similarities in research data management, and the need for regular interconsortial discussions on textual analysis methods and data. The collaboration proved valuable for interdisciplinary dialogue within the NFDI, and further collaboration between the consortia is planned.

**Keywords:** Text Analytics, Business Data, Economic Data, Text, Humanities, Business Research, Machine Learning, NFDI

## 1. Introduction

In the National Research Data Infrastructure (NFDI) in Germany, text data serves as a crucial source for analysis in the consortia dealing with humanities research (Text+) and Business, Economic, and Related Data (BERD@NFDI). Despite apparent differences between these consortia, closer examination reveals shared interests in text analytics methods, as well as diverse research questions pursued by scholars in each consortium. Recognizing the commonalities, both BERD@NFDI and Text+ collaborated to organize a conference on text analytics research (see https://events.gwdg.de/event/435/, visited 2023-07-06), while acknowledging the different research questions addressed by the scholars in each consortium. This poster abstract presents the lessons learned from the establishment of this conference.

## 2. Lesson 1: Building Common Ground through the Presentation of Research Questions Applying Textual Data

Establishing common ground between BERD@NFDI and Text+ involved presenting research questions using textual data. Concrete research questions provided insights into the respective domains without requiring introductory explanations of each discipline. These questions highlighted the applied methods for data creation, elicitation strategies, and analysis workflows. Methods include all aspects of data annotation, including natural language

processing tools. Topics of discussion ranged from standard processing, such as Named Entity Recognition, to novel research problems, such as differentiating between human and machine-generated texts used in research.

## 3. Lesson 2: Similarities in the Consortia in Manual Annotation and Automatic Processing of Textual Data

Both consortia, BERD@NFDI and Text+, employ similar tools for manual data annotation and automatic data processing. These tools include AI technologies, such as machine learning and natural language processing, as well as annotation tools. Common methods, such as measuring annotation quality and approaching deviations from annotations, are also shared. Comparable processes were observed in data preparation for textual and non-textual content from advertisements used in marketing and linguistic studies using newspaper texts.

## 4. Lesson 3: Similarities in Research Data Management Functionalities and Processes

Despite differences in user interface, data categories, and research data, BERD@NFDI and Text+ share similar methods for eliciting, processing, and managing research data. Repository systems and catalog systems are utilized for research data management, enabling data archiving, access, and metadata descriptions. While not all data can be publicly offered for download due to confidentiality concerns (proprietary data, publishing rights), catalog systems provide information on data sets and access conditions. There is potential for metadata exchange between the consortia, facilitating the exchange of ideas between both research communities.

## 5. Lesson 4: Regular Interconsortial Discussion on Textual Analysis Methods and Data

The unexpected quality of the insights generated highlights the importance of platforms for exchanging ideas on textual analysis. Such platforms enable the identification of additional synergies, the sharing of methodology experience, and discussions on cooperation. This interdisciplinary exchange should encompass analytical techniques as well as research data management.

## Conclusion

In conclusion, the collaboration between BERD@NFDI and Text+ in organizing the Joint Conference on Text Analytics provided valuable lessons for interdisciplinary dialogue within the NFDI. This collaboration will be continued.

## Funding

This publication was created in the context of the German National Research Data Infrastructure (NFDI) e.V. The authors gratefully acknowledge the NFDI funding by the Federal Republic of Germany and the 16 federal states, the consortium BERD@NFDI is funded by the Deutsche
Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460037581, and the consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460033370. Furthermore, we would like to thank all institutions and actors who are committed to the NFDI and its goals. The conference reported here was made possible by additional funding from the University of Mannheim Foundation.

## Acknowledgement

The conference was organized by the authors together with Marc Fischer (University of Cologne), Veronica Haaß (University of Mannheim), Mark Heitmann (University of Hamburg), Gerhard Heyer (Saxonian Academy of Sciences and Humanities), Erhard Hinrichs (Leibniz-Institute of the German Language and University of Tübingen), Regina Jutz (University of Mannheim), Peter Leinen (German National Library, DNB), Florian Stahl (University of Mannheim), and Andreas Witt (Leibniz-Institute of the German Language). The responsibility for any errors in this contribution lies with the authors and not with the mentioned co-organizers. The authors would also like to thank the presenters at the *Joint Conference on Research on Text Analytics.*