BERD@BW – A Science Data Center to foster Open Science in Business, Economics and Social Sciences

Sabine Gehrlein, Irene Schumm and Renat Shigapov

Mannheim University Library, University of Mannheim

The Center for Business, Economic and Related Data in Baden-Württemberg(BERD@BW) is one of the four science data centers funded by the Ministry of Science, Research and Arts of Baden-Württemberg within the digitization strategy "digital@bw". BERD@BW is aimed to improve sharing, finding and reusing unstructured and semi-structured research data in the social sciences in accordance with the FAIR principles (findable, accessible, interoperable and reusable). BERD@BW is built by the University of Mannheim and the Leibniz Center for European Economic Research (ZEW). Both institutions are experienced in infrastructure projects and in the empirical social sciences, including business and economics. BERD@BW is based on four pillars: 1) building up methodological knowledge, 2) developing tools and services dealing with unstructured and semi-structured data, 3) training and consulting with respect to legal and technical issues in research data management, and 4) engaging in national and international networking. The services and materials developed within BERD@BW are available as openly as possible on the project homepage: `https://www.berd-bw.de`.

1 Introduction

The social sciences, including business studies and economics, have a long tradition of handling structured research data in standardized mainly-tabular forms with proper metadata. These datasets often origin from public administration processes or surveys. In many cases they are sensitive and restricted in use. Infrastructure institutions (e.g., research data centers[1] and libraries) and commercial providers make the data available for academic purposes under clearly specified licenses and guaranteed authenticity. Data access and research methods with structured data are well established in both teaching and research.

Unstructured and semi-structured data, on the other hand, pose new challenges for data management and research. The licenses are harder to specify. The data authenticity

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: `https://doi.org/10.11588/heidok.00029645` veröffentlicht.

[1] organized by KonsortSWD, `https://www.konsortswd.de/en/ratswd/`, retrieved May 14, 2021.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13743 (CC BY-SA 4.0)

is trickier to guarantee. Unstructured data come in a variety of forms (pictures, text, video, and audio) from many non-standard sources (social media, web pages, and mobile phones) and in very large volumes ("big data"). To make the datasets findable, accessible, interoperable and reusable (FAIR),[2] new technologies have to be adapted and skilled staff is needed to implement suitable solutions.

To face these challenges, the Center for Business, Economic and Related Data in Baden Württemberg (BERD@BW in the following) was established by partners from infrastructure (Mannheim University Library, University-IT Mannheim, and research data center of ZEW) and research (Prof. Stahl and Prof. Kreuter from the Mannheim Center for Data Science, MCDS, and Dr. Licht from ZEW). BERD@BW is based on four pillars: 1) collecting methodological competence for modern data analysis, 2) developing the tools and services for collecting, analyzing and archiving unstructured and semi-structured data, 3) training and consulting on legal and technical aspects of research data management, and 4) networking. In the following, we describe these pillars and draw conclusions.

2 The Pillars of BERD@BW

Methodological competence. According to the publications in the top business and economic journals, unstructured data has become more and more important within the last ten years [1]. However, among the social scientists the data science and coding competencies needed to process and analyze unstructured and semi-structured data are not yet broadly spread. To improve those data science competencies among the target group of BERD@BW, knowledge about the methods and algorithms was collected and provided in an easily approachable form[3]. Machine learning, artificial intelligence and natural language processing are clearly explained without the use of too much technical jargon. A benchmarking system for algorithms dealing with unstructured data is currently under development.

Tools and services. Various tools and services are developed in BERD@BW to collect, process, archive and link semi-structured and unstructured data and to make the data available in compliance with the FAIR principles. A web-scraping engine "ARGUS"[4] is developed in order to collect unstructured data from non-standard sources. A corpus of websites of German companies is built using ARGUS [2-4] by scraping the web pages twice a year. The dynamically growing corpus is used as a prototype for archiving and providing unstructured data.[5] Social scientists are used to deal with sensitive data provided by research data centers. Access to sensitive data may be restricted due to privacy regulations and confidentiality obligations. A typical way to access the data is using a guest workstation on the premises of the respective research data center. Since the start of

[2] `https://www.go-fair.org/fair-principles/`, retrieved May 14, 2021.
[3] `https://digitaleconomy.org`, retrieved May 14, 2021.
[4] Available on GitHub
[5] The corpus „Mannheimer Webpanel" is provided by the research data center of the ZEW, see `https://kooperationen.zew.de/zew-fdz/datenangebot/mannheimer-webpanel.html`.

the Corona pandemic, physical access to many research data centers has been very limited or not possible at all. Therefore, the "BERD desktop" was developed for a secure remote access to sensitive data. The BERD desktop is a remote desktop server using a docker container and a virtual machine. To protect the sensitive data against unauthorized use, the configuration of the BERD desktop prohibits changes in system configuration, internet access, and uploads as well as downloads by the user. Data transfer into and out of the BERD desktop works only through the data administrator of the research data center. To enable users to run their analyses, the most popular tools for data analysis in the social sciences (i.e., R, Python and Stata) are provided. The first prototype of the BERD desktop in the bwCloud[6] is implemented. First tests of external users with data from the research data center of the ZEW are currently made.

Many datasets in the social sciences are still not findable, accessible, interoperable, and reusable. To improve this at least partially, a knowledge graph-based research data management infrastructure for German company datasets is developed. The Wikibase software was chosen as a backend for the infrastructural services. A test frontend in Python and JavaScript was created. Valuable historic datasets were digitized, OCR-ed (optical character recognition) and structured [5]. To speed up data integration and knowledge graph construction with Wikibase, the open source tool "RaiseWikibase" is implemented [6]. For automatic annotation services the open source semantic annotator "bbw" was designed and coded in Python [7]. The annotator performs named entity linking, property matching and type linking for tabular data without metadata using a Wikibase knowledge graph.

Training and consulting. Based on the collected knowledge and using the tools and services created, several training and consulting measures (both synchronous and asynchronous) have been initiated. The online micro workshop series "Data Literacy Snacks", which takes place during lunchtime for free, is offered as a part of synchronous training.[7] The events provide introductions into several topics of research data management tailored to the target group of data scientists and empirical researchers in business, economics and the social sciences. The first series comprises sessions about reproducible research, privacy regulations for research data and knowledge graphs in research data management. The number of registrations and participants exceeded expectations and personal feedback was very positive so far. As a part of asynchronous training and consulting the "interactive Virtual Assistant" (iVA) was developed [8] which guides users through different topics in research data management. For example, the first iVA implementation provides an interactive introduction into privacy law [8].

The iVA concept is currently under evaluation and will be rolled out to other topics relevant for the BERD community. The iVA implementations will be available as Open Educational Resources. Furthermore, a concept for an (a)synchronous online course "Good Practices for Managing Data" is developed in BERD@BW. While many discipline-specific courses in the area of data science and empirical research focus on methodology, this course

[6] `https://www.bw-cloud.org/`, retrieved May 14, 2021
[7] `https://www.berd-bw.de/snacks/`, retrieved May 14, 2021

specifically targets governance questions in research data management. The content of the course is mainly based on the Train-the-Trainer concept in research data management [9] and the concepts published by the UK Data Service [10].

Networking. A major aim of the funding program for the science data centers was fostering community-based networking and involvement into the National Research Data Infrastructure (Nationale Forschungsdateninfrastruktur, NFDI).[8] The NFDI is a nationwide initiative of all German States and the Federal Government for building up and connecting discipline-specific research data networks in order to improve research data management within those disciplines, especially long-term storage, backup and accessibility.[9] The discipline-specific networks shall work together on cross-cutting topics and engage in international research data management initiatives.

One consortium covering the social sciences is KonsortSWD, which mainly focuses on sensitive structured data managed by the research data centers accredited by the RatSWD. Based on BERD@BW, the complementing consortium BERD@NFDI[10] was formed and organized with the main focus on unstructured data. More highly-recognized infrastructure institutions joined BERD@NFDI: ZBW (Leibniz Information Centre for Economics) and GESIS (Leibniz Institute for the Social Sciences). They are the most relevant providers of research data and information infrastructure in business, economics and the social sciences in Germany.[11] IT resources (storage and computing power) are provided by the LRZ (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities), which is world-famous for its high-performance computing resources. In addition, more well-known research institutions dealing with unstructured data in business, economics and the social sciences joined the consortium. Among them are the Universität Hamburg, the University of Cologne, and the Ludwig-Maximilians-Universität München.

It is worth to mention that the co-spokesperson Prof. Bernd Bischl is a co-founder of OpenML[12], which could become a groundwork for a FAIR machine learning and data analysis platform within BERD@NFDI. The partners of BERD@NFDI are important players in international and subject-specific research data networks such as GoFAIR, the European Open Science Cloud, and the Coleridge Initiative. Therefore, the connections to these initiatives are established as well.

[8] `https://mwk.baden-wuerttemberg.de/de/service/presse-und-oeffentlichkeitsarbeit/pressemitteilung/pid/vier-science-data-centers-in-baden-wuerttemberg/`, retrieved May 14, 2021.

[9] `https://www.dfg.de/en/research_funding/programmes/nfdi/index.html`, retrieved May 14, 2021.

[10] `https://www.berd-nfdi.de/`, retrieved May 14, 2021.

[11] ZBW and GESIS run the DOI registration agency for Social Sciences research data da|ra, for example, `https://www.da-ra.de/`, retrieved May 14, 2021.

[12] `https://www.openml.org/`, retrieved May 14, 2021.

3 Conclusions

In this paper, we sketched the foundations of the science data center BERD@BW. The contributions to FAIR and open research data management in business, economics and related domains are described. Apart from collecting and processing the subject-specific data science knowledge, the tools and services, which improve collecting, storing, reusing, and providing unstructured and semi-structured data, are presented. The training and consulting concepts, developed to improve research data handling, are introduced. The sustainable development of BERD@BW is guaranteed through embedding it into the new national consortium BERD@NFDI (https://www.berd-nfdi.de). An extended international networking is expected within BERD@NFDI.

Acknowledgements

The Business and Economic Research Data Center (BERD@BW) is funded within the digitization strategy "digital@bw" by the Ministry of Science, Research and Arts of Baden-Württemberg, Germany.

Bibliography

[1] Bayerl, A., Kluge, S., Beichert, M., Stahl, F. Methods and applications of Data Science in the context of Business & Economics. Available at DigitalEconomy.org (2020), https://tinyurl.com/3vjmx9j5.

[2] Kinne, J. and D. Lenz. Predicting Innovative Firms Using Web Mining and Deep Learning, ZEW Discussion Paper No. 19-001 (2019), Mannheim, https://doi.org/10.1371/journal.pone.0249071.

[3] Krüger, M., Kinne, J., Lenz, D., Resch, B.:The Digital Layer: How Innovative Firms Relate on the Web (2020). ZEW - Centre for European Economic Research Discussion Paper No. 20-003, Available at SSRN: https://ssrn.com/abstract=3530807.

[4] Kinne, J., and Axenbeck, J. Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. Scientometrics 125, 2011–2041 (2020). https://doi.org/10.1007/s11192-020-03726-9.

[5] Gehrlein, S., Kamlah, J., Pintsch, M., Schumm, I., Weil, S.: Vom Papier zur Datenanalyse. "Neue" historische Forschungsdaten für die Wirtschaftswissenschaften. In: Heuveline, V. (ed.) E-Science-Tage 2019 : Data to Knowledge. vol. 598, pp. 140-152. heiBOOKS (2020), https://doi.org/10.11588/heibooks.598.c8423.

[6] Shigapov, R., Mechnich, J., Schumm, I. RaiseWikibase: Fast inserts into the BERD instance. ESWC2021 Poster and Demo Track (2021).

[7] Shigapov, R., Zumstein, P., Kamlah, J., Oberländer, L., Mechnich, J., Schumm, I. bbw: Matching CSV to Wikidata via Meta-lookup. In: SemTab at ISWC 2020. vol. 2775, pp. 17-26 (2020), http://ceur-ws.org/Vol-2775/paper2.pdf.

[8] Herklotz, M., and Oberländer, L. iVA: Ein interaktiver Virtueller Assistent von BERD@BW zur Aufbereitung von Rechtsfragen im Bereich Open Science. E-Science-Tage 2021: Share your research data (2021).

[9] Biernacka, K., Bierwirth, M., Buchholz, P., Dolzycka, D., Helbig, K., Neumann, J., Odebrecht, C., Wiljes, C., Wuttke, U. Train-the-Trainer Concept on Research Data Management (Version 3.0). Zenodo (2020), http://doi.org/10.5281/zenodo.4071471.

[10] Corti, L., Van den Eynden, V., Bishop, L. and M. Woollard: Managing and Sharing Research Data – A Guide to Good Practice. SAGE (2020), http://hdl.handle.net/11329/297.