Fourth Workshop on Big (and Small) Data in Science and Humanities (BigDS 2023)

Andreas Henrich[1], Naouel Karam[2], Birgitta König-Ries[3], Bernhard Seeger[4]

In the last 20 years, we have seen a continuous digital transformation in science, society, and economy. The growth of the internet and advancements in data collection have resulted in the era of Big Data, marked by a massive and continually growing amount of complex, interconnected, and heterogeneous data. Earth observation sensors, for instance, produce petabytes of data with improved spectral, temporal, and spatial precision. Social media users generate high volumes of content. The information and knowledge contained in these data have huge potential value, which, if uncovered, could aid in improving our understanding of complex systems such as earth and society, drive innovation, and empower well-informed decisions.

Thus, the importance of data has increased dramatically not only in business, but also in almost all scientific disciplines, e.g., in meteorology, genomics, complex physical simulations, bio- and environmental research, and more recently in the humanities. This led in particular to the creation of the unique NFDI (National Research Data Infrastructure), which aims to "systematically manage scientific and research data, provide long-term data storage, backup and accessibility, and network the data both nationally and internationally"[5]. NFDI began with more than 25 domain-specific consortia and projects in the area of basic infrastructure, covering a broad range of scientific disciplines from cultural sciences, humanities and engineering to life and earth sciences.

The availability of such a large volume of multidisciplinary data within NFDI and beyond leads to a rethinking in scientific disciplines on how to extract relevant information and on how to foster research. Researchers face severe challenges in leveraging data, since appropriate data management, integration, analysis and visualization tools have not been available so far. Recent advances in the development of big data technologies and the progress in machine learning and semantic technologies allow for a better computational support to deal with large amounts of heterogeneous data, and offer flexible end-to-end analytic and

[1] University of Bamberg, Media Informatics, 96047 Bamberg, Germany andreas.henrich@uni-bamberg.de
[2] Fraunhofer FOKUS & InfAI e.V., Berlin, Germany karam@infai.org
[3] University of Jena, Heinz Nixdorf Chair for Distributed Information Systems, 07743 Jena, Germany birgitta.koenig-ries@uni-jena.de
[4] University of Marburg, Department of Mathematics and Computer Science, 35032 Marburg, Germany seeger@informatik.uni-marburg.de
[5] `https://www.dfg.de/en/research_funding/programmes/nfdi/index.html`

doi:10.18420/BTW2023-41

visualization solutions for various application domains. A critical prerequisite for achieving those goals is the availability of data that is harmonized and made reusable in a sustainable and qualitative manner. This needs to be realized following the FAIR data principles, the fundamental concepts that aim to improve findability, accessibility, interoperability, and reusability of research data[6].

The need to discuss real-world problems in data science as well as the recent advances in big data technology between database researchers and scientists from various disciplines already led to the first three editions of the workshop on Big (and Small) Data in Science and Humanities (BigDS) at BTW 2015, 2017, and 2019. This year's fourth edition of the BigDS workshop co-located with the 20th Conference on Database Systems for Business, Technology and Web (BTW) accommodates the continuously growing interest in methods to efficiently and effectively manage and analyze Big Data. With workshop contributions from various disciplines, we hope to promote the dialog between domain experts and data scientists and to foster the engagement of the database community to NFDI and other important infrastructure projects.

The workshop program kicked off with Markus Stocker, who gave an inspiring keynote on machine actionable scientific information. He introduced the basic concepts, discussed the challenges, and pointed out the great opportunities for producing and sharing knowledge in research and society.

We further selected eight contributions that address different challenges in the context of data-driven processing and analytics. The papers contribute to broadly applicable technologies like provenance for spreadsheets, management and integration of geo-spatial data, ontologies, trust in AI, and user interfaces. The proposed approaches are applicable to various domains, such as ecology and digital humanities.

Two papers focus on basic methods and systems for data processing. Müller and Mertová addressed the provenance problem of data transformations in spreadsheets. Their approach creates a copy of the source data in a new worksheet and performs data transformations on this copy while referring back to the original sheet. Beilschmidt et al. presented the basic concepts of Geo Engine, a new spatio-temporal processing infrastructure that has been used in several ecological projects, including NFDI4Biodiversity. Their workflows with a spatio-temporal context offer great potential and flexibility for many applications.

There are two papers addressing data extraction and data integration in scientific applications. Bartsch et al. described an extraction process of various digital objects from different sources to create a multimodal corpus for the analysis of climate change publications. Using a variety of tools, they manage to extract images, graphs, tables or videos and annotate text. Jegan et al. described an approach to support information integration and improving the data quality by using multiple external information sources to facilitate disambiguation of geographic data. The resulting system will be used within the infrastructure of the NFDI project Text+.

[6] https://www.go-fair.org/fair-principles/

As more data is available, dataset discovery is a frequent task in daily research practice. Thus, the question of user interfaces becomes more important that is addressed in the following two contributions. Löffer et al. proposed a semantic search for biological datasets. The authors evaluated two kinds of search interfaces including free text, categories and annotating of returned research results. Their results show that users prefer interfaces with a single input field for search tasks and appreciate explanations of the results. Schildgen et al. reported on an Alexa-based NLP tool to facilitate natural language querying of databases using SQL. This also includes the translation of the query result (which is always a table) into text and voice. Such kind of easy-to-use interfaces would widely facilitate the interaction with scientific databases.

The work of Abdelmageed et al. addressed the problem of knowledge transfer on ontologies and data integration towards a concrete application. The authors developed an agricultural core ontology that is used to link general concepts to more domain-specific concepts. Bruchhaus et al. presented how trust can be introduced into big data analysis and AI. Their prototype offers a so-called trust-bus as a component in a microservice architecture.

All contributions to this year's BigDS workshop provide new domain-relevant insights and promote the use of generic as well as domain-specific methods for scientific data management and analysis. We want to thank everyone who contributed to the workshop, especially the authors, the keynote speaker Markus Stocker, the BigDS program committee, the BTW team, and all the participants. We are grateful to NFDI4Biodiversity for its financial support of the workshop.

Workshop Organizers

Andreas Henrich (Univ. Bamberg)
Naouel Karam (Fraunhofer FOKUS & InfAI e.V.)
Birgitta König-Ries (Univ. Jena)
Bernhard Seeger (Univ. Marburg)

Program Committee

Alsayed Algergawy (Univ. Jena)
Thomas Brinkhoff (FH Oldenburg)
Michael Diepenbroek (GFBio e. V., Bremen)
Jana Diesner (University of Illinois at Urbana-Champaign)
Michael Gertz (Univ. Heidelberg)
Anika Groß (Hochschule Anhalt)
Anton Güntsch (Botanischer Garten und Botanisches Museum, Berlin)
Dominik Hezel (Univ. Frankfurt)

Alfons Kemper (TU München)
Toralf Kirsten (Univ. Leipzig)
Meike Klettke (Univ. Regensburg)
Ulf Leser (HU Berlin)
Richard Lenz (Univ. Erlangen)
Ulrike Lucke (Univ. Potsdam)
Bertram Ludäscher (University of Illinois at Urbana-Champaign)
Manja Marz (Univ. Jena)
Wolfgang Müller (HITS, Heidelberg)
Thorsten Papenbrock (Univ. Marburg)
Kai-Uwe Sattler (TU Ilmenau)
Sirko Schindler (DLR Jena)
Heiko Schuldt (Univ. Basel)
Uta Störl (Fernuni Hagen)
Dagmar Triebel (SNSB, München)
Matthias Weidlich (HU Berlin)
Claus Weiland (Senckenberg Gesellschaft für Naturforschung, Frankfurt)