DiTraRe: AI on a Spider's Web. Interweaving Disciplines for Digitalisation

Anna M. Jacyszyn[1,*], Harald Sack[1], DiTraRe-Study Group[1,2], Matthias Razum[1] and Felix Bach[1]

[1]FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
[2]Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

Abstract

The recently established Leibniz Science Campus "Digital Transformation of Research" (DiTraRe) investigates the effects of a broadly understood process of digitalisation of research on a multilevel scale. The project concentrates on four research clusters concerning different topics and gathering use cases from varying scientific areas. For a multi-scale investigation these research clusters are interwoven with four dimensions, each of which approaches the tasks from a different perspective and poses its own research questions. Within this "spider's web" we are not only developing practical solutions for each use case but also seeking to find generalisations valuable to the scientific community as well as society in general. Sophisticated AI technologies, like natural language processing, knowledge extraction, and ontology engineering, are investigated within the DiTraRe project by the dimension *Exploration and knowledge organisation.* This position paper aims to describe the DiTraRe Science Campus in general as well as concentrate on its aforementioned dimension concerning implementation of AI techniques.

Keywords

digitalisation, knowledge organisation, research data management, applied artificial intelligence

1. Introduction

The ongoing process of digitalisation holds great potential to simplifying and assisting not only our daily lives but also research activities. It enables us to transform our data into machine-readable formats, to which we can later apply numerous state-of-the-art (SOTA) techniques, such as e.g. machine learning (ML) models. The recently established Leibniz Science Campus "Digital Transformation of Research" (DiTraRe) [1][1] aims to investigate and analyse processes of digitalisation in research and their corresponding effects in a multi-levelled and interdisciplinary approach. The foundations for innovative techniques of knowledge creation will be laid by developing and applying many data-driven methods.

This paper concentrates on the vision of the DiTraRe dimension *Exploration and knowledge organisation* which is developing AI methods to support DiTraRe use cases. Within the project we will significantly enhance the ability to analyse and interpret fitness data for sports research. Our work with chemists concentrates on automatising processes in the laboratories, which will result in advancing the way we can teach AI the laws of chemistry. With biomedical engineers we will use SOTA AI techniques to develop a novel method of predicting the length of stay at an intensive care unit (ICU) in a non-invasive and much quicker way. A feature to support the process of creation of a uniform platform which we will construct with climate researchers will strongly increase the re-use and availability of earth science data.

4^{th} *International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment 11/12 November 2024 - Baltimore, MD, USA; co-located with The* 23^{rd} *International Semantic Web Conference, ISWC 2024*

*Corresponding author.

Anna.Jacyszyn@fiz-Karlsruhe.de (A. M. Jacyszyn); Harald.Sack@fiz-Karlsruhe.de (H. Sack); Matthias.Razum@fiz-Karlsruhe.de (M. Razum); Felix.Bach@fiz-Karlsruhe.de (F. Bach)

0000-0002-5649-536X (A. M. Jacyszyn); 0000-0001-7069-9804 (H. Sack); 0000-0002-5139-5511 (M. Razum); 0000-0002-5035-7978 (F. Bach)

[1]DiTraRe web page, https://www.ditrare.de/en

2. DiTraRe project

DiTraRe is based on the longstanding successful cooperation between FIZ Karlsruhe – Leibniz Institute for Information Infrastructure (FIZ KA)[2] and Karlsruhe Institute of Technology (KIT)[3]. The project is organised as a matrix: four research clusters begin with one use case (UC), which has specific research questions (RQs). Then, each of the UCs is being investigated within four dimensions.

2.1. Research clusters

Protected data spaces provides UC *Sensitive data in sports science* in cooperation with KIT Institute of Sports and Sports Science (KIT-IfSS) which is developing the MO|RE data platform. The challenge within this task is to investigate how sensitive health (i.e. BMI, blood pressure) and personal (i.e. geolocation, social status) data can be published in a way that personal and data protection rights are adhered to.

Smart data acquisition begins with UC *Chemotion Electronic Lab Notebook* (ELN)[4] in cooperation with KIT Institue of Biological and Chemical Systems (KIT-IBCS). The aim of this UC is to extend and improve automatisation processes in the chemistry labs. Its challenges include i.e. high complexity of processes ongoing in a lab as well as missing methods for automated data curation.

AI-based knowledge realms includes UC *AI in biomedical engineering* as a collaboration with KIT Institute for Biomedical Engineering (KIT-IBT), which develops computer models of the human heart. In order to overcome the problem of data privacy as well as biases in databases, simulated databases are being used. However, this leads to questioning the trustworthiness and explainability of AI.

Publication cultures provides UC *Publication of large datasets* in collaboration with KIT Institute of Meteorology and Climate Research (KIT-IMK) which generates and analyses very large datasets of atmospheric data. Due to the size of these data, publication, and reuse of data are limited and very inefficient. New methods need to be developed to enable exploration and evaluation.

2.2. Dimensions

Nowadays, SOTA methods need to be used to gain necessary access of the ever-growing abundance of information. This includes e.g. natural language processing (NLP) and ontology engineering, which enable obtaining well-structured knowledge. These semantic technologies, along with other AI techniques, are investigated within the DiTraRe by the dimension *Exploration and knowledge organisation*.

The dimension *Legal and ethical challenges* is dealing with data ethics, data protection, copyright, and data law. Legal and political context are studied in detail. *Tools and processes* will concentrate on providing the researchers with software enabling to keep a data continuum, e.g. by extending the possibilities of existing data repositories. The whole process will be examined from a technical, but also societal and ethical perspective by *Reflection and resonance*. The dimension will concentrate on transparency and comprehensibility of communicating scientific findings outside and within academia.

3. AI on a spider's web: *Exploration and knowledge organisation*

3.1. Sensitive data in sports science

KIT-IfSS has developed a platform to collect, publish, and share motor performance data (MO|RE) to deal with the issues of re-usability and reliability of data originating from numerous projects [2]. Even though the sports science is greatly profiting from this repository (e.g. study between green space availability and youth's physical activity [3]), many significant research questions still remain unanswered, i.e. evolution of motor performance over time. The plan is to combine the MO|RE repository with another database (namely KonsortSWD [4]), which provides broad information on

[2]FIZ Karlsruhe web page, https://fiz-karlsruhe.de/
[3]KIT web page, https://kit.edu/
[4]Chemotion ELN web page, https://chemotion.net/

social, behavioural, educational, and economic status. We will develop a knowledge graph (KG) of the extended MO|RE database to enable sports scientists an enhanced data analysis and interpretation. Preliminary studies have proven that the usage of KGs in health research is promising [5, 6].

We will concentrate on representations of datasets combining protected and non-protected data in a single KG and investigate possibilities of an efficient access management while retaining data privacy. Our plan is also to adapt the KG to the needs of sports science in general so that researchers working in other disciplines can easily make use of the motor activity data. Our efforts will clearly advance the potential KGs have within sports research by enabling knowledge discovery from patients' personal databases, thus uncovering new theories and classifying patients' health [6, 7].

3.2. Chemotion Electronic Lab Notebook

The complexity of processes ongoing in the chemistry lab leads to the need of novel methods of data acquisition and management to expand and enhance the data flow [8]. In the last years new methods were developed to make data findable, accessible, interoperable and reusable (FAIR) via the Chemotion ELN [9]. However, the process of digitalisation is far from being completed as many adaptations of the current workflows need to be done [10]. Our goal is to support chemists with further development of the ELN to accelerate their research by creating new automatisation methods supported with AI.

Data in Chemotion repository need to be tested for completeness and consistency in order to conform to the community standards [10]. This process is currently done partially manually by users and researchers. We will apply multiple AI techniques, i.e. NLP, to automatise the process of the data curation. We are going to specifically investigate the reaction description module in the Chemotion repository. Because of its complexity the text entered by the user is troublesome to standardise. Development of a full automatisation of data curation for the ELN will significantly improve and accelerate the functioning of the entire system, as it will exclude humans from the data quality check, thus giving them more time for other more crucial research tasks. And since using the ELN is also a best practice example, supporting it in the long run will facilitate teaching AI understanding the laws of chemistry [8], i.e. by development of self-driving labs [11].

3.3. AI in biomedical engineering

Since its first clinical applications, simulations and computational modelling have become an important and regularly used method in cardiac electrophysiology [12]. ML methods are also being utilised in processes concerning high-risk patients, i.e. those admitted to an intensive care unit (ICU) [13]. Recent studies which concentrated on predicting the length of ICU stay and mortality used a significant number of variables, e.g. 17 [14] or 20 [15]. Acquisition of these measurements upon patient's admission to an ICU is time and effort consuming. The researchers at KIT-IBT are now in the process of developing a prediction tool which will use an electrocardiogram (ECG) together with only few other variables to predict mortality and length of ICU stay faster.

A recent study used only the parameters of an ECG of non-cardiac patients [16]. We will collaborate with KIT-IBT on including data of patients with heart diseases in their novel model. Publicly available Medical Information Mart for Intensive Care (MIMIC) IV database [17] together with clinical data from University of Freiburg Heart Center will be used. In our innovative approach we will concentrate on using a raw ECG as an input and employ a multi-modal large language model (multi-modal LLM) to feed it with different variables in addition to the ECG. We will investigate how far incrementally adding different types of external information (i.e. text, image) influences the outcome. This model will enable a quick and non-invasive assessment of a length of stay at the ICU which is important because of significant costs that ICU long stays yield for general health care systems.

3.4. Publication of large datasets

KIT-IMK researchers are using multiple internal and external datasets to analyse the composition of the atmosphere. The available datasets consist of highly non-uniform (e.g. dimensions of the variables

are different) and in some cases non-standardised entries where the metadata have slightly different names for the same data (see examples in the Earth Data portal[5]). This metadata chaos as well as large sizes of accessible datasets (see i.e. data from IASI satellite [18]) makes it challenging to access and use the data from within a repository. Within the DiTraRe we are aiming at creating a platform to enable easy and convenient publication of large heterogeneous datasets.

Environmental scientists have been working on a uniform platform to collect datasets from different disciplines, such as V-FOR-WaTer [19]. However, it is still under construction. Our plan is to focus specifically on KIT-IMK data and test whether AI methods such as KGs or LLMs can support functionality of such repositories. We will explore the utility of AI in structuring datasets, preprocessing data, and standardising metadata. An intriguing case is non-stationarity of the climate system in general and its impact on the possibility of an effective usage of AI methods, i.e. an ontology. Our research will provide significant contributions to creating a uniform data management platform.

3.5. Overarching activities and synergies

Along with one of the main goals of DiTraRe, we will study synergies between research clusters. Both KIT-IfSS and KIT-IBT would profit from using clinical data, and plausibly health data from wearables. This way they can extend their studies and find connections between numerous health indicators. Another connection is that KIT-IMK as well as KIT-IfSS and KIT-IBCS are making use of the RADAR platform, which is a repository developed at FIZ KA for the archival and publication of research data[6]. We are investigating possibilities of its further development with the support of AI methods, i.e. by adding a SPARQL endpoint explorer SHMARQL which enables user friendly exploration of the shape of data [20].

Additionally, our plan is to formulate and answer RQs in the area of computer science. An important topic will be the applications of LLMs and its consequences in different UCs within varying context. Thus, the DiTraRe will not only bring forward our understanding of what role the specific aspect of AI plays within the scope of the digitalisation of research, but also enables new insights about knowledge representation in LLMs [21] as well as about their efficiency in memorising and reasoning among structured knowledge [22].

4. Conclusions

In the DiTraRe we investigate the effects and influences of a broadly understood digitalisation of research both within academia and society. This paper describes in detail tasks of the DiTraRe dimension *Exploration and knowledge organisation*, within which we explore effects of applying AI methods to specific UCs originating from various disciplines. The developed solutions in the context of scientific knowledge representation and organisation will cover areas as broad as accessing databases containing private and sensitive data, improving treatment of high-risk patients, automatising data quality control for chemistry and facilitating easy re-use of large datasets for climate research. The practices which will be the outcomes of DiTraRe, will significantly improve current SOTA in the investigated areas of research as well as advance our understanding of applying AI in the process of knowledge organisation. The project is currently in its development phase and we would significantly profit from networking with other researchers looking into similar topics in scientific knowledge representation.

Acknowledgments

The Leibniz Science Campus “Digital Transformation of Research” (DiTraRe) is funded by the Leibniz Association. We would like to thank our use case partners for kindly offering their time to revise the draft and sharing their insightful suggestions on how to improve this publication.

[5]Earth Data portal, https://earth-data.de/

[6]RADAR web page, https://radar.products.fiz-karlsruhe.de/en

References

[1] M. Razum, F. Bach, S. Brünger-Weilandt, C. Scherz, F. Böhm, H. Sack, M. Volkamer, Proposal for a Leibniz ScienceCampus – Digital Transformation of Research (DiTraRe), 2023. doi:10.5281/zenodo.11109406, project proposal.

[2] K. Klemm, K. Bös, H. Kron, T. Eberhardt, A. Woll, C. Niessner, Development and introduction of a disciplinary data repository for sport scientists based on the example mo|re data: eresearch infrastructure for motor research data : Bausteine forschungsdatenmanagementempfehlungen und erfahrungsberichte für die praxis vonforschungsdatenmanagerinnen und -managern, Bausteine Forschungsdatenmanagement 1 (2024) 1–14. doi:10.17192/bfdm.2024.1.8615.

[3] C. Nigg, J. Fiedler, A. Burchartz, M. Reichert, C. Niessner, A. Woll, J. Schipperijn, Associations between green space availability and youth's physical activity in urban and rural areas across germany, Landscape and Urban Planning 247 (2024). doi:10.1016/j.landurbplan.2024.105068.

[4] R. G. D. Forum], Big data in social, behavioural, and economic sciences: Data access and research data management. ratswd output, German Data Forum (RatSWD) 4 (2020). doi:10.17620/02671.52.

[5] P. Ernst, C. Meng, A. Siu, G. Weikum, Knowlife: A knowledge graph for health and life sciences, 2014 IEEE 30th International Conference on Data Engineering (2014) 1254–1257. doi:10.1109/ICDE.2014.6816754.

[6] X. Tao, T. Pham, J. Zhang, J. Yong, W. Goh, W. Zhang, Y. Cai, Mining health knowledge graph for health risk prediction, World Wide Web 23 (2020) 2341–2362. URL: https://doi.org/10.1007/s11280-020-00810-1. doi:10.1007/s11280-020-00810-1.

[7] A. Rossanez, J. dos Reis, R. Torres, H. de Ribaupierre, Kgen: a knowledge graph generator from biomedical scientific literature, BMC Medical Informatics and Decision Making 20 (2020). doi:10.1186/s12911-020-01341-5.

[8] F. Fink, H. Hüppe, N. Jung, A. Hoffmann, S. Herres-Pawlis, Sharing is caring: Guidelines for sharing in the electronic laboratory notebook (eln) chemotion as applied by a synthesis-oriented working group, Chemistry–Methods 2 (2022) e202200026. doi:https://doi.org/10.1002/cmtd.202200026.

[9] S. Herres-Pawlis, F. Bach, I. Bruno, S. Chalk, N. Jung, et al., Minimum information standards in chemistry: A call for better research data management practices, Angewandte Chemie International Edition 61 (2022) e202203038. doi:https://doi.org/10.1002/anie.202203038.

[10] F. Tristram, N. Jung, P. Hodapp, R. Schröder, C. Wöll, S. Bräse, The impact of digitalized data management on materials systems workflows, Advanced Functional Materials 34 (2024) 2303615. doi:https://doi.org/10.1002/adfm.202303615.

[11] P. Maffettone, P. Friederich, S. Baird, B. Blaiszik, K. Brown, et al., What is missing in autonomous discovery: open challenges for the community, Digital Discovery 2 (2023) 1644–1659. doi:10.1039/D3DD00143A.

[12] M. Peirlinck, F. Costabal, J. Yao, J. M. Guccione, S. Tripathy, et al., Precision medicine in human heart modeling, Biomechanics and Modeling in Mechanobiology 20 (2021) 803–831. doi:10.1007/s10237-021-01421-z.

[13] S. Iwase, T. Nakada, T. Shimada, T. Oami, T. Shimazui, et al., Prediction algorithm for icu mortality and length of stay using machine learning, Scientific Reports 12 (2022). doi:10.1038/s41598-022-17091-5.

[14] J. Wu, Y. Lin, P. Li, Y. Hu, L. Zhang, G. Kong, Predicting prolonged length of icu stay through machine learning., Diagnostics 11 (2021). doi:10.3390/diagnostics11122242.

[15] S. Jana, T. Dasgupta, L. Dey, Predicting medical events and icu requirements using a multimodal multiobjective transformer network, Experimental Biology and Medicine 247 (2022) 1988–2002. doi:10.1177/15353702221126559.

[16] K. Erdem, I. Duman, R. Ergün, D. Ergün, The correlation between electrocardiographic parameters and mortality in non-cardiac icu patients, European Review for Medical and Pharmacological Sciences 27 (2023) 6662–6670. doi:10.26355/eurrev_202307_33136.

[17] A. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, et al., Mimic-iv, a freely accessible electronic health record dataset, Scientific Data 10 (2023). doi:10.1038/s41597-022-01899-x.
[18] M. Schneider, B. Ertl, C. Diekmann, F. Khosrawi, A. Weber, et al., Design and description of the musica iasi full retrieval product, Earth System Science Data 14 (2022) 709–742. doi:10.5194/essd-14-709-2022.
[19] M. Strobl, E. Azmi, B. Balazs, S. Bouguezzi, A. Dolich, et al., Streamlining data pre-processing and analysis through the v-for-water web portal, in: European Geosciences Union General Assembly, 2024. doi:10.5194/egusphere-egu24-10364.
[20] E. Posthumus, Sparql-shmarql, 2024. URL: https://github.com/epoz/shmarql.
[21] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bahktin, Y. Wu, A. Miller, Language models as knowledge bases?, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2463–2473. URL: https://aclanthology.org/D19-1250. doi:10.18653/v1/D19-1250.
[22] Q. He, Y. Wang, W. Wang, Can language models act as knowledge bases at scale?, 2024. doi:10.48550/arXiv.2402.14273, arXiv preprint.

A. Graphical representation of DiTraRe structure

Fig. 1 presents a schematic graph of the structure of DiTraRe. The matrix represented as a spider's web connects four research clusters (on the right, from top to bottom: protected data spaces, smart data acquisition, AI-based knowledge realms, publication cultures) with four dimensions (on the left, from top to bottom: exploration and knowledge organisation, legal and ethical challenges, tools and processes, reflection and resonance). Each of the research clusters starts with one specific use case. Dimensions are interwoven with each of them.

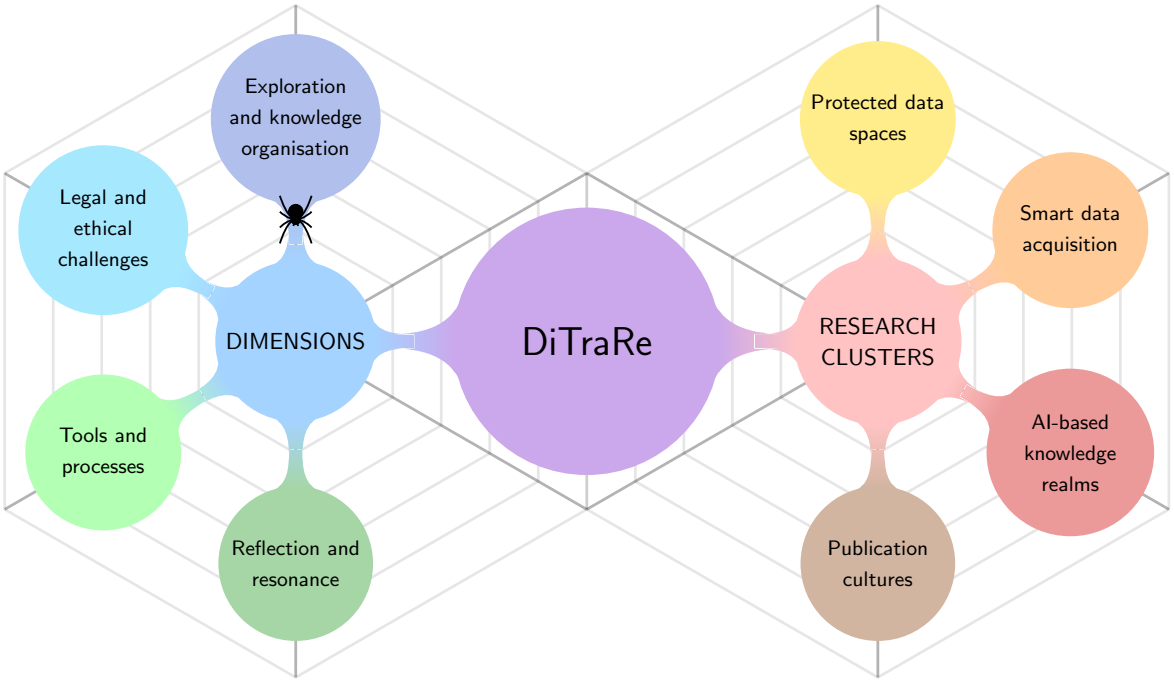

Figure 1: A schematic representation of the structure of the DiTraRe. The central node represents the project which brings together four research clusters (on the right) and four dimensions (on the left).