Frühjahrstreffen der FG Datenbanken

Report of Contributions

https://indico.rz.uni-jena.de/e/100

Contribution ID: **1** Type: **Poster**

Bridging the gap between data lakes and RDBMSs - Efficient query processing with Parquet

In the age of massive data, time-intensive loading phases make databases less viable for data exploration tasks.
Still, the highly optimized query engines of database systems are greatly beneficial for the performance of data analysis tasks.
With our research, we want to bridge this gap and provide paramount analytical performance without the need of static data loading.
Our approach enables the integration of Parquet files — one of the most used columnar file formats in data lakes — into the data processing pipeline of a database system in a convenient way.
We allow end-users to benefit from the database system performance without a costly and time-consuming loading phase.

Type of Poster

A solution

Primary author: REY, Alice

Presenter: REY, Alice

Session Classification: Poster

Track Classification: Poster

Contribution ID: **3** Type: **Poster**

VirJenDB: the comprehensive virus database based in Jena

Recent advances in sequencing technologies have resulted in a deluge of virus sequences, lately consisting mostly of SARS-Cov-2, into the global repositories such as the INSDC[1]:ENA[2] and GISAID[3]. However, many other viruses are under sequenced, and metadata is often lacking which hampers the re-use of the data[4]. As a service and upcoming Use Case of the NFDI4Microbiota consortium[5], we have developed at FSU Jena[6] the VirJenDB database, a web-based platform for the re-use and analysis of publicly available metadata and sequences from all viruses following FAIR[7] and Open Science[8] principles.

Our goals are 1) to build a lasting infrastructure with useful features based on regular feedback from virus researchers, 2) to contribute to metadata standards through the curation of the VirJenDB dataset and 3) to support virus researchers in gaining and disseminating research data management skills.

We ingested publicly available virus sequences and metadata from BV-BRC[9]; NCBI Virus[10]; ICTV[11]; and ViralZone[12]. The source data was integrated into a virus data model organized in a MySQL database with source code and documentation available on GitHub[13]. VirJenDB was built as an OpenStack project on de.NBI[14] and uses the Aruna storage system[15]. The frontend was powered by the Java REACT[16] and node.js[17] frameworks. Key features of the current VirJenDB include semantic search, taxonomy browser and download of virus sequences, statistical figures and integrated metadata. The beta version web interface can be accessed at https://virjendb.uni-jena.de.

Upcoming developments and areas of improvement include automation of data ingestion and the addition of gene annotations, sequence alignments, metagenome (mg) sequences and mg-derived genomes. Further, we plan to integrate the following tools: sequence search and automatic and community curation, as well as provide subsets of virus sequences for external use in bioinformatic pipelines. We envision a secure workbench for users to upload and analyze their own virus data with additional tools such as phylogenetic and variation analyses. VirJenDB will help to advance global virus research by integrating virus research data, and connecting researchers with relevant tools and training.

References

1. https://www.insdc.org
2. https://www.ebi.ac.uk/ena/browser/home
3. https://www.gisaid.org
4. https://www.doi.org/10.3390/v15091834
5. https://www.nfdi4microbiota.de
6. https://www.uni-jena.de
7. https://www.doi.org/10.1038/sdata.2016.18
8. https://www.unesco.org/en/open-science?hub=686
9. https://www.bv-brc.org
10. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#
11. https://ictv.global
12. https://viralzone.expasy.org
13. https://www.github.com
14. https://www.denbi.de
15. https://aruna-storage.org
16. https://react.dev
17. https://nodejs.org/en

Keywords: NFDI4Microbiota, RNA, DNA, Virus, Database, Genome, Phage, Bacteriophage, Next-

Generation Sequencing, Genome, Metadata, FAIR, RDM

Type of Poster

A solution

Primary author: CASSMAN, Noriko (Friedrich Schiller University)

Co-authors: ZIRAKSAZ, Hamed (FSU Jena); MARZ, Manuela (FMI, FSU Jena); SAGHAEI, Shahram (FSU Jena)

Presenter: CASSMAN, Noriko (Friedrich Schiller University)

Session Classification: Poster

Track Classification: Poster

Contribution ID: **4** Type: **Poster**

Datenbankherstellerrecht und Datenbankforschung

Mit diesem Poster stellen wir das Datenbankherstellerrecht vor. Hierbei handelt es sich nicht, wie man aus dem Blickwinkel eines juristischen Laien und Mitglied der Datenbankforschungsgemeinde meinen könnte, um die Rechte bei der Entwicklung einer Datenbankmanagementsoftware, sondern um die Rechte des Herstellers einer Datenbankinstanz. Auch Forschende oder Forschungsinstitutionen werden beim Forschungsdatenmanagement zu Datenbankherstellern, insbesondere wenn sie Forschungsartefakte verfügbar machen. Somit gewinnt die zugrundeliegende EU-Richtlinie aus den 1990er-Jahren an neuer Brisanz.

Unser Poster gibt einen systematischen Überblick über den rechtlichen Schutz der einzelnen Komponenten einer Datenbankanwendung. Insbesondere stellen wir das Datenbankherstellerrecht und praktische Anwendungsfälle vor, sowie anknüpfende Forschungsfragen.

Dieses Poster fasst den gleichnamigen Artikel zusammen, welcher im Datenbank-Spektrum 02/2023 erschienen ist.

Type of Poster

A challenge

Primary authors: Prof. BEURSKENS, Michael (Universität Passau); SCHERZINGER, Stefanie (Universität Passau)

Presenters: Prof. BEURSKENS, Michael (Universität Passau); SCHERZINGER, Stefanie (Universität Passau)

Session Classification: Poster

Track Classification: Poster

Contribution ID: **5** Type: **Talk**

From Research Data Management to Data Platforms: A Hugging Face Approach

Monday, March 11, 2024 5:45 PM (30 minutes)

Does research data management as we know it in the context of database research or data science need platforms like Hugging Face? Or are platforms and services such as Kaggle or GESIS sufficient? In this talk, after giving a brief overview of the core features of Hugging Face, we claim that the data research community would benefit a lot from a platform similar to Hugging Face, in particular when considering the support of the FAIR principles. We will also stress that proper infrastructures for research data management should go beyond just managing datasets and making them accessible to the research community. In particular, in view of large-scale data management, processing and analysis, it would be extremely helpful to provide researchers a platform that offers various tools and AIPs to easily interact with and explore diverse forms of data.

Type of Poster

Presenter: GERTZ, Michael (U Heidelberg)

Session Classification: Talks

Track Classification: Vortrag

Contribution ID: **6** Type: **Talk**

Schema Evolution in Research Data

Monday, March 11, 2024 4:00 PM (30 minutes)

Changes occur frequently, especially in data-driven long-term studies. Changing databases lead to the accumulation of many schemes and instances over time. However, any scientific application must be able to reconstruct the historical data to ensure the reproducibility or at least the explainability of the research results. A method is needed that allows each database version to be easily reconstructed at both the schema and data level, and data to be migrated between the different versions. Storing all versions over time is not a feasible solution, as it is often too expensive and storage-consuming. In contrast, a method that allows backward processing to earlier versions of the database guarantees the recoverability of the stored information without keeping different versions. This is the subject of our current research, where we use evolution with provenance and additional information to facilitate the reproducibility of scientific results over long periods of time. In this way, information loss can be avoided or at least reduced.

Type of Poster

Presenter: AUGE, Tanja (U Regensburg)

Session Classification: Talks

Track Classification: Vortrag

Contribution ID: **7** Type: **Talk**

Medax - a knowledge graph for biomedicine

Monday, March 11, 2024 3:30 PM (30 minutes)

Within the MeDaX project we study bioMedical Data eXploration using graph technologies. We design and implement efficient concepts and tools for integration, enrichment, scoring, retrieval, and analysis of biomedical data. Interested in data similarity and quality measures, we initiated an international community project for biomedical provenance standardisation and cooperate within the Medical Informatics Initiative (MII) to FAIRify the MII core data set. Those and other projects build the basis for development of a pipeline for knowledge graph (KG) creation from diverse data sources, for automated semantic enrichment, and for data scoring and analysis. For the MeDaX-KG prototype, we build on existing tools such as CyFHIR (generic conversion of FHIR to Neo4j) and BioCypher (harmonising framework for KG creation) and optimise graph complexity and structure by our own methods and code.

Type of Poster

Presenter: WODKE, Judith (U Greifswald)

Session Classification: Talks

Track Classification: Vortrag

Contribution ID: **8** Type: **Talk**

Terminologies in database systems

Monday, March 11, 2024 3:00 PM (30 minutes)

The use of commonly agreed terminologies is an elementary component of database systems. They have an impact on data consistency, querying and retrieval or interoperability. Creating, searching for and agreeing on a terminology to be used is a non-trivial problem, as it requires specialised knowledge and coordination processes. This presentation introduces the terminology service that deals with some of these issues.

Type of Poster

Presenter: ENGEL, Felix (TIB)

Session Classification: Talks

Track Classification: Vortrag

Contribution ID: **9** Type: **Talk**

On the Path to a Quality Indicator for Software and Data Publications for the Helmholtz

Monday, March 11, 2024 1:15 PM (45 minutes)

Research data and software publications have become a regular output of scientific work. Yet unlike more traditional text publications, widely established processes to assess and evaluate their quality are still missing. This fact prevents researches from getting the proper credit they deserve as common performance indicators often just omit this part of scientific contributions.
As part of the Helmholtz Association, the Task Group Helmholtz Quality Indicators for Data and Software Publications has been set up to develop a quality indicator to be used within the Association. The goal is to define a set of quality dimensions and attributes suitable for all branches represented in Helmholtz and raise the awareness and appreciation of research data and software publications as equally important scientific outputs. We base our work on already well-established frameworks like the FAIR principles and the COBIT Maturity Model and aim to define a graded model accounting for multifaceted nature of contemporary research.
In our talk, we will present the vision of the Task Group as well as the current state of discussions. As the definition these criteria is a continuous and dynamic process, we welcome feedback by the audience want to encourage a further dialogue within the community.

Type of Poster

Presenter: MEISTRING, Marcel (Helmholtz Open Science Office)

Session Classification: Talks

Track Classification: Vortrag

Contribution ID: **10** Type: **Talk**

Tabular Data Synthesis for Data Management

Tuesday, March 12, 2024 11:15 AM (45 minutes)

The problem of generating synthetic data is almost as old as modern research itself. However, with the advent of generative AI, new possibilities for synthesizing tabular data have emerged that go far beyond the capabilities of traditional statistical or rule-based approaches. Most of this new research comes from the ML community, where ML models need to be fed with useful training data. Since many data management use cases also require synthetic data, it makes sense to adapt these research results. Nevertheless, those use cases, such as query optimization, have different requirements than ML use cases. Requirements that are currently not met by such modern synthesizers. In this talk, we will give an overview of the current state of the art in the field of tabular data synthesis and discuss open challenges in the context of generating synthetic tabular data for data management.

Type of Poster

Presenter: PANSE, Fabian (HPI)

Session Classification: Talks

Track Classification: Vortrag

Contribution ID: **11** Type: **Talk**

Research Data Management in TIRA for Reproducible Shared Tasks

Tuesday, March 12, 2024 12:30 PM (30 minutes)

TIRA is a platform to organize shared tasks with software submissions, mostly in information retrieval and natural language processing. Due to the software submissions, TIRA allows blinded experimentation on (confidential) datasets to which participants have no access. After a shared task, the artifacts of the shared tasks, i.e., research data in the form of submitted software, inputs, and outputs to systems, or ground-truth labels, can be made publicly accessible if desired. Archiving of software and data artifacts in TIRA aims to improve experimental results' reproducibility and simplify comparisons against strong baselines in future research.

Type of Poster

Presenter: FRÖBE, Maik (U Jena)

Session Classification: Talks

Track Classification: Vortrag

Contribution ID: **12** Type: **Poster**

STRENDA DB –a Web-based Assessment and Storage Tool for Enzymology Data

The STRENDA Commission (STandards for Reporting ENzymology Data, www.beilstein-strenda.org) made up of experts from the enzyme chemistry community and supported by the Beilstein-Institut, has developed the STRENDA Guidelines in tight consultation with the community. The aim is to improve the quality of enzyme function data in the literature. Today, more than 60 biochemical journals already recommend authors to refer to these guidelines when reporting enzyme kinetics data.

To enable scientists to easily prepare data for manuscripts, the STRENDA Commission has developed a web-based portal for the direct electronic submission of data by the authors prior to publication. This portal called STRENDA DB provides an assessment tool with which authors, journals'editors and reviewers can check whether the reporting of experimental data is compliant with the STRENDA guidelines and thus matches the instructions for authors from the journals. The data entered are stored in STRENDA DB and will be made publically accessible after they have been published in a journal. More than 20 biochemical journals recommend their authors to deposit their experimental data in STRENDA DB.

However, as Findability and Accessibility of datasets stored in STRENDA DB are relatively easy to implement, making the data interoperable is still a challenge. There might be multiple ways to address this issue (and I'd be keen to learn more about this) and certainly, a reasonable way is the creation of a standardized data exchange format that allows the seamless transfer of data from the lab bench via tools to databases and back. Here, I will shed very briefly some light on EnzymeML which is a community-developed data exchange format in its first version for enzymology and biocatalysis data that also provides an API for Python and Java libraries to be integrated into both applications and databases.

Type of Poster

A solution

Primary author: KETTNER, Carsten (Beilstein-Institut zur Förderung der Chemischen Wissenschaften)

Session Classification: Poster

Track Classification: Poster

Contribution ID: **13** Type: **Talk**

From theory to practice - Advancing Research Assessment for Incentives at Charité and BIH through infrastructure

Monday, March 11, 2024 2:00 PM (30 minutes)

There is a gap between current responsible research and innovation (RRI) as well as open sciences (OS) practices and assessment practices. While
research practices and their ways of publication and dissemination have diversified, assessment practices have remained narrow –focusing on criteria of publication quantity and reputation. In my talk, I will discuss two projects. The first project is the
MERIT portal –an application and assessment software for appointments of professors. The MERIT portal introduces structured CVs including RRI as well as OS criteria and strategies to reduce the risk of bias during assessments. The focus is to strengthen quality and content-oriented assessments with the support of science-based metrics. The
second project is the Open Data LOM project. In 2019, the Charité introduced an open sciences indicator in the institutional performance-oriented funding system.

Type of Poster

Presenter: KIP, Miriam (BIH Charité)

Session Classification: Talks

Track Classification: Vortrag

Contribution ID: **14** Type: **Talk**

Democratising data analysis with Galaxy

Monday, March 11, 2024 5:00 PM (45 minutes)

Galaxy is an open-source platform that allows researchers to analyze and share scientific data using interoperable APIs and various user-friendly web-based interfaces. The Galaxy project was launched in 2005 and has since become a powerful tool for researchers across a wide range of research fields, including *omics, biodiversity, machine learning, cheminformatics, NLP, material science, climate research.

One of the key features of the Galaxy platform is its emphasis on transparency, reproducibility, and reusability. Galaxy is a multi-user environment which facilitates sharing of e.g. tools, workflows, notebooks, visualizations, and data with others. This makes it particularly easy to reproduce results in order to verify their correctness and enable other researchers to build upon them in future studies. All provenance information of a dataset, including version of used tools, parameters, execution environment are captured and can be reused or exported using standards like BCO or RO-Crate to public archives.

Type of Poster

Presenter: GRÜNING, Björn

Session Classification: Talks

Track Classification: Vortrag

Contribution ID: **15** Type: **Talk**

Exploring Computational Reproducibility in Jupyter Notebooks: Insights and Challenges

Tuesday, March 12, 2024 12:00 PM (20 minutes)

Reproducible research emphasizes the importance of documenting and publishing scientific results in a manner that enables others to verify and extend them. In this talk, we explore computational reproducibility within the context of Jupyter notebooks, presenting insights and challenges from our study. We will present the key steps of the pipeline we used for assessing the reproducibility of Jupyter Notebooks. In our study, we analyzed the notebooks extracted from GitHub repositories associated with publications indexed in the biomedical literature repository PubMed Central. Our process involved identifying the notebooks by mining the full text of publications, locating them on GitHub, and attempting to rerun them in an environment closely resembling the original. We documented reproduction success and exceptions and explored relationships between notebook reproducibility and variables related to the notebooks or publications, including results related to programming languages, notebook structure, naming conventions, modules, dependencies, etc. Furthermore, we will discuss the common issues and practices, identify emerging trends, and explore potential enhancements to Jupyter-centric workflows. Through this comprehensive examination, we aim to provide actionable insights and practical strategies for researchers striving to enhance the reproducibility of their work within the Jupyter notebook ecosystem and contribute to the ongoing dialogue surrounding reproducibility and computational methodologies in scientific research.

Type of Poster

Presenter: SAMUEL, Sheeba (Friedrich Schiller University)

Session Classification: Talks

Track Classification: Vortrag

Contribution ID: **16** Type: **Talk**

Problems and Issues in Biodiversity Data Infrastructures

Tuesday, March 12, 2024 9:00 AM (30 minutes)

The current biodiversity crisis has triggered an extreme need for a better understanding of the network of life on Earth. Efficient data management is crucial in biodiversity and is the backbone for a digital twin of past, present, and future life. The Research Data Commons (RDC) is the central cloud-based information system architecture of NFDI4Biodiversity, the consortia of the NFDI (Nationale Forschungsdateninfrastruktur) offering reliable biodiversity data and services for improving the conservation of global biodiversity.

This talk introduces the essential components of the RDC and provides an overview of research problems and issues we faced during its first development phase. As biodiversity is a data-intensive discipline with many heterogeneous small and large data sources following various metadata formats and collected from different research communities, the RDC faces massive data integration problems. Moreover, the derived data products also must obey specific criteria like the FAIR data principles. In summary, we see plenty of opportunities for the database community to address challenging research questions in an area highly relevant to society.

Type of Poster

Presenter: SEEGER, Bernhard (U Marburg)

Session Classification: Talks

Contribution ID: **17** Type: **Poster**

Revisiting the process of Knowledge Graph generation with the integration of LLMs

In recent years, the advent of Large Language Models (LLMs) has transformed both natural language processing (NLP) and knowledge representation. With vast pre-trained parameters and advanced neural architectures, these models show remarkable results in generating human-like text. In knowledge representation, ontologies serve as fundamental frameworks for organizing and representing knowledge across domains. These structured frameworks serve as the basis for constructing comprehensive knowledge graphs (KGs). KGs, in turn, provide a robust mechanism for linking diverse information and enabling sophisticated data analytics and reasoning. Creating ontologies and KGs require considerable time and effort, typically involving domain expertise and many design decisions. In this poster, we explore the use of LLMs for creating KGs. We present our semi-automatic pipeline for KG construction for the deep learning techniques used in scholarly publications in the biodiversity domain, with the use of open-source LLMs. We present our insights and challenges in the automatic construction of knowledge engineering in terms of prompt sensitivity, and repetitive and inaccurate answers.

Type of Poster

A challenge

Primary authors: KOMMINENI, Vamsi Krishna (Friedrich Schiller University Jena); SAMUEL, Sheeba (Friedrich Schiller University Jena); KÖNIG-RIES, Birgitta (Heinz Nixdorf Chair for Distributed Information Systems)

Session Classification: Poster

Track Classification: Poster

Contribution ID: **18** Type: **Poster**

Reproducibility of Deep Learning pipeline method information using a Multi-modality approach

Scientific publications have enormous amounts of information and serve as the main pillar for advancing knowledge across various disciplines. Recently, many sectors and disciplines have been employing Deep Learning (DL) models due to their popularity. However, manually extracting DL method information from publications is becoming tedious with the ever-growing published literature. On the other hand, validating and verifying this information is a pivotal step for checking the reproducibility of the DL pipeline in scientific publications. In this work, we leverage the multimodal information (text, figures, tables, graphs, etc.) to automatically retrieve the method information of DL pipelines in scientific publications using the suite of open-source models, including Large Language Models (LLMs) and computer vision models. We will present the initial results from the text modality of DL method information from biodiversity scientific publications drawn using open-source LLMs.

Type of Poster

A challenge

Primary author: KOMMINENI, Vamsi Krishna (Friedrich Schiller University Jena)

Co-authors: AHMED, Waqas (Friedrich Schiller University Jena); KÖNIG-RIES, Birgitta (Heinz Nixdorf Chair for Distributed Information Systems); SAMUEL, Sheeba (Friedrich Schiller University Jena)

Session Classification: Poster

Track Classification: Poster

Contribution ID: **19** Type: **Poster**

The FAIR data principles from a repository perspective - BEXIS2 status and outlook

With the acceptance of the FAIR Data principles in the research community, the requirements and standards of data publications have changed significantly. While the FAIR principles are explicitly targeted at metadata and digital resources such as APIs, workflows, ontologies, and models, these digital objects can not be made FAIR without supporting infrastructure services that are themselves FAIR.

We are developing BEXIS2, an open-source, community-driven, web-based research data management system. In 2021, we conducted the self-assessment using the FAIR indicators, definitions, and criteria provided in the FAIR Data Maturity Model. The self-assessment results indicated that BEXIS2 remarkably conforms and supports FAIR indicators.

In our poster, we show the results of the FAIR self-assessment, our current developments with regard to improvements of the FAIR data principles, and our very next ideas for improving these aspects.

Type of Poster

A challenge

Primary authors: OSTROWSKI, Andreas (Friedrich-Schiller-Universität Jena); FÜRSTENAU, Cornelia (Friedrich-Schiller-Universität Jena); SCHÖNE, David (Friedrich-Schiller-Universität Jena); PETZOLD, Eleonora (Friedrich-Schiller-Universität Jena); ZANDER, Franziska (Friedrich-Schiller-Universität Jena); HOHMUTH, Martin (Friedrich-Schiller-Universität Jena); THIEL, Sven (FSU Jena); KÖNIG-RIES, Birgitta (Heinz Nixdorf Chair for Distributed Information Systems)

Session Classification: Poster

Track Classification: Poster

Contribution ID: **20** Type: **Poster**

Enabling Semantic Tools for Interdisciplinary Research

Research has become increasingly reliant on extensive data. The integration, sharing and reuse of research data poses a significant challenge, particularly in the context of interdisciplinary collaborative projects. An essential objective for a research infrastructure dedicated to data management is to facilitate efficient data discovery and integration of diverse data sources. This pressing need for FAIR data requires, besides persistent identifiers and data citation rules, common standards and shared vocabularies, thesauri and ontologies. These knowledge artifacts, referred to as terminologies, often exist in disconnected and distributed forms. The work presented in this poster describes our terminology repository and service, enabling a unified access, development, and maintenance of terminologies within biodiversity and environmental sciences. We show how BiodivPortal supports semantically enhanced components and applications in the context of our Research Data Commons.

Type of Poster

A solution

Primary author: KARAM, Naouel (Institut für Angewandte Informatik (InfAI))

Session Classification: Poster

Track Classification: Poster

Contribution ID: **21** Type: **Poster**

FAIR Assessment Tools: An evaluation of assessment tools of data sets according to the FAIR principles

Since the publication of the FAIR principles in 2016, they have become increasingly important and various tools have been developed to help assess published data with regard to compliance with the FAIR principles. There is a wide range of fair assessment tools currently available, from simple printable PDF checklists to fully automated tools that only require a DOI or URL to perform the assessment. Researchers hoping for feedback on how to optimize their own dataset with regard to the FAIR principles have different requirements than data stewards who need a quick overview of the quality of the datasets published in the repository. In order to get an orientation as to which tools are suitable for which user group and which question, we evaluated the FAIR assessment tools available in the period from July to August 2022. In our evaluation, we considered the following aspects, among others: the duration of processing, the target group of the tool, whether prior knowledge (in the field of IT and RDM) is necessary for using the tool and for understanding the results.

The poster summarizes the evaluation of the FAIR assessment tools by assigning them to four categories: Fully Configurable Tools, Automatic Tools, Improved Survey Tools and Regular List Tools. The categorization gives users an overview and thus supports them in selecting the right tool for their needs.

Type of Poster

A solution

Primary authors: ASSMANN, Cora (Friedrich-Schiller-Universität Jena); REX, Jessica; LANG, Kevin; NEUTE, Nadine; GERLACH, Roman

Session Classification: Poster

Track Classification: Poster

Contribution ID: **22** Type: **Poster**

Large-Scale Analysis of Heterogeneous Earth Observation Data

In recent years, the amount of data made available in the earth-observation domain has increased exponentially. In 2022, for example, data released from the observations of eight Sentinel satellites amounted to 6.64 petabytes [2]. Now, researchers all over the world are using these vast amounts of resources to further improve our understanding of the world. In their journey, the researchers often use code notebooks like Jupyter while conducting their analyses [1]. However, analyses have shown [3] that the code quality in these notebooks is generally mediocre. As analyses grow larger using more, higher-resolution datasets, it is now necessary to find new methodologies for handling the data.

In the past, it was feasible to simply download all the analysis-ready data (ARD), store it locally and run the required analysis. Now, as the amount of data to be processed has increased significantly, this approach of always downloading data to a local machine is not practical anymore. The main approach to address this issue is by pushing the computation to the cloud. Usually, these cloud platforms provide the user with a code notebook interface, an EO dataset catalog, and optionally a web interface for exploring datasets. However, as each pipeline has to contain some code that loads data, improving the pipeline by splitting up the execution pipeline in order to reduce latency and overall resource utilization can have a huge impact on users being able to perform their research.

To solve these issues, we want to create an algebra for Earth Observation queries. Based on this algebra, we will build an optimizer that distributes the workload between the data provider, platform provider, and user in order to minimize the query latency while maximizing efficiency. This is important for several reasons: First, it increases the productivity of an Earth Observation scientist. Second, it decreases the amount of resources necessary to complete the workload. And finally, it caters to the FAIR (findable, accessible, interoperable, reusable) data principles as it allows other scientists to validate datasets and perform analyses more efficiently.

References

[1] Caprarelli, G. et al. 2023. Notebooks Now! The Future of Reproducible Research. Earth and Space Science. 10, 12 (2023), e2023EA003458. DOI:https://doi.org/10.1029/2023EA003458.
[2] Castriotta, A.G. 2023. Copernicus Sentinel Data Access Annual Report Y2022. Technical Report #1. European Commission.
[3] Wang, J. et al. 2020. Better code, better sharing: on the need of analyzing jupyter notebooks. Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results (Seoul South Korea, Jun. 2020), 53–56.

Type of Poster

A challenge

Primary author: DUSELLA, Gereon (DIMA@TU Berlin)

Co-authors: Dr PANDEY, Varun (DIMA@TU Berlin); Prof. MARKL, Volker (DIMA@TU Berlin)

Presenter: DUSELLA, Gereon (DIMA@TU Berlin)

Session Classification: Poster

Track Classification: Poster

Contribution ID: **23** Type: **Talk**

Snowflake Berlin

Monday, March 11, 2024 6:15 PM (30 minutes)

Im Vortrag wird Snowflake kurz vorgestellt und Herausforderungen im Bereich Datenbanken aufgezeigt, an denen wir derzeit arbeiten. Auch kurz das Snowflake Academia Programm wird vorgestellt.

Type of Poster

Presenter: JUNGHANNS, Dirk (Snowflake)

Session Classification: Talks

Track Classification: Vortrag

Contribution ID: **24** Type: **not specified**

Sitzung der GI Fachgruppe Datenbanken

Type of Poster

Presenter: KLETTKE, Meike

Session Classification: Fachgruppentreffen

Track Classification: Vortrag

Contribution ID: **25** Type: **not specified**

Wikidata as a FAIR and multilingual interface to the research ecosystem

The practice of data sharing is slowly but surely reaching further and deeper into scholarly realms, and an increasing share of such data meets at least some of the FAIR Principles. On that basis, it is increasingly possible for research data to be found and used by people and processes with no close relationship to the original research context, which opens up both opportunities and challenges. This contribution focuses on opportunities, specifically those provided by integrating research resources with Wikidata, a FAIR and public-domain database providing general reference information spanning all domains of scholarly knowledge. Aligned with thousands of databases and with terminologies in many languages, it can serve as a multilingual interface to the research ecosystem, including to resources originally available in a very small number of languages or even just one.

Type of Poster

A solution

Primary author: MIETCHEN, Daniel (MaRDI & Wikimedia)

Session Classification: Poster

Track Classification: Poster

Contribution ID: **26** Type: **Poster**

Ocient Hyperscale Data Warehousing

tbd

Type of Poster

A solution

Primary author: STOLZE, Knut (Ocient)

Session Classification: Poster

Track Classification: Poster

Contribution ID: **27** Type: **not specified**

Towards FAIR Data in Legal Domain

In this work, we explore the pivotal role of legal interoperability in facilitating the sharing and reusability of data across diverse domains. In particular, we focus on the challenges within the legal context, delving into issues related to diverse data types, potentially sensitive information, copyright concerns, and licensing intricacies. This work navigates through the complexities of implementing FAIR principles in the legal domain, emphasizing the need for a robust legal framework. We highlight the importance of legal instruments such as regulations, directives, policies, and international agreements, the presentation elucidates how these instruments guide critical aspects such as data ownership, licensing, and data protection. These measures ensure the alignment of legal frameworks with FAIR principles, fostering a culture of responsible and interoperable data sharing.

Type of Poster

A challenge

Primary authors: Prof. ALGERGAWY, Alsayed (University of Passau, Chair for Data & Knowledge Engineering); Mr KIRGEYEV, Bakhtiyar (University of Passau, Chair for Data & Knowledge Engineering)

Session Classification: Poster

Track Classification: Poster

Contribution ID: **28** Type: **not specified**

Welcome

Monday, March 11, 2024 1:00 PM (15 minutes)

Presenter: KÖNIG-RIES, Birgitta (Heinz Nixdorf Chair for Distributed Information Systems)

Session Classification: Talks

Contribution ID: **29** Type: **not specified**

Flashtalks

Tuesday, March 12, 2024 9:30 AM (15 minutes)

1 Minute Teasers presenting the posters

Session Classification: Talks