Federated Infrastructures Report 2024

DIG-UM Annual Meeting 2024

Kilian Schwarz and Matthias Hoeft

DIG-UM

Federated Infrastructures – The mission

- Fostering the federation of compute infrastructure in order to enable data taking, data processing and data archiving – this includes large data volumes and data of large diversity – including the required network backbone with sufficiently high bandwidth
- Creation of a ErUM wide federated science cloud with large central commonly used computing infrastructures, automation and workflows transparent to users, easy findability and access of data, computing and workflows, using standardised, preferably industry compatible tool sets and a single sign on infrastructure (AAI = Authentication and Authorisation Infrastructure)
- We do that in close collaboration with the other DIG-UM topical groups, with Helmholtz DMA ST1 and NFDI (especially PUNCH4NFDI and the MC WG)

Federated Infrastructures – current status

- In the mailing list there are about 20 members
  - from 6 ErUM communities
    - still dominated by particle physics
  - from 11 institutes
    - slightly dominated by DESY
- Meetings mainly on demand and predominantly virtual

Federated Infrastructures – white paper

- In order to gather and define community requirements with respect to federated infrastructure, a white paper has been written with authors and contributions from almost all ErUM communities
- The white paper is openly accessible on the erum-data-hub web page
  - https://erumdatahub.de/federated-infrastructures/
- The white paper has been quoted by BMBF in the announcement of the call

Federated Infrastructures in Research on Universe and Matter: State of Play

Editors: Markus Demleitner (Universität Heidelberg) and Kilian Schwarz (DESY) for DIG-UM*

October 16, 2023

Abstract

As an output of the DIG-UM Topic Group on Federated Infrastructures, this document tries to provide a concise and necessarily subjective overview of the state of play of digital research infrastructures in the domains covered by DIG-UM's eight communities with a particular focus on Germany. Its main goal is to help the community members to understand the practices and technologies already established in the participating domains. It may also be useful to identify progress made as DIG-UM advances.

1 Introduction

Part of DIG-UM's mission is to improve the interoperability of the research data infrastructures in Germany within the sectors of natural sciences represented through ErUM-Data's committees. It is the purpose of this paper to investigate what contribution the topic group can (or should) make in practice.

Die forschungspolitischen Prioritäten und die wissenschaftlichen Themenfelder berücksichtigen unter anderem Empfehlungen aus den Veröffentlichungen „Challenges and Opportunities of Digital Transformation in Fundamental Research on Universe and Matter“[6] des ErUM-Data-Hubs und „Federated Infrastructures in Research on Universe and Matter: State of Play“[7] von DIG-UM (Digital Transformation on Research of Universe and Matter), der Selbst-

Federated Infrastructures – coordination meeting

- Aachen, November 4, 2024
- 29 registered participants
- Close overlap seen between FI and RDM
- Target:
  - bring together scientists and the topical groups FI and RDM
  - see how many applications exist and coordinate between FI and RDM which applications fit where
  - Give room for collaboration among consortia
  - Give room for scientists to find fitting consortia

This is an event organized by the DIG-UM Topic Group Federated Infrastructures with spport of the ErUM-Data-Hub.

The BMBF is promoting better exploitation of the scientific potential of research data by supporting the interdisciplinary development of tools and the exchange of knowledge in the ErUM-Data programme. The planned schedule for upconing ErUM data calls is as follows: Call announcement in October 2024, applications to be submitted by January 2025, funding to start in autumn 2025.

In this workshop we would like to bring together scientists who are planning to submit a proposal related to federated infrastructures or who participate already or would like to join such a proposal. In particular, we invite proposal consortia that have already formed or are in the process of forming to attend the workshop and coordinate their applications. We also explicitly invite individuals who are not yet part of such a consortium.

We plan to have introductory talks by the Federated Infrastructures and Research Data Management groups.
Established consortia are invited to present their plans and indicate areas where they see potential for collaboration with partner consortia. We also invite participants to present proposals for possible consortia that have not yet formed.

Federated Infrastructures
– examples of consortia currently applying

- Current state of applications in Federated Infrastructures
- Selected examples
  - Federated Storage Infrastructure (data lakes, efficient data access, dynamic data caches, token based file access, …)
  - SUSFECIT (Federated Compute Infrastructure, green scheduling, accounting)
  - Federated Analysis Facilities (AAI, user interface, information service, analysis environment and scaling, workflows, training)
  - Görlitz Data Centre for Astronomy (staff development, data centre operation, data transfers, software development, community relations)
  - In planning: Sustainable Data Centre
- The general idea is that consortia collaborate in order to build together a federated ErUM Science Cloud

Community feedback: Thank you Hub!

- Besides supporting the topical workshops and meetings (including locations and infrastructure), the Hub also helps to keep connections to the community and to the other DIG-UM topical groups.
- „Good outreach and impact“

Big Data Analytics Report 2024

DIG-UM Annual Meeting 2024

Jan Steinheimer and Thomas Kuhr

DIG-UM

Almost 2 years into first funding round – Big topics for the next round

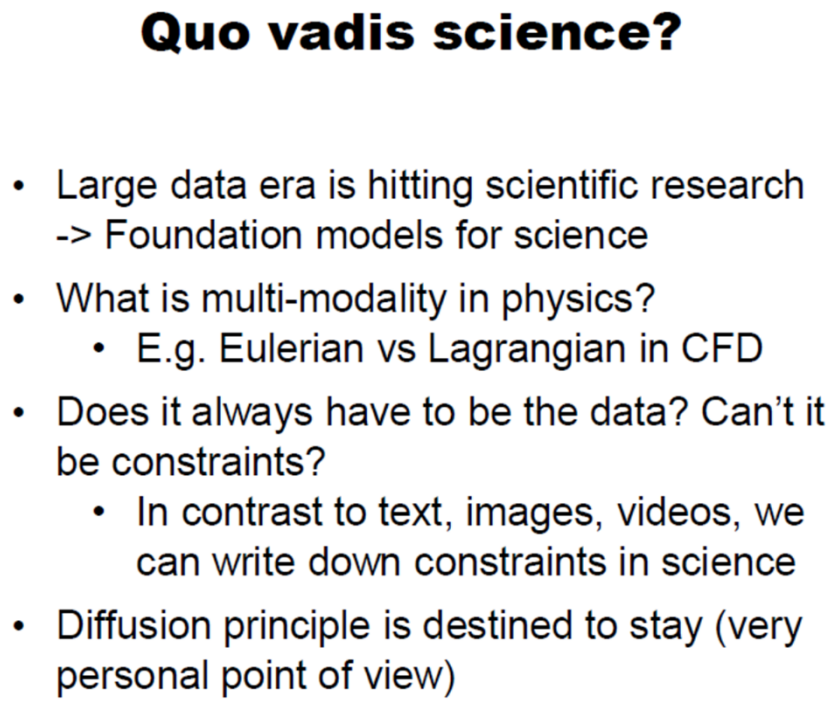
- We have 10 consortia which are running.
- Participation of all ErUM-communities.
- The majority of consortia at least planned to submit a new proposal.

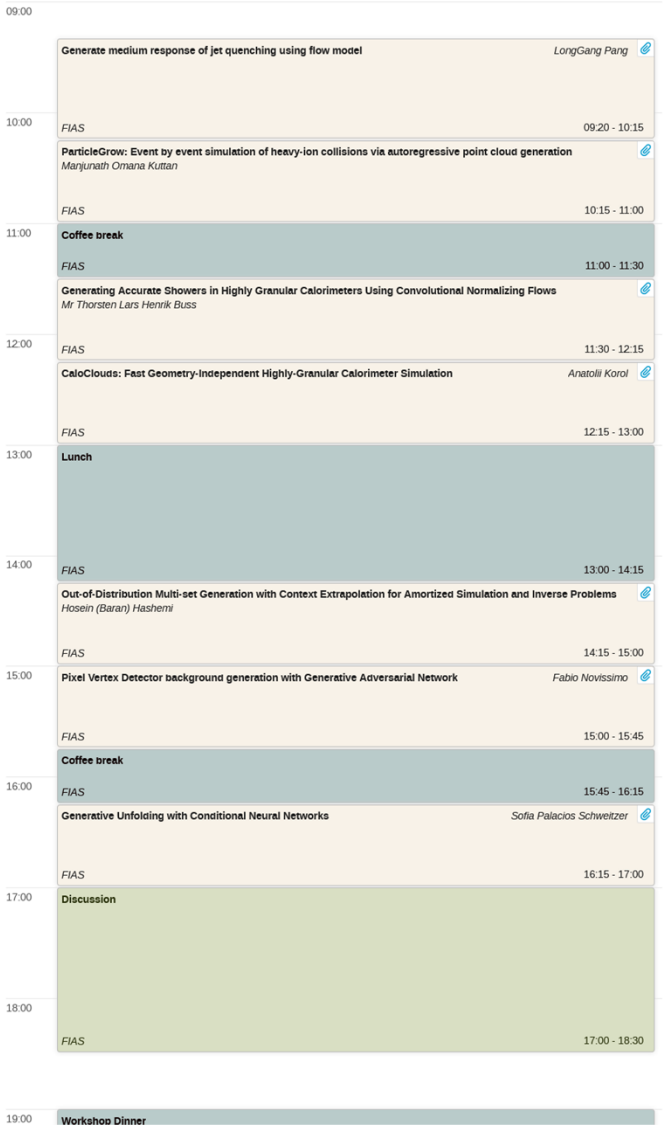
Based on workshops, TG meetings ((Bi-)Monthly TG ~ 10 participants.) and feedback on planned consortia:

- AI on dedicated hardware (e.g. FPGA) for fast and big data.
- Generative models.
- Sustainable and efficient software and algorithms.
- Large Language Models and how they can transform scientific work.

Workshops 2024 (supported by Hub and TG)

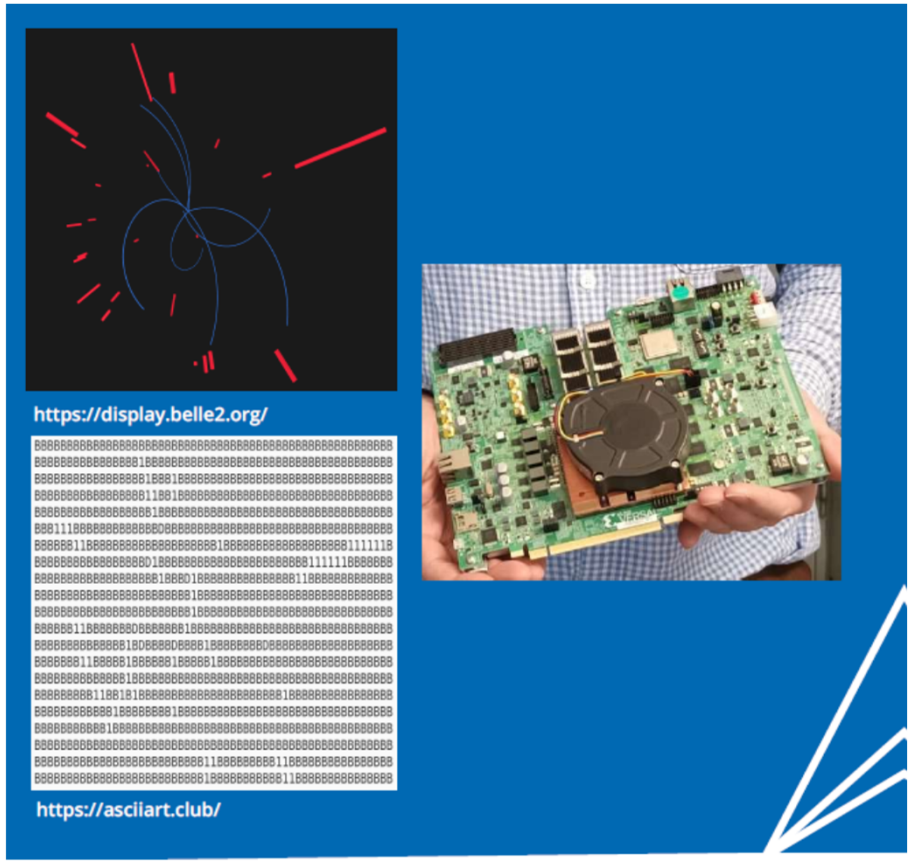
- HuB and TG supported workshops this year
- March: Generative Models at FIAS (Frankfurt)
- Contributions from various fields – growing importance





Workshops 2024 (supported by Hub and TG)

- April 8–11 : Workshop on Realtime Machine Learning (Giessen)
- Large interest in DL on FPGA for ‘realtime’ analysis.



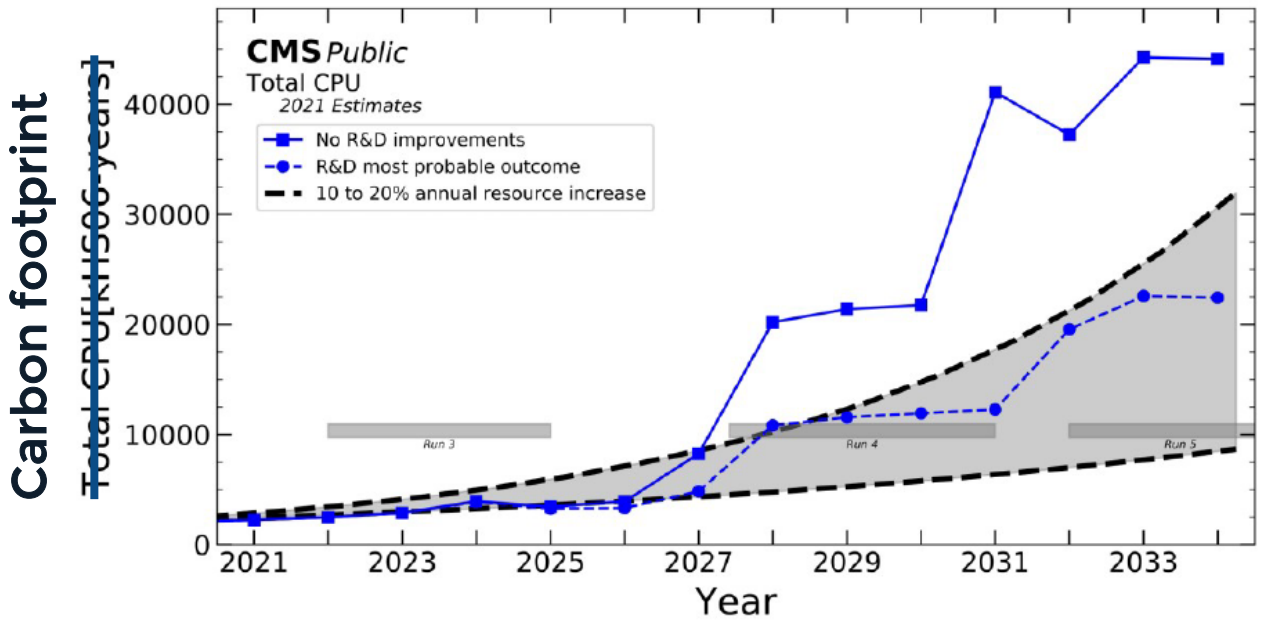
Closing comments (Sören Lange):

- Triggerless DAQ systems is a modern trend (4 experiments reported at this workshop);
- Sending data with Terabytes/s (CBM ~1 TB/s, ALICE >3 TB/s);
  pushing all data to CPU farm is tempting (“easier” than FPGA programming)
- System-on-a-chip (i.e. FPGA interfaced with ARM processors) changed the FPGA world, makes it much easier to get e.g. NN training data to the FPGA (compared to “pure” FPGAs); non-expensive boards available (ZYNQ) and AMD/Xilinx kept the approach for newer (and more expensive) platforms
- Surprising ideas presented:
  - “half autoencoder” (CMS), score is taken from the bottleneck instead of loss function from behind the decoder
  - anomaly detection using decision trees instead of autoencoder;
  very tempting, does not need multiplications (DSP slices) but only “if statements”
- Versal is the new Porsche

IG-UM

Sustainable research

- Already 2023 the issue was published.
- The next round of funding may take the suggestions into account.



Resource-aware Research on Universe and Matter: Call-to-Action in Digital Transformation

Ben Bruers[1], Marilyn Cruces[2], Markus Demleitner[3], Guenter Duckeck[4], Michael Düren[5], Niclas Eich[6], Torsten Enßlin[7], Johannes Erdmann[6], Martin Erdmann[6*], Peter Fackeldey[6], Christian Felder[8], Benjamin Fischer[6], Stefan Fröse[9], Stefan Funk[10], Martin Gasthuber[1], Andrew Grimshaw[11], Daniela Hadasch[9,12], Moritz Hannemann[8], Alexander Kappes[2], Raphael Kleinemühl[13], Oleksiy M. Kozlov[14], Thomas Kuhr[4], Michael Lupberger[15], Simon Neuhaus[13], Pardis Niknejadi[1], Judith Reindl[16], Daniel Schindler[17], Astrid Schneidewind[8], Frank Schreiber[18], Markus Schumacher[19], Kilian Schwarz[1], Achim Streit[20], R. Florian von Cube[20], Rodney Walker[4], Cyrus Walther[9], Sebastian Wozniewski[17], Kai Zhou[21]

Community feedback: Thank you Hub!

- Besides supporting the topical workshops and the annual TG meeting, the Hub also helps to make other BDA related workshops known to the community.
- „Good outreach and impact“

Next BDA annual meeting

- Save the date: March 13-14 in Munich.
- Thank you: Thomas Kuhr and Christian Stieghorst for helping with the organization.

Contact information:

https://wiki.erumdatahub.de/de/mailing-lists

Research Data Management Report 2024

DIG-UM Annual Meeting 2024
12/04/2024

Gernot Maier
for the DIG-UM – Topic Group RDM

DIG-UM

DIG-UM – Topic – Group RDM

KET

Markus Elsing
CERN

KfB

Pierre Schnizer
HZB

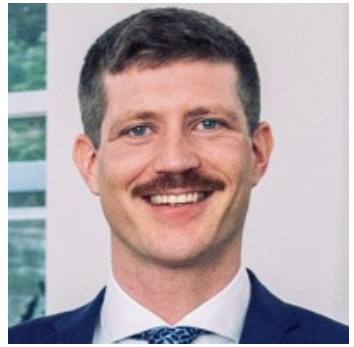
KAT

Gernot Maier
DESY

KFN

Michael Schulz
TUM

KFS



Hans-Georg Steinrück
FZJ/RWTH

Co-Chairs
since 04/2024

Former Chair
KFN

Astrid Schneidewind
FZJ

Former Deputy Chair
RDS

Monica Valencia-Schneider
U Köln

THANK YOU!!!

Mailing list

erum-data-research-data-management@lists.rwth-aachen.de

https://lists.rwth-aachen.de/postorius/lists/erum-data-research-data-management.lists.rwth-aachen.de/

DIG-UM

RDM in the ErUM context

Research Data Management (RDM)

- Highly relevant for FAIR data, open data, open science, knowledge transfer, efficient scientific progress, sustainability, …
- Subject of many complementary initiatives – NFDI, HGF, state initiatives, EOSC, …
- For efficient resources utilization: define boundaries and overlaps

In the ErUM context

Organization, preparation, and provision of data for analysis by scientists

DIG-UM

Identified interrelated needs for communities

Based on DIG-UM workshops, TG RDM meetings, and feedback from the ErUM committees (especially before BMBF-ErUM-Data-Strategiegespräch)

1. Development of (modular) systems for research data management
2. Development of systems for managing and accessing metadata
3. Development of data workflows for more efficient execution of experiments and their control
4. Development of systems for efficient life cycle management
5. Development of systems for the automation and management of access rights

Sustainability beyond the project phase & education & user-friendliness

DIG-UM

Examples of current application consortia

- **Physics-LLM**

  lead by Tim Ruhe, TU Dortmund

- **ASAP::O**

  lead by Mikhail Karnevskiy, DESY

- **HEPModel – accessible reproducible likelihood models across disciplines**

  lead by Mikhail Mikhasenko, CERN

- **Modular systems (e.g.) for FAIR data & systems for managing and accessing metadata**

  initiated by GSI

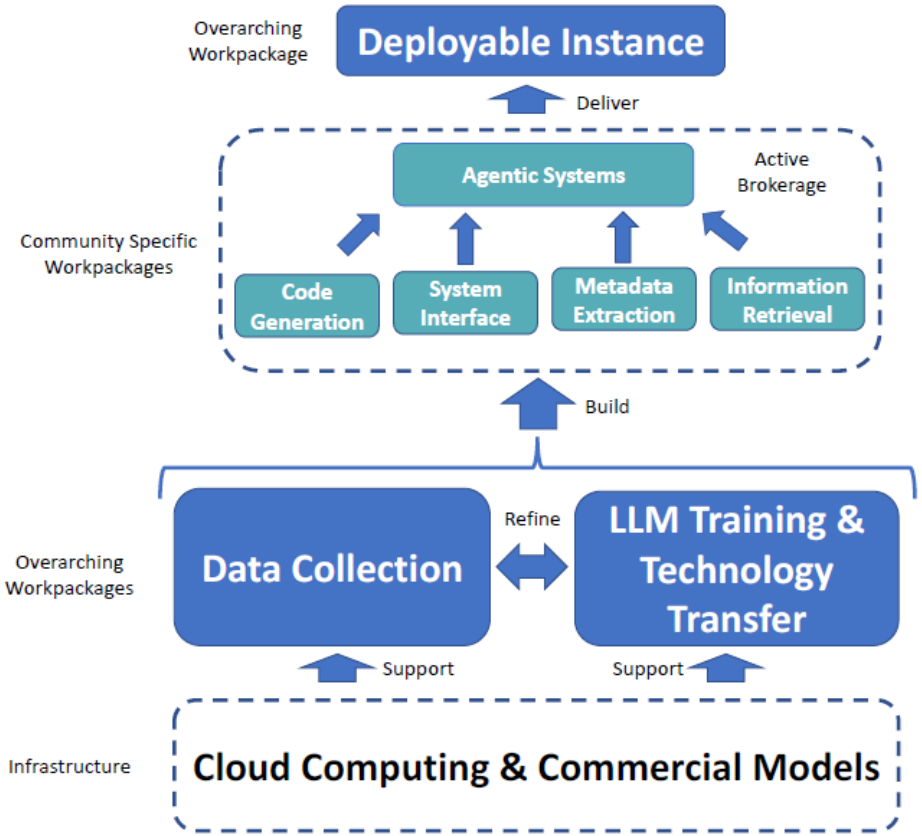
- **…**

DIG-UM

Examples of current application consortia

- **Physics-LLM**
  lead by Tim Ruhe, TU Dortmund
- ASAP::O
  lead by Mikhail Karnevskiy, DESY
- HEPModel – accessible reproducible likelihood models across disciplines
  lead by Mikhail Mikhasenko, CERN
- Modular systems (e.g.) for FAIR data & systems for managing and accessing metadata
  initiated by GSI
- …

Application of Large Language- and Foundation Models for RDM in Physics



DIG-UM

Examples of current application consortia

- Physics-LLM
  lead by Tim Ruhe, TU Dortmund
- **ASAP::O**
  lead by Mikhail Karnevskiy, DESY
- HEPModel – accessible reproducible likelihood models across disciplines
  lead by Mikhail Mikhasenko, CERN
- Modular systems (e.g.) for FAIR data & systems for managing and accessing metadata
  initiated by GSI
- …

Messaging broker system to facilitate transport of 2D detector data at high rates

State of the art of ASAP::O

- Data sent as messages by "producers" over network to distributed "receivers“
- Sharing workload of analyzing and storing these data
- Optimized for 2D synchrotron imaging & diffraction data
- Mbs at ≥ 1000Hz (data rates = Gb/s)

New developments

- Adaption to multiple other use cases from other communities
- Neutron imaging & time-of-flight diffraction, photon science
- Extension of functionality towards the synchronization of multiple data streams and higher event rates (MHz)
- Deployment and easy installation for many more applications in the future

DIG-UM

Thank you
& the ErUM-Data-Hub

Mailing list

erum-data-research-data-management@lists.rwth-aachen.de

https://lists.rwth-aachen.de/postorius/lists/erum-data-research-data-management.lists.rwth-aachen.de/

Annual Workshop

- Welcome
- DIGUM / Erum Data Intro

Pierre Schnizer, Tim Ruhe et al.
für DIGUM Topic Group User Interface

DIG-UM

User Interface: Motivation

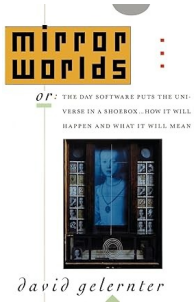
Src: ancient stone bridge in peoloponnesos, athensgreece.com

We are drowning in information
we are starving for knowledge

Rutherford D. Roger

Mankind was always better in acquiring data than interpreting it.

-- Mirror Worlds
David Gelernter



User Interface: Motivation

- Large scale data → exponential growth rate
- HEP phyics data rates → other ERUM Communities

- Challenges for users (typical response)
  - Get their data
  - Reduce their data
  - Archive their data
  - Get measurement result

- Users: have fairly good idea of their job
- Needed
  - Standardised task → off load
  - Clean, stable interface
  - „Hang their job on the trellis”

Good User Interfaces:
- Lower entry barrier
- Broadens user base
- Simplifies their job execution

Dieter Rams: Principles of good design
- innovate
- Maked product useful & understandable
- Unobstructive / honest
- Long lasting
- Thought trough to the end

Workshop: work flow engines

Reana:
„deCERNified“ for Astronomic community

Operated at AIP Kubernetes

User test case:
- Decoding timepix 3 data stream
- Using Reana workflow engine

Results
- User: straightforward to manage
- Split up of task
- Scalable to large data sets

- Straightforward to learn

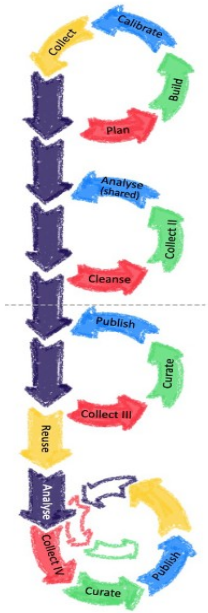
Use cases: e.g. KfS
- Experiments roughly a week
- Data rate Gb/s, Evaluation: combine TB
- Analysis chain:
  - Online analysis → adjustment of measurement plan
  - Offline
- Review and reevaluate e.g. after 6 month

Workshop: work flow engines

Harry Enke, Using diverse and distributed infrastructure in PUNCH4NFDI

What facilitates user interaction?

Data life cycles in Physics : not really cycling the same data



- Instrument conception, construction
  - expected observational properties
  - dvelopment of instrumental capabilities
  - construction
- Operations : Data taking
  - Instrument calibration
  - Operation modes
  - Quality monitoring
- Data evalutaion / calibration …
- Data curation
  - Data evalutation,
  - Removal of instrument charcteristics + artifacts
  - Data curation / producing science ready data (NFDI)
- Data exploitation
  - Public (or group based access) data exploitation
  - FAIR data domain. (NFDI)
  - Interdisciplinary exploitation (NFDI)

PUNCH4NFDI | ErUM Data UI | 17/09/2024

Workshop: large language models

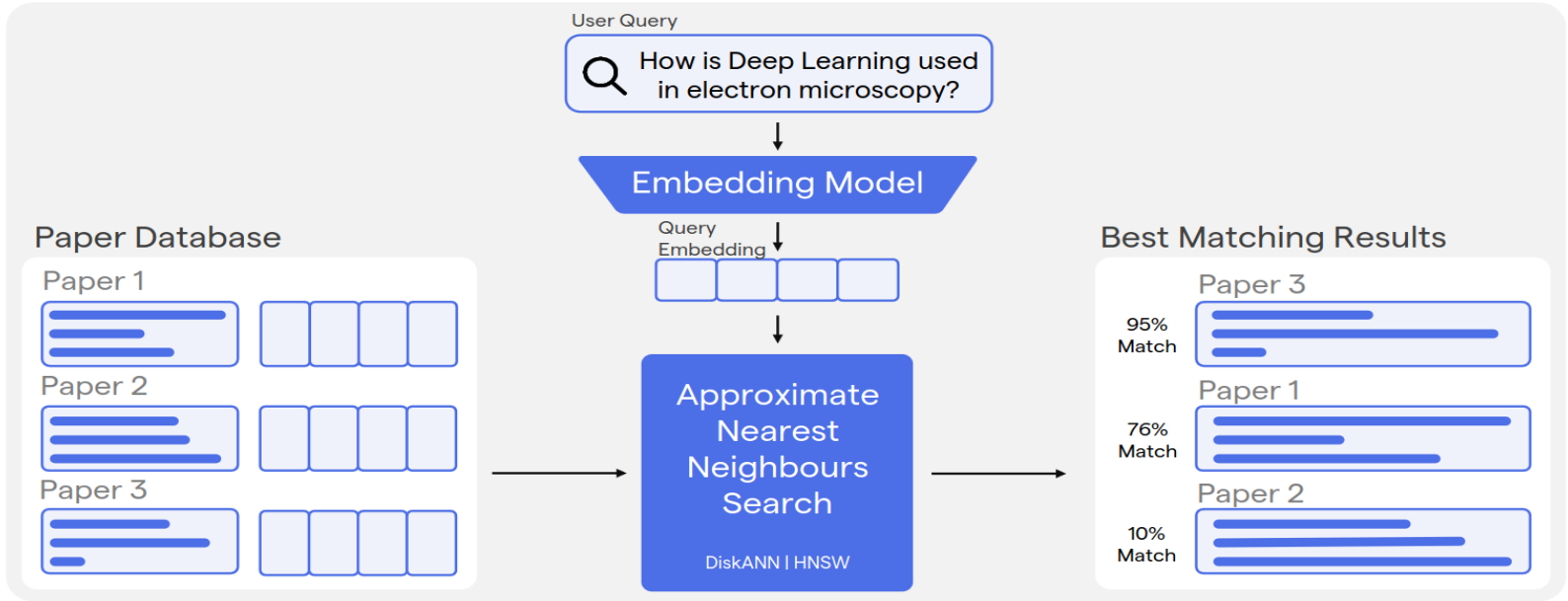
Target:

- Interact with the computer in natural language
- Make it digest your data

- Thus:
  - Intelligent prompt design
  - Follow up on relevant material

Computers are useless.
They can only give you answers

Attributed to Pablo Picasso



User interface: Quo vadis?

Next steps

- Make work flows wide spread
  Lower their entrance barrier
- Online analysis of 1 –100 GB/s data streams
  → accessible to „university research group teams“
- Empower the users

Human computer interface

- Science question in natural language
- **R**ead-**E**valuate-**P**rint-**L**oop redesign?
  - Tell computer what has to be done
  - Computer responds how it wants to do it
  - Gives reasoning and credibility analysis of its access
  - Gives alternatives
- Let user be “decision maker”

How Can I Support the Effort?

Bi-weekly Zoom-Meetings!

Join the Email List:

https://lists.rwth-aachen.de/postorius/lists/erum-data-user-interface.lists.rwth-aachen.de/

https://www.flaticon.com/free-icons/knowledge-sharing

Knowledge Distribution

Judith Reindl (KfSI)

&

Dirk Lützenkirchen-Hecht (KFS)

DIG-UM

der Bundeswehr Universität München

BERGISCHE UNIVERSITÄT WUPPERTAL

Topics for future workshops

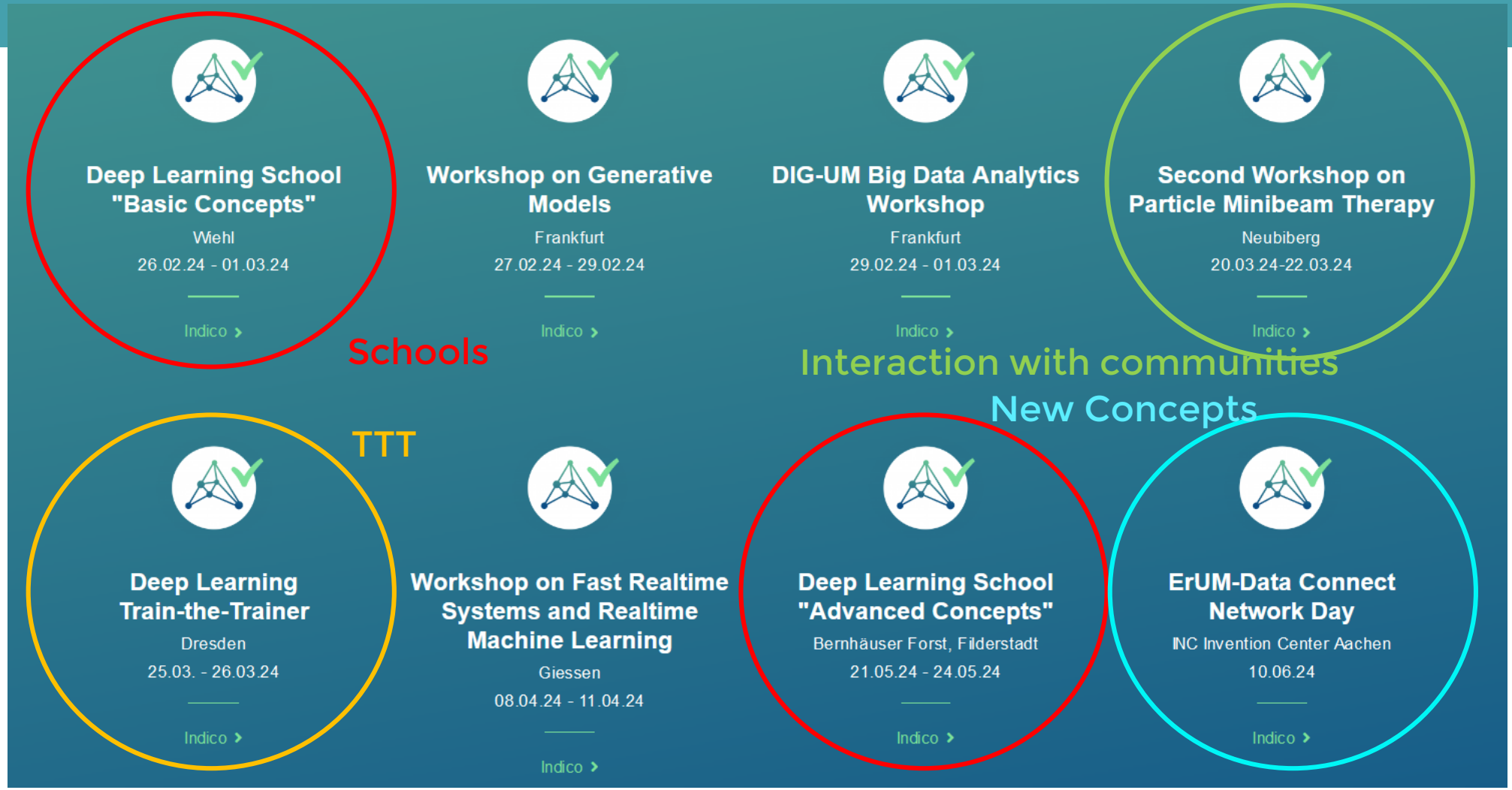
- Continue existing activities (Deep-learning schools, TTT wo...

- Start new formats:
  - Programming schools
  - Developers Workshops
  - AI-generated programming code
  - ...

Your input is needed!

Last slide from DIG-UM Annual Meeting 2023

Retrospective – Workshops & Schools 2024



Retrospective – Workshops & Schools 2024

Retrospective 2024

2024: Again 15 Events with >500 participants

2 deep learning schools

7 expert workshops

1 Train the Trainer workshop

+ 2 programming events

+ new formats

+ online events

Geographic proximity influences attendance

Still an issue – but due to broader distribution (north /south / east) + online events: ⇒ more participants from underrepresented regions

Participants do not represent the ratio of members in the ErUM communities

Researching the needs of under-represented communities, especially KFN and KFS, interaction with communities

⇒ Better in 2024 – more specialized topics

(e.g. Machine Learning School for X-ray & Neutron Science 9/24, KFSi-Minibeam Workshop 3/24)

Outreach

Involving the general public

Podcast

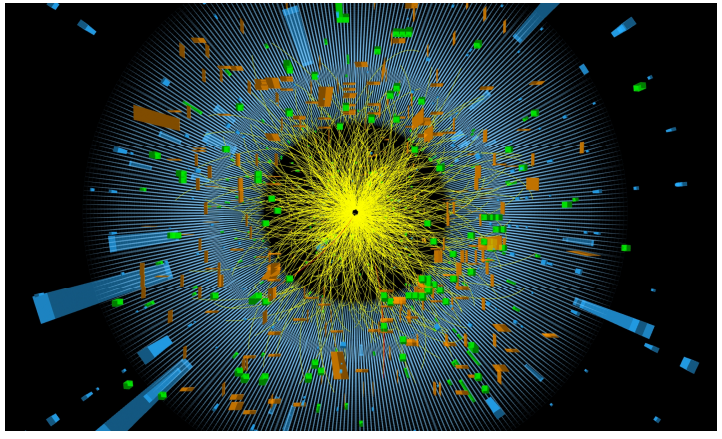
Anchoring knowledge distribution in the tenders

Two more are still to come in 2024

Retrospective – Strategy Meeting 23./24.1.2024

… many very exciting & constructive discussions with BMBF & PT DESY

Topics: - sustainability (… re-use, synergies, added values, collaborations (e.g. NFDI))
- availability
- metadata
- data management
- teaching activities
- …



Bundesanzeiger
Herausgegeben vom
Bundesministerium der Justiz
www.bundesanzeiger.de

Bekanntmachung
Veröffentlicht am Freitag, 11. Oktober 2024
BAnz AT 11.10.2024 B1
Seite 1 von 11

**Bundesministerium
für Bildung und Forschung**

**Richtlinie
zur Förderung von Verbundforschungsvorhaben zu den Themenfeldern
„Software und Algorithmen mit Schwerpunkt auf Künstlicher Intelligenz und Maschinellem Lernen“,
„Forschungsdatenmanagement“ und „föderierte Digitalinfrastrukturen“
zur Erforschung von Universum und Materie (ErUM)**

Vom 30. September 2024

1 Förderziel, Zuwendungszweck, Rechtsgrundlage

1.1 Förderziel

An den großen Forschungsinfrastrukturen der naturwissenschaftlichen Grundlagenforschung werden Erkenntnisse gewonnen, die einen wichtigen Beitrag für unsere langfristige Zukunftsvorsorge bilden. Sowohl die wissenschaftlichen Ergebnisse als auch die entwickelten Technologien und Methoden bilden die Grundlagen für Innovationen von morgen und übermorgen. Auch für den digitalen Wandel nimmt die Erforschung von Universum und Materie eine Schlüssel-

DIG-UM

Retrospective – Strategy Meeting 23./24.1.2024

Call from BMBF (30.09.2024) / page 4:

Arbeitspakete zu den Maßnahmenfeldern „Qualifizierungsangebote“ und „Austausch, Kommunikation und Transfer“ des Aktionsplans ErUM-Data können beantragt werden. Dies berücksichtigt unter anderem mögliche Formate wie Meetings, Konferenzen und Workshops sowie Personalaustausch zwischen Hochschulen, Forschungsinfrastrukturen, Rechenzentren sowie der Wirtschaft.

– Zu den „Qualifizierungsangeboten“ zählen unter anderem Ausbildungsmaßnahmen für Nachwuchswissenschaftlerinnen und Nachwuchswissenschaftler sowie Weiterbildungsmaßnahmen für etablierte Wissenschaftlerinnen und Wissenschaftler, die sich auf ErUM-spezifische Aspekte oder Nachhaltigkeit beschränken. Dabei sollen auch moderne Formate aufgegriffen werden. Die Einbindung von Nachwuchswissenschaftlerinnen und Nachwuchswissenschaftlern in koordinierende Aufgaben wird ausdrücklich begrüßt.

– Zu „Austausch, Kommunikation und Transfer“ zählen Aktivitäten, die sowohl den Austausch mit den relevanten Akteuren als auch mit der Öffentlichkeit steigern. Dies kann unter anderem durch Arbeitstreffen, Workshops und Konferenzen oder auch Gastaufenthalte bei Verbundpartnern erfolgen. Für den Dialog mit der Öffentlichkeit sind verschiedene Maßnahmen und Formate möglich, von erprobten und innovativen Formaten der Wissenschaftskommunikation bis zu Projekten im Bereich der Bürgerwissenschaften.

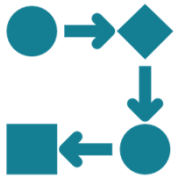
If you apply, don‘t forget to apply for related funding!

Preserving knowledge and promoting young talents

Data-Science topics in Universities

- anchoring of related content in the degree programmes
- Teaching digital skills
- Start as early as the Bachelor's programme

(e.g. strategy meeting with KFS, DFG, BMBF in Bad Honnef, stressing the need of “Data Experts”)



“tailored-learning” Concepts

Start with materials collections on the ErUM-Data-Hub!

- Problem-based learning
- …

Qualification of the next generation

- Coaching offers
- Help for self-help
  (e.g. working out loud)
- …

Topics for future workshops

Continue & strengthen existing activities:

- Deep-learning schools
- TTT workshops
- expert workshops
- Programming schools
- Developers Workshops

Develop new activities:

- workshops for specialized communities
- tools / workflows for education
- …

Your input is needed!