Zoommeeting Leitungen der Topic Groups 17.3.2022, 8:30-9:30

FI, Kilian Schwarz, Markus Demleitner, BDA -, RDM, Astrid Schneidewind, Monica Valencia-Schneider, UI, Tim Ruhe, KD Dirk Lützenkirchen-Hecht, Katrin Link, SP, Martin Erdmann, KET Günter Quast

Kurzbericht BMBF Review ErUM-Data-Hub 8-Mar-22

Teilnehmer:innen BMBF: Dr. V. Dietz, E. Lilienthal, R. Dieter, D. Kunold; PT.DESY: Dr. W. Ehrenfeld, Dr. D. Hoppe; Prisma: Prof. Dr. H. Dosch - DESY, Prof. Dr. E. Elsen - DESY, Prof. Dr. R.-D. Heuer - CERN, Prof. Dr. S. M. Schmidt - HGF Dresden-Rossendorf, Prof. Dr. M. Tolan – Göttingen, Prof. Dr. B. Kanngießer – Berlin, Prof. Dr. J. Wambsganß – Heidelberg; ErUM-Data-Hub: Dr. B. Murphy – Kiel, Prof. Dr. U. Katz - Erlangen-Nürnberg, Prof. Dr. M. Erdmann Aachen

Insgesamt gab es positives Feedback zu den bisherigen Aktivitäten und den aktuellen Planungen des ErUM-Data-Hubs:

1. Der Bereich Aus- und Weiterbildung „Digitale Kompetenzen stärken“ wurde allseits sehr stark unterstützt und Impact zugetraut.
2. Die Gestaltung der Schnittstellen Helmholtz-Zentren und Universitäten ist noch nicht ausgereift.
3. Bei NFDI & ErUM-Data müssen wir Überlapp und Synergie klären, um keine divergierenden Entwicklungen zu erzeugen.
4. Sehr positiv haben wir aufgenommen, dass die genaue Verwendung der Mittel Flexibilität hat und weiter besprochen werden kann.
5. Am Projektende müssen klar verständliche „Deliverables“ formuliert werden.
6. Erwartungsmanagement: dem BMBF nur „realistisch Machbares“ in Aussicht stellen.

Als Konsequenz zu Punkt 3. soll im nächsten Digitization Board 7.4.22 (direkt nach dem NFDI-Meeting am 5.4.22) eine Agenda für ein gemeinsames Meeting ErUM-Data & NFDI formuliert werden und die Planungen für ein Meeting im September initiiert werden.

Status, plans, progress on Federated Infrastructures: Kilian Schwarz, Markus Demleitner

Fokus dieses Meetings waren die Themen und Verknüpfungen der Topic Group Federated Infrastructures. In einem gemeinsamen Effort mit dem Resource Provider Board entsteht ein detailliertes Dokument über Bedarfe und vorhandene Ressourcen in den 8 ErUM-Komitees. Dieser Prozess ist dank des großen Engagements der Beteiligten schon weit gediehen und enthält zahlreiche Fakten als substantielle Informationsbasis. Daraus lässt sich einerseits die Heterogenität der Communities ablesen, andererseits sind aber bereits gemeinsame Herausforderungen identifiziert.

Mehrere Thesen wurden in der Diskussion formuliert:

1. Unsere Kernaufgabe als Leitungspersonen der 5 DIG-UM Topic Groups ist es, ein strategisches Gedankengut zu entwickeln, durch dessen zukünftige Umsetzung die ErUM-Forschung über Maßnahmen der Digitalen Transformation noch bessere Fortschritte machen kann.
2. Welche Strukturen (Projekte der NFDI, EOSC, ErUM-Data etc) die Strategien in konkrete Maßnahmen umsetzen, ist nicht entscheidend. Entscheidend ist das Endergebnis zum Wohl der Forschung in ErUM.

3. Wir benötigen Themen-geführte Leitgedanken insbesondere für NFDI und ErUM-Data, um in den Entwicklungsarbeiten voneinander zu profitieren, das Damokles-Schwert einer Doppelförderung zu vermeiden und keine Doppelstrukturen zu verfolgen.
4. Die breit gestreuten Fördermittel sind herausfordernd. Einerseits ist es von Vorteil, viele Forscher:innen an den Entwicklungsprozessen zu beteiligen. Andererseits müssen wir als inhaltlich strategische Leitung genau im Blick behalten, wie wir das unter 1. formuliertes Ziel umfassend erreichen können. Alle Förderungen und Projekt-Entwicklungen sollen dafür möglichst gut zusammenpassen.
5. Integrationen verlaufen natürlicherweise vertikal, weil wir in internationalen Experimenten und Umgebungen arbeiten: von ganz großen Strukturen reicht die Integration zu den einzelnen Forscher:innen in den ErUM-Arbeitsgruppen. Das BMBF versucht eine horizontale Integration durch die Verbindung sehr heterogener ErUM-Communities. Die Umsetzung horizontaler Strukturen erweist sich als möglich, ist allerdings mit großen Herausforderungen verbunden, wie wir alle gemeinsam in DIG-UM erleben.
6. Synergien und Schnittstellen zwischen den ErUM-Topical WGs müssen in bilateralen und multilateralen Gesprächen gefunden werden. Ebenso gibt es Synergien mit der NFDI und entsprechenden Arbeitsgruppen, die gelebt werden müssen. Einen ersten NFDI-Workshop mit ErUM-Beteiligung gibt es am 5. April (https://indico.desy.de/event/33410/)
7. Hinweis auf kanadische Strukturen https://www.canfar.net/en, https://www.computecanada.ca/home

Nächstes Meeting: wegen des Train-the-Trainer Workshops am 31.3. findet das nächste Digitization Board Meeting am 7.4., 8:30 statt.

ErUM-Data/DIG-UM
Federated Infrastructures meets DIG-UM

Kilian Schwarz
Markus Demleitner

Federated Infrastructures Mission

- Providing a distributed computing 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
- In future German computing resources need to support a larger number of users and therefore need to be combined to federations. ErUM scientists have to be able to access and use these resources in an easy and efficient way. The mid term target therefore needs to be to create an ErUM Data Science Cloud consisting of such a merger of German computing resources.

Federation

- How do we understand federation
  - 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)
- Why do we want to federate ?
  - Synergy
  - Cost efficiency
  - Make it easy to exploit a wider range of different resources
  - Facilitate data sharing
  - Optimise resource usage by increasing the number of potential users
  - Optimise resource usage by increasing the number and diversity of use cases
  - Avoid lock-in to specific providers
  - Avoid trouble due to localised funding issues
  - Global trend
- What to consider
  - Diversity of data sets and experiments
  - Interoperability
  - Connections to NFDI, EOSC, synchronisation of many parallel efforts, industry

Initial meeting

- On November 30, 2021
- Participants:
  - Kilian Schwarz (KhuK)
  - Markus Demleitner (RDS)
  - Alexander Schmitt (KET)
  - Jacob van Santen (KAT)
  - John Bulava (KAT)
  - Simon Heybrock (KFN)
- Joined later
  - Anton Barty (KFS)
- Missing:
  - KFSI
  - KFB

New members since community meeting according to mailinglist

- Andreas Haungs, KIT, KAT
- Erik Bründermann, KIT, KFB
- Boyang Yu, Uni-München
- Christoph Wissing, DESY, KET
- Dirk Duellmann, CERN, IT
- Oliver Freyermuth, Uni Bonn, KET
- Wienemann, Uni Bonn, KET
- Hartmut Stadie, Uni Hamburg, KET
- Johannes Lange, Uni Hamburg, KET
- Bridget Murphy, Uni Kiel, KFS
- Thomas Kreß, RWTH Aachen, KET
- Thomas Kuhr, LMU München, KET
- Wolfgang Ehrenfeld, DESY

erum-data-federated-infrastructure@lists.rwth-aachen.de

Concrete work program

- We decided to start with writing a document collecting the state of art in our 8 comittees
- In this document we would like to identify
  - Current computing and storage infrastructures
  - Already federated infrastructures
  - Current issues which need to be adressed
  - Infrastructures which are planned to be federated

Work in progress

- Starting point has been made by RDS
- File template exists
- communities started writing
- Document repository exists in github
  - https://github.com/msdemlei/dig-um-fi-init
- Plan: March 18 document build, then circulate via mailing list

Status of communities
KAT

- Established Gamma-ray observatories (H.E.S.S., VERITAS, …) store data at few sites
  - Data are accessed and transfered conventially using tools as ssh, ftp, Globus, ..
- Other observatories (CTA, Icecube, …) make heavy use of the US based OpenScience Grid
- CTAO not yet decided, but DESY will be one of the CTAO data centres
- User identification process with Grid certificates still cumbersome

Status of communities KET/KHuK

- Existing federated infrastructures
  - WLCG
    - world‘s largest computing Grid, 170 computing centres in more than 40 countries
    - Supported by EGI (EU), OSG (US), regional Grids
- Technologies employed
  - Most important middleware stacks in use:
    - ARG, gLite, UNICORE, dCache, Globus Toolkit, XrootD
- Current issues/future challenges
  - Data volume and complexity will increase dramatically
  - Approach: concept of a science cloud with IaaS
- Infrastructures to be federated
  - Large computing centres with dedicated CPU and storage systems
  - Opportunistic resources at HPC centres, commercial Cloud systems, scientific Clouds, GPU clusters
- Future technologies
  - Are being developed in funded ErUM data projects as e.g. IDT-UM and FIDIUM

Status of communities KFN

- Current compute and storage infrastructures
  - ESS:
    - Slurm batch cluster with 99 compute and one GPU node
    - 2.1 PB file systems (ZFS/GPFS)
- Already federated infrastructures
  - None, but access via ESS based AAI
- Current issues
  - Unifying AAI backend at ESS
- Infrastructure to be federated
  - Transparent access to ESS resources via AAI
  - Federating storage backends
  - Data portal to search and access data following FAIR principles

Status of communities KFS

- Current compute and storage infrastructures
  - Largely seperate infrastructures on-site of the facilities
- Already federated infrastructures
  - DESY and European XFEL host hardware in same data centre at DESY
    - Shared:
      - AAI
      - File transfer services
      - Tape archiving systems
    - Compute and storage systems logically separated but
      - Both GPFS and dCache
      - Shared Maxwell cluster
- Current issues
  -
- Infrastructure not (yet) federated
  - Resources at home institutions as
    - Universities, MPG, industry

Status of communities RDS

- Existing federated infrastructures
  - Virtual Observatory (VO)
    - Federated data infrastructure
    - Registry/global discovery
    - Some 25000 interoperable services
    - Hundred of millions of data sets
    - Standards for searching and manipulation of data, interacting with data bases, data formats and semantics
  - Technologies employed
    - Relational data bases (Postgres), metadata scheme, end-user components
  - Current issues
    - How to deal with data which are too large to move ?
    - Interoperable techniques for collections of arrays
    - How to deal with 1000 data collections behind a single access URL
    - Data modelling for more complex data structures
    - Institutional basis for German contribution
  - Infrastructures to be federated
    - Compute platforms next to immovable data should be made interoperable and discoverable
    - authentication

Status of communities identification of common interests

- Federated and interoperable AAI
- Token based user identification
- Federated data storage infrastructure (Data Lake ?)
- How to deal with large data volumes
- ...

Status of discussion at Community meeting

- About 20 participants
- Gathering community overarching requirements
  - Fed-Inf is producing a corresponding document
  - Resource and Data Provider Board does a similar thing
  - Both shall be presented as basis for requirement investigation to DB
- Wish list
  - Federated Science Infrastrutures (Compute & Storage) with all resources from all communites usable by all communities
  - Accounting
  - Federated data centre (curated data sets federated and interoperable)
  - Federated user access (AAI)
- Important: collaboration with NFDI and related projects
- In meeting also MERIL was mentioned (http://portal.merit/eu/merit/)

synergies

- Borderlines to topical groups
  - BDA
  - RDM
- Synergies with other topical groups
  - BDA, RDM, UI, knowledge distribution
  - BDA: software & algorithms need to run on our FI
  - RDM: we need to do RDM on our FI
    - Bilateral meeting was requested by FDM
  - UI: users need to interact with our infrastructure via UI
    - Concrete use case: AAI
    - What can be handled with AAI groups and federated AAI ?

Overlap with NFDI

- NFDI is currently creating a consortium for base services. One of them probably AAI
- Largest overlap with section „common infrastructures“
- And the working groups therein:
  - Data integration (Data Lake), identity management (AAI), Federated Multicloud

NFDI/Federated Multicloud

- Cross community access to data
- Interoperability
- Compatibility
- Common platform for data handling and management
- Federated decentralised solution preferred
- Common set of standards (APIs and tools)
- Solution should follow FAIR principles
- Data privacy and protection rules should follow highest standards

==> large overlap: propose to invite representative of NFDI group to our meeting
Need to make sure that things fit together

Synergies with NFDI

Workshop on NFDI tools / services / synergies between physics-related consortia and others

Tuesday Apr 5, 2022, 9:00 AM → 5:00 PM Europe/Berlin

Description In this 1-day workshop, representatives of the physics-related and other NFDI consortia (PUNCH4NFDI, DAPHNE4NFDI, FAIRMat, NFDI4Phys, but also MaRDI, NFDI4Culture etc.) will discuss among themselves and with representatives from ErUM/DIGUM, Helmholtz and others the issue of tools, synergies and services in the NFDI and in the relevant scientific disciplines.

The goal of the workshop is to foster the exchange on existing competences and solutions and, based on this fuller picture, to discuss knowledge exchange and the development / setup of common tools and services. The workshop will consist of presentations introducing competences and solutions with ample discussion time, and it is intended as a start of a regular productive exchange between NFDI, ErUM etc. Main topics to be addressed will be (but open for further suggestions):

- Analysis workflows and distributed computing
- Software repositories for common software solutions
- Metadata schemas
- AAI and other collaborative tools

More information coming soon.

Please register - the ZOOM connection will be distributed only to registered participants.