ErUM-Data/DIG-UM Federated Infrastructures

Kilian Schwarz
Markus Demleitner

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

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 the 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.

*: Erforschung von Universum und Materie

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 = Authentication and Authorisation Infrastructure)
- Why do we want to federate ?
  - Synergy
  - Common use of infrastructure/software/tools
  - 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 Nationale Forschungsdateninfrastruktur (NFDI), European Open Science Cloud (EOSC), synchronisation of many parallel efforts, industry, existing board, organisations, communities

synergies

Synergies with other topical groups

- Big Data Analytics (BDA), Research Data Management (RDM), User Interface (UI), knowledge distribution
- BDA: software & algorithms need to run on our Federated Infrastructure
- RDM: we need to do RDM on our Federated Infrastructure
  - Bilateral meeting was requested by RDM
- UI: users need to interact with our infrastructure via UI
  - Concrete use case: AAI
  - What can be handled with AAI groups and federated AAI ?

members

- On the mailing list about 20 members
  - from 6 communities
    - dominated by Particle Physics
  - from 11 institutes
    - dominated slightly by DESY

Concrete work program

- We wrote a document collecting the state of art in our 8 comittees
- In this document we identified
  - Current computing and storage infrastructures (to be identified in collaboration with Resource Provider Board)
  - Already federated infrastructures
  - Current issues which need to be adressed
  - Infrastructures which are planned to be federated
  - Basis for future discussions
- The document will be published on the Erum-data-hub web page
- And lateron in Springer journal „Computing and Software for Big Sience“
- In 2023 a workshop on „federated infrastructures“ is planned

Status of communities identification of common interests

- How to deal with large data volumes
- …
- And wish list
  - Federated Science Infrastrutures (Compute & Storage) with all resources from all communites usable by all communities
  - Accounting
  - Federated data infrastructure (curated data sets federated and interoperable)
  - Federated user access (AAI)
    - Token based user identification

- Important also:
  - collaboration with NFDI and related projects
  - Close coordination with existing boards and communites (e.g. GridKa OB)

White paper - introduction

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

DIG-UM Topic Group Federated Infrastructures

December 5, 2022

Abstract

As a first 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 physics represented through ErUM-Data's committees. It is the purpose of this paper to investigate what this can (or should) mean in practice.

Interoperability between data services, the necessary condition of federation, is of course a very desirable property from a user perspective. Services with common interfaces mean that users will not have to learn new techniques when moving between service providers. It means that their software continues to work as they use data and services from different sources, quite typically also that they get to choose between multiple implementations of a standard (e.g., in different languages, on the

White paper - conclusion

7 Conclusions

A first analysis of the provided answers shows already that a federated and interoperable authentification and authorisation infrastructure (AAI), and a federated data infrastructure as, e.g., a Data Lake and an understanding how to deal with large data volumes is prioritised very highly by many ErUM-Data communities. This is supported by the upcoming challenges provided especially by the HiLumi-LHC. Here the communities can and must learn from each other considering also the fact, that the state of having federated infrastructures up and running and also the focus on what federated infrastructures have been and still need to be implemented is quite different in the participating communities. Moreover, the federation of infrastructures needs to be informed by requirements and constraints of other DIG-UM Topic Groups. For instance, an analysis workflow designed within Big Data Analytics needs to take into account the capabilities of a federated infrastructure; metadata generated through Research Data Management will most certainly help implementing useful and rich archive systems.

One should also consider here that going from generic (bytes) via structured (formats) to disciplinary (data models) this means decreasing ease of federation (anyone