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

Markus Demleitner (University of Heidelberg) and Kilian Schwarz (DESY)
in the context of
DIG-UM Topic Group Federated Infrastructures

February 4, 2023

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 local machine vs. in the "cloud", etc). Very generally, interoperability reduces lock-in to particular service providers and thus also increases the ability of researchers to (re-) use data from different sources. This is why interoperability is a pillar of the FAIR[1] principles. An additional benefit of interoperability is that it reduces the mental overhead on researchers when moving from one infrastructure to another, thereby making their research more efficient and increasing science output.

For data providers and service operators, the consideration is more complex. While it is true that once a standard is established and solid server-side software support exists, their implementation is generally simpler than a new design from the ground up. On the other hand, establishing standards and writing portable, reusable software is a major investment over a custom solution. Where standards are not already well established or required by funding agencies, the Principal Investigators (PIs) of the projects that produce that data – it is these that service providers primarily talk to in the design phase – may actually prefer a custom solution, for instance because offering their data for easy consumption by third parties is not a priority for many PIs. Indeed, this step is often seen as an added cost unless required by the community.

In the present case, the situation of the service operators is even more complicated, in that their "customers", the researchers, will, if at all, ask for "vertical" integration, i.e., interoperability with the standards established in their field, be they

[1] cf. Wilkinson et al, 2016, "The FAIR Guiding Principles for scientific data management and stewardship", Scientific Data **3**, 160018.

formally agreed or informally set by widely used software. Since in both cases the standards are set within regional or global collaborations, they are both specific to disciplines while at the same time international in reach.

The Topic Group's mandate, along with the BMBF's goals, is horizontal federation, i.e., making infrastructures of different disciplines on the national level interoperable and then combining them into a seamless whole. This concept of a Federated Science Cloud, consisting of computing, storage, and archive resources of all eight ErUM communities in Germany, certainly appears attractive.

To establish that, a questionaire has been provided to the participating communities in which they were supposed to answer the following questions:

- What federated infrastructures are already existing which can be used as a basis for future implementations?
- What technologies and standards are employed?
- What issues and challenges are interconnected with federated infrastructures?
- What infrastructures are planned to be federated?
- What approaches and solutions do already exist for building community - spanning federated infrastructures?

The information provided shall be used as a common starting point for building a community-spanning Federated Science Cloud in Germany and to work out a common wish list which is of interest for the majority of the communities. Moreover, this document is supposed to help in the preparation for the upcoming Federated Infrastructure call, in particular with a view to harmonising the proposals responding to it.

The body of this paper is written by topic group members and is aiming for a – necessarily subjective – survey of the state of play in the various disciplines. This survey will then inform some preliminary conclusions in sect. 7.

The material here occasionally needs to be technical, and discipline-specific jargon can not always be avoided. We hope that the glossary in appendix A will help to establish a common background.

2 Astroparticle Physics

by John Bulava, DESY – Andreas Haungs KIT/KAT

2.1 Current Needs and Use of Federated Infrastructures

The data produced in astroparticle physics observations is stored and distributed in different ways. In its methodological approaches to analysis, astroparticle physics is very close to experimental particle and nuclear physics, but the results are often interpreted by astronomical approaches. Similarly, computing and its requirements are very diverse.

Compared to, e.g., high energy physics, astroparticle physics has wider set of diverse tool chains, data sources, and observatories. The major infrastructures in Germany currently contributing to Big Data and Big Computing requests are the Gamma-ray Astronomy Observatory CTA, the Cosmic Ray experiment Pierre Auger Observatory, the Neutrino Observatory IceCube, the future gravitational wave detector Einstein-Telescope, the future Dark Matter experiment DARWIN emerging from XENONnT, the neutrino mass experiment KATRIN, and the Double Beta Decay Experiment LEGEND. Information is obtained in particular by linking the data of different experiments (Multi-Messenger Astroparticle Physics), in which the digitalization of the research field plays an important role.

Each of these research infrastructures has its own computing concept, but often a co-use of the WLCG centers at the research centers and individual universities with particle physics involvement.

For some established observatories such as H.E.S.S. or the Pierre Auger Observatory, data is stored conventionally at a few sites and is accessed and transferred conventionally using standard protocols/tools such as ssh, ftp, and Globus. While flexible and easy to implement and extend, this paradigm does not efficiently accommodate growing numbers of users, and relies on each user having explicit access to the site in question. Other observatories, such as CTA (planned) and IceCube, make heavy use of the (US-based) OpenScienceGrid, as well as ad-hoc federated computing contributions from member institutions. OpenScienceGrid is based on GlideinWMS, which in turn uses HTCondor. Ad-hoc federation of smaller and heterogeneous computing clusters at member institutions is provided by https://github.com/WIPACrepo/pyglidein, which also uses the HTCondor glidein mechanism to provision slots to a central pool Software and other shared data resources are provisioned mainly through CVMFS. Data movement is handled by custom middleware using GridFTP as a transport.

These concepts have worked until now, as the demands on funded data centres have been negligible compared to the use by particle physics. In this framework the astroparticle physics community has gained experience in using Tier-1 and Tier-2/Tier-3 HTC centers as well as operation and first use of specific HPC centres (in particular for the existing gravitational wave detectors). However, this will change within the current decade, as new and large infrastructures will come into operation and astroparticle physics will also move into the area of exabyte data management.

2.2 Current Issues, Future Challenges

The federated model (as foreseen for most of the future research facilities in astroparticle physics) has many advantages, not least of which is an improvement in data transfer and computational throughput. Nonetheless, some users have difficulty navigating through a heterogeneous pool of resources and adapting their workflows to different environments. Debug cycles can be extremely long, as there is often no direct line of contact between the person responsible for the workflow and the remote site administrator. Making effective use of opportunistic computing resources requires active and engaged site administrators, as well as clear service-level agreements.

Another (minor) disadvantage with the grid model is the cumbersome user identification process to obtain grid certificates. For the CTAO, the particular grid implementation and access policy is as yet undecided, but four sites (of which DESY is one) have been designated as CTAO data centers.

Astroparticle physics and its scientists have to be integrated in the modern FAIR data life cycle where the federated infrastructures are an important pillar of an efficient computing concept.

Federated infrastructures are required to consist of dedicated CPU and storage systems at large computing centers, but also analysis resources as well as access and available resources at HPC centers. This includes also commercial cloud systems and extended GPU clusters. In addition, for an efficient and robust combination of data from different observatories (multi-messenger approach), the operation of specific software adapted to the needs of astroparticle physics and further development of dedicated software tools is needed. Here, too, sustainable synergies with other communities – especially with particle physics – can be achieved. However, this requires the willingness and support of all sides to adapt the already existing solutions to the specific requirements of astroparticle physics.

2.3 Possible Solutions, Future Challenges

The various challenges described could be countered by the concept of a variable data lake.

After a more detailed needs and gap analysis, a next step could be the development of a data lake (eventually based on XRootD) where a real-time monitoring com-

ponent needs to be included. Extended data storage systems with a flexible access and authorisation system used in the data lake, as well as corresponding protocols, need to be developed. All existing data sources, including data from astroparticle physics observatories, must be efficiently integrated into the data lake via open protocols. It must also be possible to easily integrate existing computing systems into the built-up data lake prototypes, including the globally distributed WLCG computing system. Therefore, techniques will be developed to integrate all existing computing resources, both dedicated experimental resources and opportunistic resources, HPC and cloud systems as well as dedicated analysis centres, to the data lake in a performant way. Users from all communities must be able to easily access the relevant data in the data lake.

Whereas the astroparticle physics data lake can be integrated in a large-scale federated infrastructure, astroparticle physics needs in addition a dedicated infrastructure, where a cross-observatory analysis and data center is to be set up. Structurally, this would be conceivable as a dedicated extension of a Tier-1 center with a hardware requirement of approximately 1 to 2 MEUR per year for the astroparticle physics in Germany.

3 Elementary Particle Physics Hadron and Nuclear Physics

by Alexander Schmidt, RWTH Aachen
Kilian Schwarz, DESY

3.1 Existing Federated Infrastructures

Over the last decades the field of experimental high-energy particle physics as well as the large experiments of the hadron and nuclear physics community ALICE at CERN and the *FAIR* experiments at GSI in Darmstadt have experienced a dramatic inflation in the data volume, in the required resources to process them, and in the complexity of managing the processed data. Both the ALICE experiment during LHC Run 3 as well as the CBM and PANDA experiments at FAIR will take data in the order of magnitude of TB/s. Real-time event reconstruction and online event selection and to some extend also higher-level processing of the stored data will be done via large compute facilities on site, e.g. the Green IT Cube at GSI. The high-energy particle physics experiments at the CERN Large Hadron Collider (LHC), especially ATLAS and CMS, will even reach the Exabyte region during the high luminosity phase (HL-LHC). These experiments at CERN have played a pioneering role in the development of the Worldwide LHC Computing Grid (WLCG). This project has been founded to handle and process the data produced by the LHC experiments. Also other HEP collaborations such as Belle II (KEK, Japan) are closely associated to WLCG and make use of the WLCG infrastructure and services. Still today the WLCG is the world's largest computing grid. It consists of around 170 computing centers in more than 40 countries. It is supported by many associated national and international grids across the world, such as European Grid Infrastructure (EGI) and Open Science Grid (OSG), based in the US, as well as many other regional grids, which are transparently federated in the WLCG. The WLCG consists of four layers, or "tiers"; 0, 1, 2 and 3. Each tier provides a specific set of services. The tier-0 is at CERN and is responsible for the safe-keeping of the raw data (first copy). Thirteen tier-1 centers are responsible for a proportional share of raw and reconstructed data, large-scale reprocessing and safe-keeping (custodial storage) of corresponding output, distribution of data to tier-2s, although the roles have evolved to be less strictly distinct in recent computing models. The tier-2s are typically computing centres at universities and other scientific institutes, which can store sufficient data and provide adequate computing power for specific analysis tasks. They handle analysis requirements and proportional share of simulated event production and reconstruction.

3.2 Technologies Employed

Currently the WLCG defines four component layers, networking, hardware, middleware and physics analysis software. The most important middleware stacks used in the WLCG have been developed within the context of various European Middleware Initiatives (e.g. EGEE, EDG, ...) having provided Grid middleware such as ARC, gLite, dCache, the Globus Toolkit (developed by the Globus Alliance), and the Virtual Data Toolkit. Nowadays these software stacks have developed further and the large LHC experiments have implemented Virtual Organisation-specific software stacks on top. For data, job, and workflow management the ALICE experiment is using the Grid middleware AliEn on top. For job management ATLAS uses PanDA, while CMS uses Global Pool and CRAB. LHCb has developed DIRAC which is an open software which has been adopted by many communities outside of LHCb, e.g. Belle II. As general technique used by all LHC experiments the job agent or glidein architecture has been established in order to pull the jobs from the central workload management systems to the local batch systems of the participating centres.

A common storage middleware at LHC is dCache which is deployed by the majority of tier-1 centres and more than 60 tier-2 centres worldwide and is developed within a collaboration with DESY being the lead centre. Additionally for data storage and data transfer XRootD is deployed. XRootD is used by the particle physics experiments at CERN as well as to some extent by GSI/FAIR experiments. The data management software used by many experiments is the Rucio framework which was originally developed by ATLAS.

Analysis frameworks of the large high energy physics experiments as well as many smaller nuclear physics experiments are traditionally based on the ROOT framework with the main developing centre being CERN. ROOT provides also the file structure in which the data are being stored.

Many experiments also provide analysis frameworks based on Python. A community - driven project with the aim of providing Particle Physics at large with a corresponding ecosystem for data analysis is Scikit-HEP.

3.3 Current Issues, future challenges

Currently, the computing resources are covered by computing centers provided by the Helmholtz association as well as the Max-Planck society and resources at various universities.

In the realm of LHC computing, the data volume and complexity will increase dramatically with the upcoming high-luminosity phase of the collider. A substantial increase of computing resources will be required. Even taking more efficient software and new technologies into account, the risk of a resource gap must be mitigated.

A remarkable increase of data volume and complexity is already starting now, with LHC Run 3. While the high-luminosity phase for LHC starts in 2029, the ALICE and LHCb experiments will be run at significantly increased luminosities already during Run 3, starting in 2022.

These challenges will require fundamental adjustments to the LHC computing models which will have a strong impact on the German contributions.

Also Belle II and the experiments at the GSI FAIR facility, as well as new large scale projects in astro-particle physics will have similar requirements.

These additional computing requirements come along with increased costs with respect to acquiring resources and operating them. When considering the current energy crisis on the market this means that these costs are actually increasing dramatically. In order to counteract that energy efficient methodologies need to be developed and applied. Data centres have to become energy efficient with the Green IT Cube at GSI being a positive example. Also research with respect to AI technology in data management is required as well as timely application. Using such techniques workloads, maintenance issues, the necessity to use human resources can be minimised, which means that in the end the operational costs can be significantly reduced.

One challenge the particle physics experiments and the large hadron and nuclear physics experiments have in common is to do data management and data analysis following the *FAIR* principles. Open data, open science, and community overarching data analysis is a central challenge which needs to be addressed. Next to data archiving and preservation one challenge at HEP is also the necessity to be able to reanalyse legacy data with the consistent software and calibration constants.

Another central challenge, this time specifically for the hadron and nuclear physics community, is the support of the many small experiments and communities. They are not as far advanced as the larger experiments in using federated computing and storage infrastructures, neither in applying *FAIR principles* and nor in corresponding software development.

How to meet these challenges is described in section 3.5.

3.4 Infrastructures to be Federated

The resources to be used consist of installations with dedicated CPU and storage systems at large computing centers, but also analysis farms at smaller institutes as well as only temporarily available resources at partner institutions and HPC centers (including supercomputing resources of the Gauss-alliance). But also commercial cloud systems and cloud access at universities and research facilities need to be included. This means that the resources to be used are characterised by a great heterogeneity. The use of GPU clusters is an attractive option that is investigated as well. In order to make efficient use of accelerators as GPUs the experiment code needs to be adapted. This is an ongoing effort and it depends on the experiment to what extent code can be ported to GPUs and what speedup can be achieved. For the robust operation of experiment specific software at such a variety of heterogeneous resources a high level of abstraction of work flows is required and the development of dedicated software tools is needed. This is necessary in order to allow the use of a multitude of resources spread across multiple locations without the special expertise of the local user community. The use of more federated and commonly used resources will also allow to use them in a more flexible way to increase the utilisation and occupancy of resources.

3.5 First approaches for solutions, future technologies

To meet these challenges, new technologies as well as a new distribution of the resource provisioning will be needed.

Currently it is being discussed that the mass data storage could be provided by a small number of sites in the context of so-called "data-lake" models. Such a concentration of the storage would allow a cost reduction for both personnel and resource provisioning. This would foresee that DESY, GSI and KIT would provide the majority of the WLCG mass storage.

In 2021 the NHR alliance was founded, a cooperation of German HPC centers to support science at universities. The KET community devised a plan to exploit those NHR resources for HEP computing and to gradually migrate the resources at the university tier-2 centres to the NHR sites for compute and the HGF sites for storage until the start of HL-LHC in 2029.

This will also require adaptations for the operational models of the NHR, because the WLCG requires fixed pledged resources and is not able to accommodate the current project based resource allocation models. This transformation will require further discussions at the political level.

New technologies for the access and use of existing and future computing resources will be needed. The new technologies evolve around the concept of a "science cloud" with the approach to implement complete infrastructures as services (IaaS, Infrastructure as a Service). This way, not only individual applications, but also complex infrastructures consisting of a multitude of services can be provided virtually.

One central point is to optimise the computing models and infrastructures in place. The central data cloud is to be accessible in a transparent and efficient way by heterogenous compute infrastructures consisting of classic pledged Grid resources, but also opportunistic and heterogeneous resources as HPC centres and Cloud infrastructures, which is where the virtualisation techniques mentioned above need to come into place. In addition dynamic disk caches need to be provided in order to provide efficient data access from opportunistic resources. Also efficient resource usage would increase by providing central and larger compute and storage facilities which can be jointly used by several scientific communities. This is also facilitated by the fact that due to the possibilities provided by modern bandwidths and access protocols the tasks being applied to individual tier centres in WLCG become more complex and flexible.

Beyond that one also needs to start thinking about paradigm shifts in order to be able to reach the targeted goals: one needs to be able to deal with the higher complexity of requirements and infrastructures, and one needs to deal with ecological and economical constraints that are now becoming a real issue. The real challenge lies in finding an optimum with respect to the measures applied and the achievements one can get out of them.

One way ALICE and FAIR are addressing the challenges provided by the necessity to process huge amounts of data in short time is to reconstruct a significant amount of these data already at the online compute facilities O2 and the Green IT Cube on site, also in order to do data reduction in real time. In addition to that ALICE is providing dedicated Analysis Facilities to the community which are optimised for high speed data processing, such as the GSI Analysis facility.

Since 2017 the German ErUM community (Erforschung von Universum und Materie) has started to work on common solutions to the challenges listed above. The pilot project IDT-UM (2017-2021) is being continued with the FIDIUM project (2021-2024), in which the intelligent techniques for abstraction, data management and accounting will be developed.

IDT-UM started developing overlay batch systems via which dynamic science clouds can be brought into existence. In addition to that dynamic disk caching systems have been developed so that a first step for efficient data analysis via dynamic science clouds has been done. These works are to be continued in the FIDIUM project. Central tasks here are ongoing development of the overlay batch system including data aware scheduling, to provide compute sites in a box with minimum administrative overhead as well as accounting and monitoring tools. A second field of development are real time monitoring systems for data lakes, efficient integration of dynamic disk caches and usage of parallel ad hoc file systems as caches in HPC centres. This will be complemented by data replication and placement methods for data lakes and usage driven data management and access methods. Based on these developments prototype data lakes will be provided. Another central point is tests, documentation and generalisation of production and analysis environments.

For data archiving and preservation there are solutions in place as data management software as, e.g., Rucio, which triggers outsourcing cold data to tape. For FAIR data analysis there are already experiment-specific solutions. On top of that NFDI consortia as PUNCH4NFDI are developing standardised solutions. These need to be well coordinated with activities in the European Open Science Cloud (EOSC) and related EU projects such as ESCAPE.

4 Neutron Research

by Tobias Richter, Simon Heybrock, Kareem Galal, European Spallation Source ERIC

4.1 State of federated infrastructures

Overall the German neutron landscape consists of about 1500 researchers (often in university groups) and a small number of national, European and international facilities. Federation happens on a national and European facility level. External services (for example from EOSC) play a small role.

For the users of neutrons there are a few examples of federated infrastructures in a operational state. Facilities in the PaNOSC and ExPaNDS EOSC projects have started to open their data catalogues to harvesting by B2Find and OpenAire via OAI-PMH. That means that information on experimental data older than the facility-specific embargo period (typically around 4 years) is findable in federated cross-discipline metadata repositories. In addition the two EOSC projects develop a federated domain specific search API and portal. Another example of federation is the UmbrellaID AAI infrastructure that enables user authentication with the same credentials in a number of participating computing services.

Currently data processing capacity and storage is hosted by facilities in isolation. Data is often transferred away to local university infrastructures by the researchers and analyzed and compare with data from other experiments there.

4.2 Interest in federated services

Users of neutron facilities would benefit from:

- catalogue services that combine metadata from experiments carried out at different facilities and in their own labs in a consistent view
- data sharing services with authentication to collaborate on dataset with local and external collaborators without having to move the data
- data processing and analysis capabilities that work transparently on data irrespective of the location
- being able to fulfil their data preservation requirements by using federated long term archiving

Availability of a scalable long term archiving service could also be welcome as a federated offering by the neutron research facilities. The same is true for compute capabilities for offline analysis and simulations.

5 Synchrotron Radiation

by Anton Barty, DESY

5.1 State of federated infrastructures

The photon science large scale facilities generate significant raw data volumes and currently maintain largely separate computing infrastructure located on-site at each of the facilities. This is a result of the funding model whereby each facility is separately funded to support the computing needs of the facility users.

This development is largely historical. Computing services have grown locally as detector demands increased. IT services are linked to local data acquisition where each facility is the "Tier 0" centre for data, and increasingly computing on that data. Computing has been seen as a part of the local facility infrastructure rather than a federated service across infrastructures. Each facility has its own hardware, login and job scheduling systems. Resourcing is driven by separate funding streams for each of the facilities.

At the same time, DESY and the European XFEL host their hardware in the same data centre at DESY and have some ability to share resources. The compute and storage systems are financed separately and are logically separated in the data centre even though they share AAI infrastructure, file transfer services, tape archiving systems and the like. Both run on GPFS for high speed storage and use dCache to manage tape archiving, use the same common login authentication and remote access gateways, and hence can locate compute nodes within the shared Maxwell cluster. Opportunistic use of free CPU resources is possible as data from one facility can be processed on nodes owned by the other facility due to shared AAI and GPFS infrastructure; however, each facility gets priority over its own resources when needed. The arrangement is better described as symbiotic co-existence than a true federation of resources. The resources provided to the user community at their home institutions (universities, MPG, industry) are definitely not federated

5.2 Already federated computing and storage infrastructures

Broadly speaking, the large scale facilities do not use federated computing and storage infrastructures. The reason is that each facility is funded to provide storage and computing for the facility users, not the users of other facilities – and since there are always fewer resources than needed there are no spare resources to share. A secondary reason is that high data volumes mean that significant computing is needed on site anyway for stable data acquisition and prompt processing, necessitating some form of on-site data centre regardless of federation.

One example of shared infrastructure that does work is the Maxwell cluster at DESY. Maxwell is a significant compute and storage resource located at DESY which serves the needs of both DESY photon science (FLASH/PETRA-III) and European XFEL, as well a heterogeneous collection of other researchers at DESY including the CryoEM user facility CSSB at DESY. This facility is centrally located with each institution contributing resources for their respective user community. Due to a common login and portal, shared use of resources is possible as data and servers from all institutions share a common entry point, AAI, disk space, tape archiving, etc. In practice, each facility has priority use over its own resources and can make opportunistic use of resources from the other facilities when available. Software and other infrastructure is shared, and since the cluster consists of heterogeneous resources it is possible to make use of specific configurations if needed (e.g., large memory machines, multiple GPU machines, many-core machines, especially for testing). This model seems to work well largely because the facilities involved share the common computer centre at DESY.

There have been several initiatives overseas to make use of shared infrastructure for the photon science community, notable among which are the DOE facilities in the US, the SLS in Switzerland, and Max-IV in Sweden. Here, the facilities were encouraged to use national supercomputing facilities for their offline and long-term computing needs, with mixed success.

In the US, for example, there was an initiative and pilot project for the DOE funded light sources to use the DOE-funded supercomputing facility at NERSC to process their data. This ran into the following problems:

- The need for immediate experiment feedback necessitated significant on-site computing resources in any case.
- Large amounts of data needed to be transported over large distances with high reliability. One instrument at a light source can produce 1 PB of data a day, and there may be 30 such instruments at one facility. Data transport must be prompt and reliable.
- The access model for supercomputing facilities did not match the needs or expectations of photon science users (long job turnaround times of up to a week versus the expectation of short turnaround times for data optimisation and even interactive operation).
- The supercomputers themselves were optimised for calculation and not processing large volumes of data with high data throughput rates
- Using a supercomputer highly optimised for parallel computation to perform multiple small single-threaded jobs with little code optimisation from photon science users was realised to be a poor use of the supercomputing resources.
- The knowledge barrier to entry into a supercomputing facility environment was largely above all but the most specialist research groups. Simply put, the facility was designed to serve highly computer literate users, and not the heterogeneous set of non computer specialists in the photon science community. It was simply too complex to use for most people, and too complex for their needs.

The conclusion here is that such use of a central supercomputing facility such as already exists in many national infrastructures may be technically possible, but it is not necessarily well suited to the needs of the photon science community, nor is it necessarily a good match or good use of such a resource.

5.3 Current issues which need to be addressed

By and large, the Maxwell cluster at DESY works relatively well for experiments on site at DESY. One could consider using an expanded version of such a resource as federated offline computing for all photon and neutron facilities in Germany. This would have the benefit of common infrastructure, common environment and software stack, and a 'single point of entry' for all national facility users. It helps that most of the photon and neutron facilities in Germany fall under the same funding umbrella. If designed to serve the photon and neutron community it may avoid the trap of trying to do absolutely everything for all communities.

5.4 Computing and storage infrastructures we would like to federate

The photon and neutron communities are working together within the NFDI under the banner of Daphne4NFDI. We therefore share several common views and goals on shared and federated infrastructure, including:

1. AAI – Allowing employees from other institutes and from the PaN community access in a transparent manner to photon science compute resources would be highly advantageous. At the moment, each user requiring access to data or resources has to be granted an authenticated account. This adds significant administrative overhead which could be avoided though data and compute access through authenticated AAI from the home institute. This was supposedly addressed as part of the UmbrellaID PaNData initiative. However, UmbrellaID has seen limited adoption outside of the proposal and user office systems, and is not used for authentication to data or compute infrastructure.

2. Storage – Federating storage backends as done with the Maxwell cluster could save resources, particularly for long term archiving. The requirements to preserve data are becoming onerous in terms of cost (due to the large data volumes involved) and a common solution could be highly advantageous.

3. Federated compute resources – The Maxwell cluster at DESY has been rather successful and could be applied more widely to the PaN community, provided the issue of near-real-time WAN data transfer is addressed, and an appropriate funding model devised. Note that there would still be a need for onsite computing for real time feedback at each facility.

4. Standardised environments and software stacks would help users moving from facility to facility.

5. Data portal (scicat) – The Scicat data portal allows users to search and access data from current and previous experiments. This should include electronic user log books and metadata capture. Using the data portal ensures data is accessible according to the FAIR principles. Federating this tool will ensure data is searchable and accessible across multiple research institutes, in a coordinated and simplified manner. We are currently working with the NFDI to try and harmonise this.

6 Astronomy and Astrophysics

by Markus Demleitner,
Univ. Heidelberg

6.1 Existing Federated Infrastructures

The Virtual Observatory (VO) is an international effort to run and develop a federated data infrastructure in Astronomy that is held together by a set of data and protocol standards, continually developed since the early 2000s. Consisting of a Registry, some 25000 interoperable services (which still roughly match data collections) comprising hundreds of millions of datasets (spectra, images, and the like) and hundreds of billions of table rows, and a set of clients and libraries consuming these services, it is widely used in the astronomical community.

One central assumption behind the VO is that users should not normally interact with data services – machines should do that on their behalf. This is to facilitate using many different services at the same time ("global discovery"; here, a client can easily query hundreds of services, where no human would fill out hundreds of web forms) or consecutively (avoiding "lock-in" to specific services or resources; standard services save exploration time).

A *VO service* thus is a network-accessible endpoint defined by our *standards*, giving access to one or more data collections. To illustrate the VO's approach, here is a hypothetical session:

Suppose a user requests "images of Barnard's star in x-rays". This is how this proceeds in the VO:

1. A client asks a searchable registry: Give me resources that
   - serve images,
   - have data in the x-ray part of the spectrum, and
   - have data around $\alpha = 269.45$, $\delta = 4.693$ (i.e., the current location of Barnard's star in the sky).
2. The Registry responds with metadata for the services matching these criteria.
3. The client now goes to each service returned and asks it for data
   - covering the position $\alpha = 269.45$, $\delta = 4.693$ and
   - intersecting the spectral range $0.1 \cdots 120$ keV of photon energy.
4. Each server responds with one set of metadata per matched image. The client turns this into some representation for the user.
5. The user picks images based on the metadata (e.g., observation date, sensitivity…).
6. The client retrieves image (or parts of them) and makes them available for further processing.

Here is a brief overview of the standards that have been developed in the context of the VO effort:

Searching for data: Images (SIAP), spectra (SSAP), objects (SCS), spectral lines (SLAP), generic datasets (ObsCore).

Remote manipulation: SODA, specifying cutouts, rescaling, etc., to avoid pulling data not relevant to the user.

Interacting with databases: Access using TAP, common query language ADQL.

Formats: Table exchange using VOTable, complex spherical geometries with MOC, multiscale images with HiPS.

Registry: Registry Interfaces for the architecture, VOResource, VODataService, TAPRegExt, SimpleDALRegExt for the metadata format, RegTAP for how to search it.

Semantics: Light semantics of physical quantities (UCD), Unit syntax, vocabulary maintenance, $\sim$ 15 vocabularies.

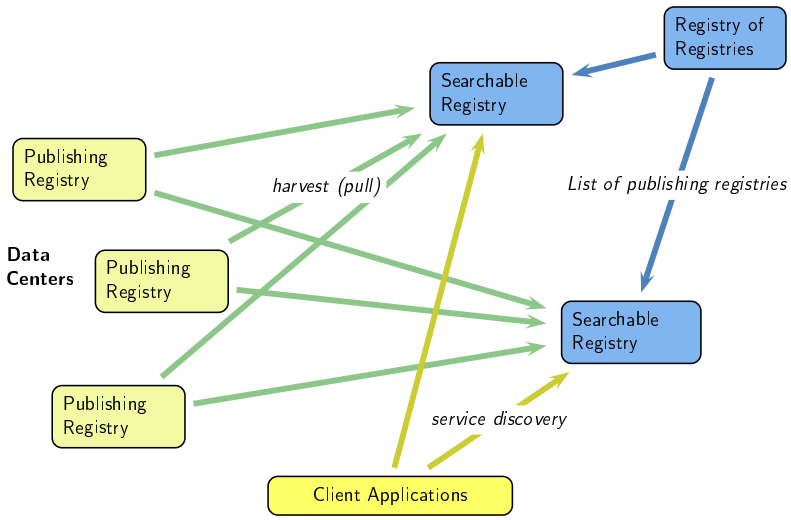

Figure 1: The architecture of the VO Registry: Service discovery happens using *searchable registries*. While anyone can run one, in practice there are three major operators (GAVO, ESA, STScI) that clients can go to as they choose. These searchable registries get their data by OAI-PMH harvesting *publishing registries* operated by the data providers. They know where to go because all (currently about 45) publishing registries are listed in the one central infrastructure of the VO, the *RofR* (Registry of Registries).

Other: SAMP for assembling complex environments from simple building blocks, VOSpace as an object store, VOEvent for disseminating information in transient phenomena in the sky.

The full list of IVOA standards is available from the IVOA document repository[2]. The IVOA also takes up a large number of third-party standards, in particular many IETF and W3C standards including HTTP, RDF, and XML, as well as FITS and OAI-PMH, or the Unified Astronomy Thesaurus maintained by journal publishers, librarians and learned societies.

A cornerstone of the VO is the Registry; Fig. 1 illustrates its architecture, designed both to avoid operation-critical central components and in the VO spirit of letting, in principle, anyone offer any sort of service. The only part in the VO Registry that is actually a central singleton is the Registry of Registries, and that can, if necessary, be unavailable for days without noticable impact on VO users.

6.2 Technologies Employed

Central to most of what the VO does are relational databases – most projects run Postgres, but many other RDBMSes are in use, too. This is partly because a relevant part of our science data ("catalogues") is relational in nature, partly because our discovery interfaces for datasets are written in terms of relational representations.

Some rather basic structures were developed within the discipline; a good example is MOC and the associated HiPS, techniques for the fast representation of rather arbitrary regions on spheres and information linked to such regions (think research-quality Google Earth).

Our Registry's metadata scheme is significantly more elaborate than DataCite's, in particularly adding metadata on tables and services. Mapping VOResource to DataCite is rather straightforward but loses almost all the metadata that actually makes the VO work.

Really central to the VO are end-user components. These can roughly be grouped into Browser-based ones (e.g., ESA sky for general browsing, WIRR for registry discovery), Applications (which, for deployment convenience, are mostly written Java;

[2] http://ivoa.net/documents

popular examples include TOPCAT and Aladin), and libraries for VO access (e.g., pyVO and astroquery from Python, which by now are the most popular programmatic interfaces among end users, STIL from Java). In particular the early VO had a tendency to forget about these client parts, and whenever it did that, things quickly went awry.

For worked-out use cases for VO operations, see VO Text Treasures[3].

6.3 Current Issues

One major challenge is how to deal with data too large to conveniently move. While with the SQL derivative ADQL and the query-transporting Table Access Protocol TAP we have a highly successful model for how to bring expressions from relational algebra to the data, similarly clear and interoperable techniques for array-like data, let alone collections of arrays (e.g., images, dynamic spectra, complex time series) have yet to be developed; that these techniques will probably involve Turing-complete formalisms that are hard to reason about even for machines exarcerbates the situation.

Right now, observatories facing the problem of data that is hard to move tend to offer services based on containers or ipython backends, but of course these are neither interoperable (in the sense that a, say, ipython notebook could be easily transferred from one service to another) nor discoverable, and at least compared with ADQL queries (that will still work even if they were written 10 years ago) they will break rather quickly due to evolving dependencies.

One thing we are still struggling with was the early design of identifying data collections and services, where, say, a collection of spectra was registered as a service for searching spectra. This was initially convenient but became increasingly cumbersome as services started to publish multiple data collections – there are several services in today's VO offering more than a thousand data collections each behind a single access URL – and data collections became available through multiple interfaces – it is rather common today for datasets to be published through both a typed interface (SIAP, SSAP) and through a TAP-published table containing standard metadata for observational data (obscore). Rectifying this early misdesign while widely deployed clients will break on straightforward fixes has proven to be a major challenge.

What has not really worked in the VO so far is data modelling. While data models have been proposed for both concrete data products (e.g., Spectra) and physical concepts (e.g., positions, photometry), takeup for whatever prescriptions were derived from these has been low, and clients essentially do not consume them. On the other hand, there is a clear need for having more complex data structures described interoperably. There is an effort underway to rectify the situation by a stronger formalisation of modelling work using an interoperable subset of UML ("VO-DML") and standard serialisation of instances. On the other hand, given the predominance of relational databases in the VO, direct modelling in relational terms (e.g., ObsCore, RegTAP, EPN-TAP, Datalink) has been rather successful.

For Germany, a central problem with respect to the VO is that there is essentially no institutional footing. Where France has the CDS, Canada the CADC, the US at least MAST and IRSA, the UK the WFAU, and Italy INAF's Trieste data centre, the German contribution to the VO has so far relied mostly on BMBF and EU project funding, which is both uncertain and strongly fluctuating. An institutional basis would make our contribution significantly more sustainable and credible.

6.4 Infrastructures to be Federated

As said above, compute platforms next to immovable data should be made interoperable and discoverable, but that clearly is a hard problem without simple solutions. Perhaps going some smaller steps initially, e.g., using techniques like ArraySQL, will help to find paths towards more general and simultaneously viable solutions.

[3] https://dc.g-vo.org/VOTT

A constant problem since the early days of the VO has been authentication. It has not been particularly pressing so far, as open data is rather common in astronomy, but with advanced services like persistent uploads on TAP services and computing platforms, the lack of interoperable, widely implemented, federated authentication is becoming a hindrance. What standards have been written in this field (SSO, Credential Delegation), are either not sufficiently constraining to enable implementations against them or are not implemented widely enough.

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 can deal with raw storage, far fewer with, say, FITS files, and still fewer with XYFITS from radio interferometry). Moreover, compared to raw data fewer people will *want* to read astronomical tables, still fewer will have reason to analyse visibilities.

Hence, fortunately what is most easy to federate also promises the highest return, and that is probably what we should start with.

While good motivations for building federated infrastructures are using synergy, potentially reducing cost, making it easier to exploit a wider range of different resources, facilitating data sharing, optimising resource usage by increasing the number of potential users and also the diversity of use cases, and avoiding lock-in to specific providers, it is also required that the communities need to develop a common understanding of:

- Which infrastructures are already operated in the respective communites? Which of them are already part of international federations and bring along corresponding boundary conditions?
- Which technologies are being used in federation and service provision? This is particularly interesting because compatible fundamental technologies might lead the way to low-cost solutions.
- What are the experiences with these infrastructures? Are there lessons learned?
- Which further infrastructures should be made interoperable and/or federated?
- Are there criteria to decide when horizontal integration is beneficial (e.g., common requirements or technologies) and when the problems (e.g., necessarily different metadata schemes, entirely disjunct workflows) outweigh possible benefits?
- Do we want interoperable resource/data collection discovery across disciplines? And if so, why and how? See `https://github.com/msdemlei/cross-discipline-discovery` for drafts of user stories, also PUNCH overarching use cases:

https://www.punch4nfdi.de/use_cases/use_case_class_4____user_story/,
https://www.punch4nfdi.de/use_cases/use_case_class_4/

- What is the funding model which can be applied for infrastructures spanning different communities? Science communities are typically funded to provide resources for their own research communities, while a federation of resources requires allowing others to access the same resources in competition with the own research community.
  In addition to that the complex funding structure (regional, state, federal, European levels) requries a particularly close interaction between the players both on the side of the funders and on the side of science.

As a concluding statement it needs to be stressed here that reliable, and long-term funding for data centres is a precondition for useful and sustainable federation.

A Glossary

CADC Canadian Astronomy Data Centre, a Victoria, BC-based institution managing data publication for the Canadian astronomy community (and a good deal beyond that) that GAVO would consider a model for how such a thing should be organised.

CDS Centre de Données astronomique de Strasbourg, a French data centre that keeps and curates must astronomical data published in tabular form.

CTA Cherenkov Telescope Array.

CVMFS CERN Virtual Machine File System, a technology to distribute a central software installation globally via a squid caches.

DataCite A technology for minting persistent identifiers for data artefacts that comes with a minimal and cross-disciplinary metadata schema for describing data resources and relationships between them.

EGI European Grid Initiative.

ESA European Space Agency.

GAVO German Astrophysical Virtual Observatory, the German contribution to the global VO effort.

GPFS General Purpose File System, an IBM-developed cluster file system.

IETF Internet Engineering Task Force, the body that manages the development and evolution of the internet's core standards.

IVOA International Virtual Observatory Alliance, the body that manages the development and evolution of standards in the VO.

LHC Large Hadron Collider, a 20 km long circular particle collider operated at CERN.

MAST see STScI.

OAI-PMH Open Archives Initiative Protocol for Metadata Harvesting, a widely-deployed mechanism to exchange resource metadata allowing incremental harvesting.

OSG Open Science Grid, a Grid initiative of US resource providers. OSG provides a middleware distribution that is also used elsewhere, primarily in America.

PI Principal Investigator; used here to designate the researchers producing data and/or consuming services, typcially within a specific project of limited duration.

RDBMS Relational Database Management System, a class of software allowing efficient querying and manipulation of tabular data.

RDF Resource Description Framework, a W3C standard to describe semantic resources like vocabularies and ontologies.

SQL Structured Query Language, the de-facto standard for writing queries against RDBMSes.

SSO Single Sign On, a group of technologies that usually features a single service authenticating users and granting them tokens giving access to many other access-controlled resources.

STScI Space Telescope Science Institute, a Baltimore, MD-based facility publishing data from several NASA instruments, e.g., through its MAST archive.

TAP Table Access Protocol, a VO standard letting users execute queries in a SQL-like standard language on remote servers (cf. sect. 6).

UML Unified Modeling Language, a set of technologies and standards designed to facilitate reasoning about software and its development.

VO Depending on context, this could be Virtual Observatory (cf. sect. 6) or Virtual Organisation, a concept in the Grid's authorisation infrastructure.

W3C World Wide Web Consortium, the body that manages the evolution of Web standards like HTML, XML, CSS, RDF, and so on.

WLCG Word Wide LHC Computing Grid, a collaboration of resource providers to support the resource needs of the LHC experiments.