Journal of Internet Services and Applications
a SpringerOpen Journal

RESEARCH Open Access

A data infrastructure reference model with applications: towards realization of a ScienceTube vision with a data replication service

Morris Riedel[1]*, Peter Wittenburg[2], Johannes Reetz[3], Mark van de Sanden[4], Jedrzej Rybicki[1], Benedikt von St. Vieth[1], Giuseppe Fiameni[5], Giacomo Mariani[5], Alberto Michelini[6], Claudio Cacciari[5], Willem Elbers[2], Daan Broeder[2], Robert Verkerk[4], Elena Erastova[3], Michael Lautenschlaeger[7], Reinhard Budig[7], Hannes Thielmann[7], Peter Coveney[8], Stefan Zasada[8], Ali Haidar[8], Otto Buechner[1], Cristina Manzano[1], Shiraz Memon[1], Shahbaz Memon[1], Heikki Helin[9], Jari Suhonen[9], Damien Lecarpentier[9], Kimmo Koski[9] and Thomas Lippert[1]

Abstract

The wide variety of scientific user communities work with data since many years and thus have already a wide variety of data infrastructures in production today. The aim of this paper is thus not to create one new general data architecture that would fail to be adopted by each and any individual user community. Instead this contribution aims to design a reference model with abstract entities that is able to federate existing concrete infrastructures under one umbrella. A reference model is an abstract framework for understanding significant entities and relationships between them and thus helps to understand existing data infrastructures when comparing them in terms of functionality, services, and boundary conditions. A derived architecture from such a reference model then can be used to create a federated architecture that builds on the existing infrastructures that could align to a major common vision. This common vision is named as 'ScienceTube' as part of this contribution that determines the high-level goal that the reference model aims to support. This paper will describe how a well-focused use case around data replication and its related activities in the EUDAT project aim to provide a first step towards this vision. Concrete stakeholder requirements arising from scientific end users such as those of the European Strategy Forum on Research Infrastructure (ESFRI) projects underpin this contribution with clear evidence that the EUDAT activities are bottom-up thus providing real solutions towards the so often only described 'high-level big data challenges'. The followed federated approach taking advantage of community and data centers (with large computational resources) further describes how data replication services enable data-intensive computing of terabytes or even petabytes of data emerging from ESFRI projects.

Keywords: Reference model, ScienceTube, Data infrastructure, Replication

*Correspondence: m.riedel@fz-juelich.de
[1] Juelich Supercomputing Centre, Juelich, Germany
Full list of author information is available at the end of the article

1 Introduction

'A fundamental characteristic of our age is the rising tide of data - global, diverse, valuable and complex. In the realm of science, this is both an opportunity and a challenge [1]. As this quote suggests there are many opportunities and challenges in the steadily increasing amount of scientific data and there are a wide variety of rather high level reports, press releases, or statements given to support this claim. The following other examples are taken from high-level recommendations by the e-Infrastructure Reflection Group (e-IRG)[a] report on data management: *'Encourage the development of non-discipline-specific frameworks and information architectures for interoperable exchange of data...support communities for the definition of their requirements...'* [2] and *'Ensure that besided hardware and services, digital objects deserve infrastructure components in their own right: ... persistent linkage of research data... policies for long-term preservation of data, maybe focused into dedicated centers...'* [2]. Another example from the European e-Infrastructure Forum (EEF)[b]:*'Data archiving and curation is a common need for several of the ESFRI projects'* [3].

But in many cases, less concrete information is given about concrete activities that link to these high-level reports and goals while at the same time have roots with bottom-up activities in order to ensure that solutions are provided that are really needed in science in the next decade. For example, the high level expert group on scientific data puts the following high-level vision towards 2030: *'Our vision is a scientific e-infrastructure that supports seamless access, use, re-use, and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure'* [1]. But how will such an infrastructure look like and what would be the first steps towards the implementation of such a vision? How it will work together with the wide variety of existing local, regional, national, or even pan-European data infrastructures? Would it be possible to compare at least roughly these different existing data infrastructures in order to explore synergies and what mechanism is provided for that? How can community centers with domain-specific expertise and data centers with large-scale processing expertise work together to perform data-intensive computing and perhaps even contribute to the avoidance of 'big data' by investigating, for example, data de-duplication approaches?

These are all questions for which the answer is not evident while we need to acknowledge that some fragmented answers exist to some of these questions. This paper provides more consistent answers to the aforementioned questions with bottom-up solutions driven by scientific end-user needs while at the same time being linked to the aforementioned high-level vision. This high-level vision where *'...data themselves become the infrastructure...'* [1] is broken down into more concrete functionality entities and their relationships that sum up to a *'data infrastructure reference model'* in order to enable comparison with the wide variety of existing solutions in the field as well as providing a more clearer picture of how high-level visions might be realized. The absence of such a more abstract scientific-oriented data infrastructure reference model is diametral to the fundamental design principles of software engineering and has thus lead to numerous different non-interoperable architectures as part of fragmented data infrastructures in the last decade.

In order to advance from the abstract to the more concrete, concrete scientific end-user requirements are analyzed arond a *'data replication service'* use case that drives the design of reference model- derived concrete architectures. This contribution will describe how these concrete architectures are implemented as part of the European Data infrastructure EUDAT project [4] and which data infrastructure integration issues we need to overcome. The approach thus followed in this contribution is *'putting the scientific end-user into the driving seat'* meaning that the architecture is guided by the high-level vision and more clearly defined by using a specific use case around a *'data replication service'* as a driver. Scientific communities that are interested in this use case include members of the European Network for Earth System Modelling (ENES)[c], the European Plate Observing System (EPOS)[d], the Common Language Resources and Technology Infrastructure (CLARIN)[e], and the Virtual Physiological Human (VPH)[f].

This paper is structured as follows. After the introduction, Section "Motivation" provides the motivation for our work and sets it in context to the broader view on scientific data challenges in the next couple of years. Section "Data infrastructure reference model" then introduces the data infrastructure reference model and the high-level goals driving its derived architectural activities. The concrete use case around a data replication service is introduced in Section "Data replication service use case", while Section "Architecture for data replication use case" describes how architectural core building blocks guided by the reference model are derived from this particular use case. After the survey of related work in Section "Related work", this paper ends with some concluding remarks.

2 Motivation

The introduction already provided several examples of high-level recommendations, plans, agreements that motivates a *'real deployment of solutions'* and *'actions from scientific stakeholders and their partners that can make a difference'* within Europe in order to cope with the rising tide of scientific data. This can not be done from scientific communities in isolation from large-scale data centers in Europe, nor are such data centers able to reach out into the scientific communities alone. Instead, the high level expert group on scientific data suggests to *'develop an international framework for a Collaborative Data Infrastructure (CDI)'* [1]. The CDI high-level vision in turn follows a *federated model* that takes advantage of the benefits of community and data centers.

The high level expert group on scientific data report further outlines that data-intensive computing is related to *scientific workflow executions*. Processing-intensive activities such as *data mining* that are considered as common services across scientific domains are required to be available at large data centers. These data-intensive computing activities in many cases requires processing powers that raise the demand for High Performance Computing (HPC) resources. On the other end of the scale, High Throughput Computing (HTC) resources can be also powerful when used with data analysis tools (e.g. MapReduce [5]). Although many community centers (i.e. middle layer) also provide computing power at their centers (e.g. HPC at DKRZ), the larger data centers in Europe complementary offer unique computing capabilities towards the peta-scale performance range (e.g. HPC at JSC). This is the reason why *compute resources* are modelled alongside the 'data replication service' since data is replicated to data centers that enable additional large-scale processing.

3 Data infrastructure reference model

This section aims to provide a *'frame of reference'* in order to systematically approach the high-level challenges listed in the previous section. But these challenges need to be better specified and understood in order to provide concrete solutions for them. A series of workshops [6,7] have been performed with scientific users that face data challenges in order gather requirements for a federated data infrastructure.

A reference model approach is used to understand the requirements in context to each other. Such a reference model is an abstract framework for understanding significant entities and relationships between them within a service-oriented environment [8] such as a service-oriented data infrastructure. The OASIS SOA reference model [8] is taken as the foundation of the work and is illustrated in Figure 1. This illustration shows that a reference model itself is not directly tied to standards, technologies, or other concrete implementation details. A wide variety of user communities lead to many existing concrete data architectures while at the same time these all share common problems that can be expressed with abstract entities on the reference model level. A reference model guides concrete derived architecture work such as the reference architecture that is typically based on standards, profiles, or specifications. Concrete derived architectures and their implementations are expected to vary from each other, but being described with a reference model enables a better comparison between them.

Figure 1 provides pieces of information how the elements in this contribution are connected with each other by indicating which sections contribute to which reference model elements. The grey parts in this figure have been added to the original from [8] in order to provide even more information how the contributions in this paper on different architectural levels contribute to the overall 'ScienceTube' vision and its derived reference model.

3.1 ScienceTube vision and user requirements

The outcome of the workshops [6,7] held all point to a vision we refer to as *'ScienceTube'* by which we not intend to give it a clear product name. Instead this 'ScienceTube' vision rather stands for a key goal with several aligned activities in this contribution and as being partly implemented in EUDAT.

The requirements expressed by user communities that shaped this vision all sum up to the fact that *'everything is inter-linked data accessible in a lightweight Web-like fashion'* as shown in Figure 2. The vision goes far beyond just 'videos' as data what the 'Tube' might indicate and instead rather connects scientific data and makes it easily accessible to scientific users. The core functionality of the illustrated GUI is a *'scientifc data viewer'* that is able to dynamically switch depending on which data is currently viewed (e.g. scientific datasets, paper, etc.). This viewer needs to understand the different data structures of the scientific data in question.

Figure 2 also illustrates that there are *'recommendation systems'* pointing to the 'most viewed data today' such as papers that are inter-linked with scientific measurement data obtained from a large device (e.g. telescope). Other requirements have been *'rating systems'* in order to encourage trust for users over a long period and to support the reputation of those that share their data. The latter is particularly crucial and addresses a clear challenge by encouraging user communities to share their scientific data with others. Complementary to the data itself, there is a lightweight access to data processing power and available storage resources. All these requirements raise the demand for a strong service backend with various technical functionalities.

3.2 Reference model blueprint and design

Figure 1 introduces the key elements of the reference model parts and this section briefly describes its major design foundations. Figure 3 provides an abstract blueprint via a conceptual view that incorporates key principles of reference models listed in [8]. The reference model is *'abstract'*, because its entities and concepts are an abstract representation of the entities (e.g. data-oriented services) that exist in production data infrastructures.

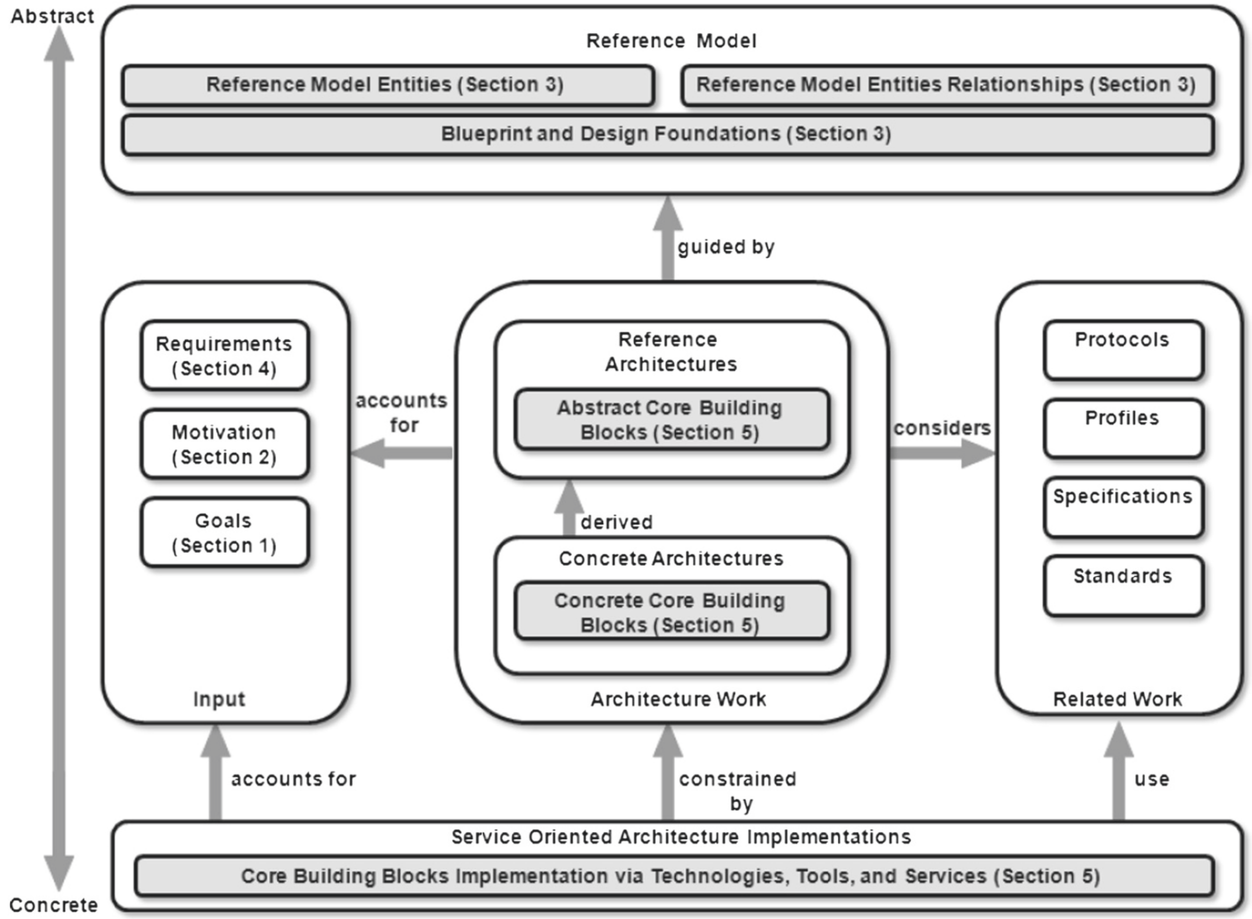

Abstract
Reference Model
Reference Model Entities (Section 3)
Reference Model Entities Relationships (Section 3)
Blueprint and Design Foundations (Section 3)
guided by
Requirements (Section 4)
Motivation (Section 2)
Goals (Section 1)
Input
accounts for
Reference Architectures
Abstract Core Building Blocks (Section 5)
derived
Concrete Architectures
Concrete Core Building Blocks (Section 5)
Architecture Work
considers
Protocols
Profiles
Specifications
Standards
Related Work
accounts for
constrained by
use
Service Oriented Architecture Implementations
Core Building Blocks Implementation via Technologies, Tools, and Services (Section 5)
Concrete

Figure 1 OASIS Reference Model for Service Oriented Architectures is used as foundation (modified from [8]).

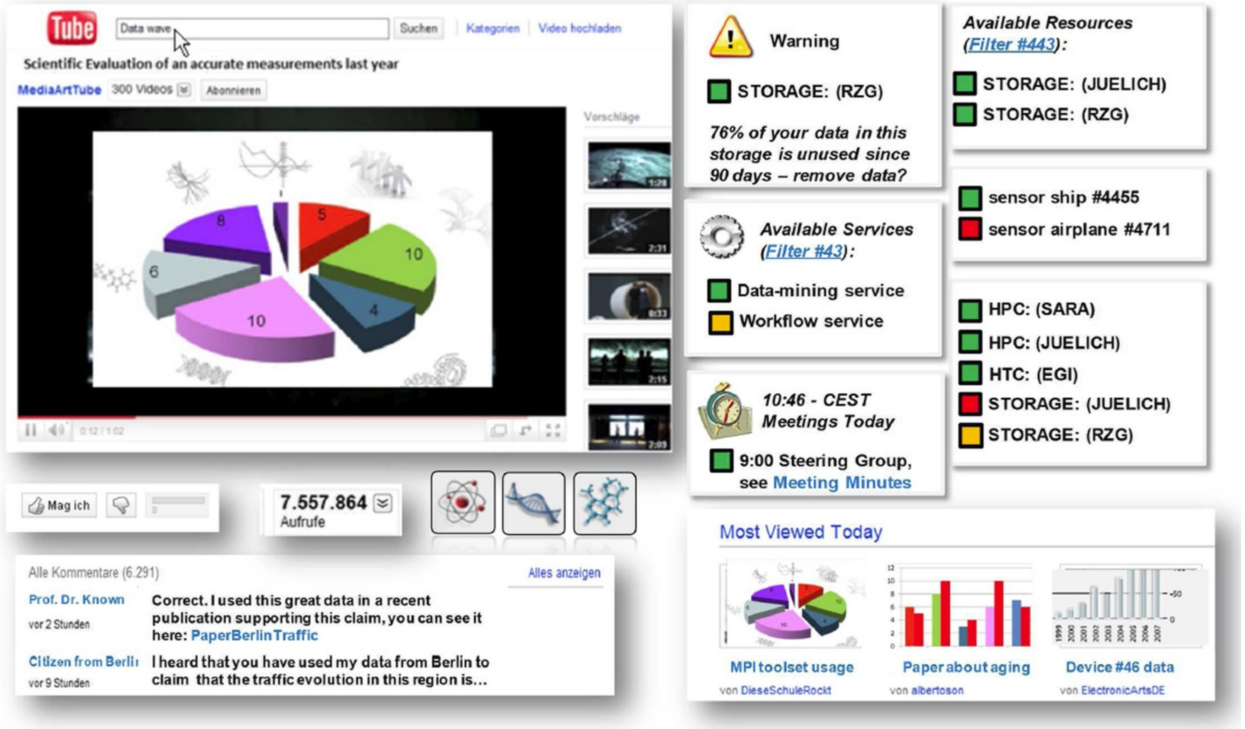
The entities follow a service-based approach and actual data services deployed on infrastructures may have certain performance characteristics, but the concept of the individual service types are relevant and not the particular deployment. Figure 3 illustrates not a particular reference installation and also not a concrete deployment of implemented reference architecture core building block services. Instead the focus is on the concept of an abstract service entity and its relationships to other entities. Another principle followed is that the reference defines entities *'within a particular problem domain'* that is set as services around the particular high-level *'Science-Tube'* vision and the underlying concrete data infrastructure. Another key principle of the reference model design is that it is *'independent of specific standards, technologies, or implementations'* making it *'technology-agnostic*.

Figure 3 illustrates the major design foundation of the reference model that follows an overall 'federated infrastructure approach' in the blueprint. This means that entities are deployed on community and data centers and that result is a truly *'collaborative infrastructure'*. The infrastructure is thus not created *'for'* the scientists but instead together *'with the scientists'* that is a major change of views compared to many methods from IT infrastructures applied in the last decade.

3.3 Reference model entities and relationships

The reference model defines entities for key functionality and relationships between them. The major first entity of the reference model are *'virtual workspaces'* that represents a kind of *'workbench'* where services and resources are conveniently accessed. This entity is illustrated in Figure 3 as a virtual overlay of the backend functionality existing at data centers. It is accessed via lightweight Web-based GUIs as shown in Figure 2.

The 'virtual workspace' maintains a *'list of profiles'* to enable scientists to configure different 'workbenches'. This enables scientists with different roles to configure their 'virtual workspace' as needed for each role (e.g. organizational, research project, etc.). *'Pre-configured filters'* shows only a limited set of services instead of potentially hundreds of available domain-specific infrastructure services.



Correct. I used this great data in a recent publication supporting this claim, you can see it here: PaperBerlinTraffic

I heard that you have used my data from Berlin to claim that the traffic evolution in this region is...

76% of your data in this storage is unused since 90 days – remove data?

Figure 2 The overall goal is to achieve a 'ScienceTube' based on federated data infrastructures to support science.

Another filter can be applied to scientific measurement devices and to data archives to enable a focussed view on information. The 'virtual workspace' informs about existing *'end-user quota'* on data resources. It also informs scientists about their *'community time'* on rare resources such as large-scale computational resources on which time is granted after peer-review processes. This includes scheduled time periods as part of large experiments (e.g. collider beam) or devices (e.g. telescope). The elements of the 'virtual workspaces' entity are directly connected to federated security methods based on *'user access policies'*. This includes identity management and authentication, but also authorization techniques while at the same time retain a local access control by resource providers.

A large part of the 'virtual workspace' entity are various *'service adapters'*. This collectively stands for multiple adapters that bring functionality from services into the workspace. Examples include integrated clients for data-mining services or simple lookup services or integrated APIs for submitting data-intensive computational jobs on different types of compute resources. Domain-specific adapters for scientific workflow services (e.g. WebLicht [9]) can be provided or more general adapter for widespread storage technologies (e.g. iRODS [10]). Other 'service adapters' tackle the *'data wave issues'* highlighted in Section "Motivation" such as data recommendation systems or data de-duplication check services. Another element that visualizes all the previous features is named as *'core functions'* with dynamic Web 2.0 functionality or mash-ups.

Table 1 lists elements of the 'virtual workspace' but there are also a wide variety of other entities. The *'Application Server'* entity represents the functionality that hosts the virtual workspaces and parts of Web-based Virtual Research Environments (VREs) to make them accessible to scientists. As shown in Figure 3, the 'ScienceTube' features are thus only one element of VREs alongside many other VRE elements (e.g. other services). *'Archives/Repositories'* are another entity that collectively represent the broad spectrum of existing scientific domain-specific data. This is deeply connected to data models encoded in the archives or the valuable meta-data assigned to scientific data-sets. The *'storage'* technologies entity is also a crucial part of the reference model. This provides access to different types of storage (e.g. tapes, disks, etc.).

Also *'compute resources'* are an important entity since data-intensive processing must be conveniently possible using High Throughput Computing (HTC) resources (e.g. with MapReduce techniques) or High Performance Computing (HPC) resources (e.g. for complex climate prediction simluations). Large-scale sources of measurement data are summarized as 'Devices', including ships with sensors or large-scale telescopes.

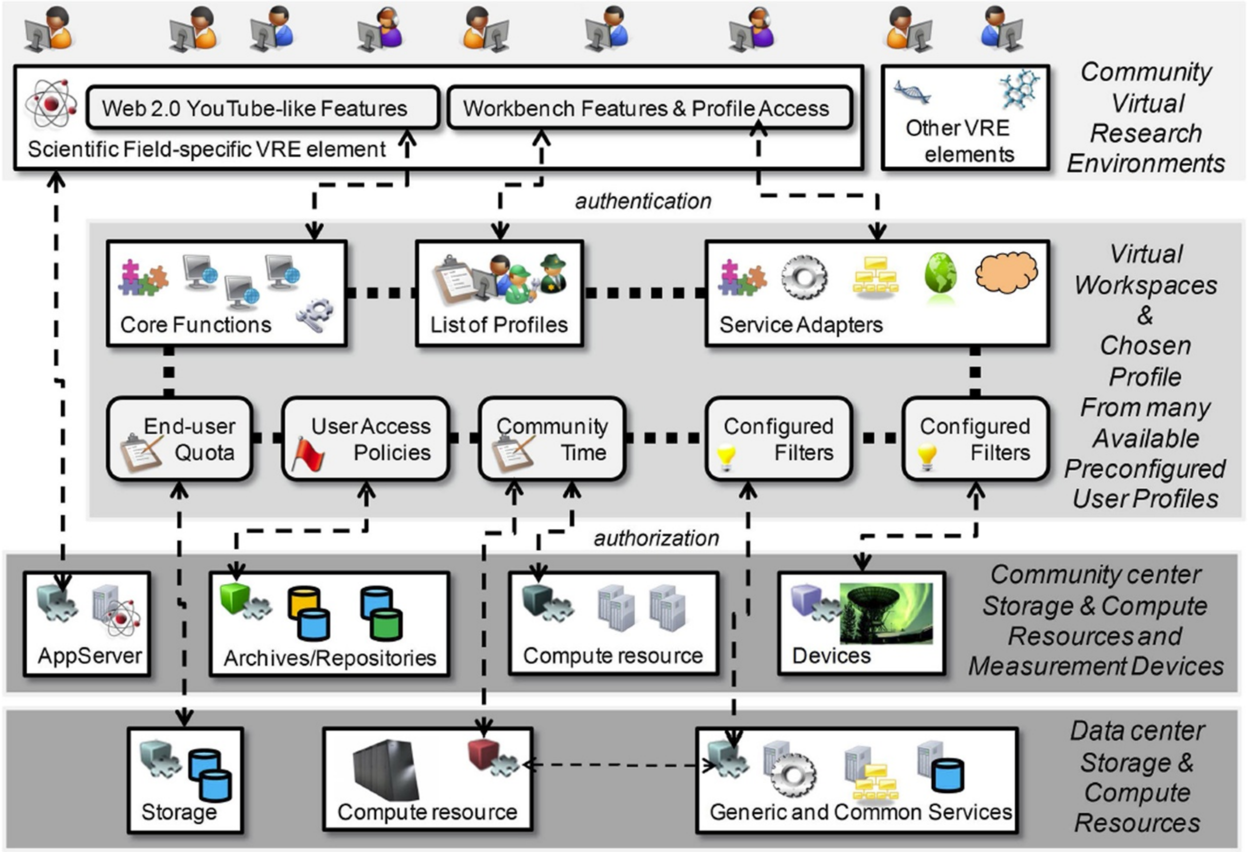

Web 2.0 YouTube-like Features
Workbench Features & Profile Access
Scientific Field-specific VRE element
Other VRE elements
Community Virtual Research Environments
authentication
Core Functions
List of Profiles
Service Adapters
End-user Quota
User Access Policies
Community Time
Configured Filters
Configured Filters
Virtual Workspaces & Chosen Profile From many Available Preconfigured User Profiles
authorization
AppServer
Archives/Repositories
Compute resource
Devices
Community center Storage & Compute Resources and Measurement Devices
Storage
Compute resource
Generic and Common Services
Data center Storage & Compute Resources

Figure 3 Conceptual view of the reference model and its entities derived from the high-level goal 'ScienceTube'.

The final entity refers to a broad spectrum of services required to run a production data infrastructure with *'Generic and Common Services'*, which in many cases are often also inter-disciplinary in nature. Firstly, infrastructure support services are required such as monitoring of services and resources as well as ticketing system for user help desks and support. Many scientific communities share the demand for a persistent identifier (PID) service such as the European Persistent Identifier (EPIC)[g] service. This service provides the functionality to assign each data object as part of the data infrastructure a unique ID making the data clearly referencable. Services (e.g. Shibboleth Identity Providers) that support the creation of a federated AAI infrastructure model. Accounting services are needed in order to support quotas and tracking the resource usage. Scientific workflow services (e.g. WebLicht [9]) that are of interest for a broader range of user communities are also part of the 'generic and common services' entity.

4 Data replication service use case

The reference model follows a 'federated model' between user community centers and data centers. Each of the centers have different specialities and boundary conditions leading to questions we need to answer in order to define a more concrete architecture. Hence, in order to derive a more concrete architecture guided by this reference model setup, the requirements from real end user communities need to be more clearly analyzed.

This analysis follows a process in which each of the different use cases drives the architectural design layout thus emphasizing again that end-users decide what services are provided, while the exact how is rather driven by technical constraints. This prevents the creation of big architectures that are not underpinned with user requirements leading to unnecessary services that are not really required. Instead, our approach ensures a slim architecture since the concrete architectures derived from the abstract reference model are underpinned by user requirements of scientific stakeholders.

While this contribution focusses on the *'data replication use case'*, there are other use cases that are tackled in parallel and will be published in other contributions. Examples of other user cases include 'data staging' from one center to another in order to perform computational activities or the 'simple store use case' that aims to work

Table 1 Reference model entities overview

Name	Short Description
Virtual Workspaces	(VW1) List of Profiles (VW2) Configured Filters (VW3) End-user Quota (VW4) Community Time (VW5) User Access Policies (VW6) Service Adapters (VW7) Core Functions
Application Server	(AS1) Hosting 'Virtual Workspaces' and Virtual Research Environment
Archives Repositories	(AR1) Scientific domain-specific datasets including different data models and meta-data for its description
Storage	(ST1) Storage technologies providing convenient access to different types of storage capacity
Compute Resources	(CR1) Seamless access to HPC and HTC resources
Devices	(D1) Large scientific devices as sources of measurement data
Generic and Common Services (Inter-disciplinary Services)	(G1) Infrastructure Support Services (G2) Persistent Identifier Services (G3) Federated AAI Services (G4) Accounting Services (G5) Workflow Services (G6) Other services...

on a dropbox-like functionality. Other use cases is the 'handling of metadata' and 'authorization and authentication infrastructure' that both are needed by end-users. All these use cases will together shape an architecture for a data infrastructure tuned for stakeholder needs following a federated model. One of the goals to create numerous architecture work elements is to better compare the infrastructure architecture with those of others in order to seek synergies (e.g. commonly deployed services) with other data infrastructures (e.g. PanData [11]) and to support the common understanding in the complex arena of 'big data'.

4.1 Scientific community stakeholders

The general stakeholders are projects emerging from the ESFRI roadmap and other larger scientific communities such as the VPH network of excellence. The stakeholders of the EUDAT task force 'safe data replication' [12] are summarized in Table 2. While a more thorough description is provided in [12], the reason why a data replication service is needed is to improve data curation and accessibility. The added value is thus to replicate data from community data centers to other large data centers to improve the reliability and access to computational resources. Several scientific communities work closely together with large data center representatives in order to understand and implement such a 'data replication service'. In the context of the EUDAT task force, only several data-sets are replicated in order to bring the 'date replication service' towards production meaning that some communities choose to replicate data from some of their specific scientific projects and not all of them.

The ENES community is involved via the Deutsches Klimarechenzentrum (DKRZ)[h] and have interest to replicate data to the Juelich Supercomputing Centre (JSC)[i] in Germany and the IT Center for Science (CSC)[j] in Finland. Another community involved is EPOS represented by the Instito Nazionale di Geofisica e Vulcanologia (INGV)[k]. They aim to replicate their data to the data centers of CINECA[l] in Italy and SARA[m] in Netherlands. Another scientific community is CLARIN with representatives of the Max Planck Institute for Psycholinguistics (MPI-NL)[n]. This community intends to replicate data to Rechenzentrum Garching (RZG)[o] in Germany and SARA in Netherlands. Finally, VPH is involved via University College London (UCL)[p] that want to replicate data to CINECA in Italy, SARA in Netherlands, and Poznan Supercomputing Centre (PSNC)[q] in Poland.

4.2 Use case analysis

Interviews have been performed with the stakeholders in the previous paragraph in order to better understand the demand for a 'data replication service' and a structured analysis of their outcomes is given in Table 3. This use case analysis presented in Table 3 is derived from a method by Malan and Bredemeyer [13] that has proved to be effective as part of our XSEDE infrastructure architectural design process [14].

Table 2 Participating community and data center

Scientific Community	Community Centers	Data Centers
CLARIN	MPI-PL	RZG, SARA
ENES	DKRZ	JSC, CSC
EPOS	INGV	CINECA, SARA
VPH	UCL	CINECA, SARA, PSNC

Table 3 'Data replication service' use case analysis

Use Case I	Data Replication Service
Description	Create and automatically execute a safe replication policy on specified data objects in the infrastructure between community and data centers
References	EUDAT Newsletter April 2012 [12]
Actors	(A1) Scientific Community Center Site Manager (CCSM)
Prerequisites (Dependencies, (Assumptions)	(P1) Federated infrastructure approach (P2) Each data object has unique PID (P3) Data access permissions remain (P4) Federated AAI concepts adopted
Steps	(S1) Create policy P for M replications (S2) Specify the target data centers C (S3) Exclude centers X from policy (S4) Define replicated data lifetime T (S5) Policy is saved and executed (S6) Policy-based replication is done (S7) Data objects are safely replicated (S8) Replicated data stored in local Long-term Archives (LTAs)
Variations	(V1) Existing policy P might be updated by the CCSM (V2) Additional manual data replication if needed by the CCSM
Quality Attributes	(QA1) Reliability of replication (QA2) More optimal data curation (QA3) Better accessibility of data
Non-functional	(NF1) Usability for creating and executing the replication policy
Issues	(I1) Federated AAI concepts work still work-in-progress

Table 3 clarifies the actors that play a role in context and provides a clear step-wise view on the required functionality based on specified assumptions. The variations are listed to have those activities noted that are not directly in-line with the previous described step-wise approach (e.g. manual interventions by actors). Another part of the analysis are the quality attributes that also communicate benefits for the 'safe replication service' users and often related non-functional requirements (e.g. usability, reliability, etc.). Finally, issues in realizing the production setup of the use case are ongoing research activities (e.g. Federated AAI).

4.3 Derived stakeholder requirements analysis

Interviews have been performed with scientific community representatives that are the stakeholders of the 'data replication service' [12]. The summary of the requirements have been analysed and set into context of the reference model and possible derived reference architecture core building blocks. Table 4 lists the identified core building blocks in context of the requirements stated by the stakeholders.

5 Architecture for data replication use case

This section aims to provide more concrete details about various *'architecture work'* derived from the 'data replication service' use case guided by the reference model introduced in Section "Data infrastructure reference model" as frame of reference. Figure 1 shows the different levels of architecture work, including a reference architecture with abstract core building blocks and a more concrete architecture. This section will also provide information about implementation activities using technologies, tools, and services. This includes challenges faced while implementing this architecture setup for production summarized as *'data infrastructure integration activities'*. The reference model in Section "Data infrastructure reference model" gives the basic blueprint with abstract entities and relationships that are more concrete in this section and focussed on the specific use case of a 'data replication service'. The resulting architecture thus does not describe the whole data architecture nor all services that will be available in EUDAT that provides functionality for many other required service use cases too. Instead, it can be considered as a first step in providing a more concrete architecture description of a reference model with aligned architecture work specifically based on the real 'data replication service' use case. The architecture work accounts for the motivation raised in Section "Motivation", the overall goal described in Section "Introduction" and the specific requirements raised in Section "Data replication service use case". Subsequent activities during the course of the next years will provide more concrete architecture work elements for other service use cases (e.g. data transfer, meta-data search, etc.) and over time produce a more detailed reference architecture.

5.1 Use case derived reference architecture

This paragraph introduces the reference architecture illustrated in Figure 4 being derived from the use case analysis results from Section "Data replication service use case" and being guided by the reference model design in Section "Data infrastructure reference model". The reference architecture provides several abstract core building blocks in order to remain technology-agnostic thus contributing to the software engineering principle that the architecture should be seperated from its implementation

Table 4 'Data replication service' use case requirements

Architecture Core Building Block	'Data Replication Service' Stakeholder Requirements
Mature and Extensible Storage Technology	(R1) The technology required must be mature to guarantee a highly available and robust service. (R2) CCSMs need to have functionality which data objects and collections need replication. (R3) A policy-based feature enables that all centers can be audited (e.g. DSOA) in order to establish trust with clearly described policy rules. (R4) Powerful policy functions are required to enable CCSMs to specify M replicas to be stored for N years.
Persistent Identifier Service	(R5) Each data object in the data infrastructure should have a clearly assigned PID (R6) The use of PIDs for replicated data objects enables CSMs to know whether the replicas are identical with the source.
Monitoring Service	(R7) The infrastructure services must be monitored in order to obtain information about their production status.
Federated AAI APIs	(R8) CCSMs need replicas to be accessible by users while maintaining the access permissions as defined by the originating community center.
Local Long-term Archiving	(R9) The storage technology should require as little changes as possible on the community data organization side that is already established around the local LTAs.
Common Services	(R10) Ticketing service and help desk support should be established for end-users.
Web-based Workbench	(R11) Virtual workspaces making the services accessible to users.

as best as possible. The core building blocks itself are put in context to the reference model entities in Table 5. One of the reference model design foundation is the *'federated approach'* that is applied to the core building blocks of the data architecture as shown in Figure 4. The positioning of core building blocks is obtained from the use case that aims to replicate data from community centers to data centers.

5.2 Technology assessment

Table 6 provides an overview of the assessed technologies for defining the implementation of a concrete architecture driven by the 'data replication use case'. The listed technologies are either already available at centers including their knowledge or are known to be stable and mature technologies in order to prevent new developments to the most possible degree.

The core building block (RA1) requires some development efforts in order to integrate all the different service functionalities and to make it accessible in a Web-based workbench. The (RA2) data models with meta-data are also existing at the scientific community centers already but are part of the architecture to underline their importance in the data infrastructure and their possible numerous relationships to different storage services. In fact, several investigations in the organization of (RA2) points to the use of file system hierarchies for scientific data (e.g. DKRZ) and there are expected benefits when using a storage technology in conjunction with this organization scheme. Also, (RA4) referring to local long-term archiving solutions are largely already existing at the different centers as well, but needs to work together with the storage technologies of choice.

In terms of (RA3), there are several mature and extensible storage technology systems available in the academic field of infrastructures such as the Disk Pool Manager (DPM) [15], dCache [16], and iRODS [10]. The different involved centers have experience with all of these technologies and all of them seem to be mature and usable. But when taking into account the stakeholder requirements listed in Table 4, in particular the policy-based functionality, the iRODS technology is the most appropriate system. As early evaluations reveal, iRODS also integrates well with existing local LTA technologies (e.g. file systems).

In terms of (RA5), there are two persistent identifier services of interest named as DataCite/DOI [17] and the EPIC service. Several partners of the EUDAT scientific communities (e.g. CLARIN) and centers (e.g. RZG) have already experience with EPIC as part of the REPLIX project [18] and our work is build on top of this approach. Also in terms of (RA6), the NAGIOS system [19] with its extensibility using a probe-based approach was choosen as a monitoring solution, because expertise was existing at the majority of centers. For (RA8), the ticketing

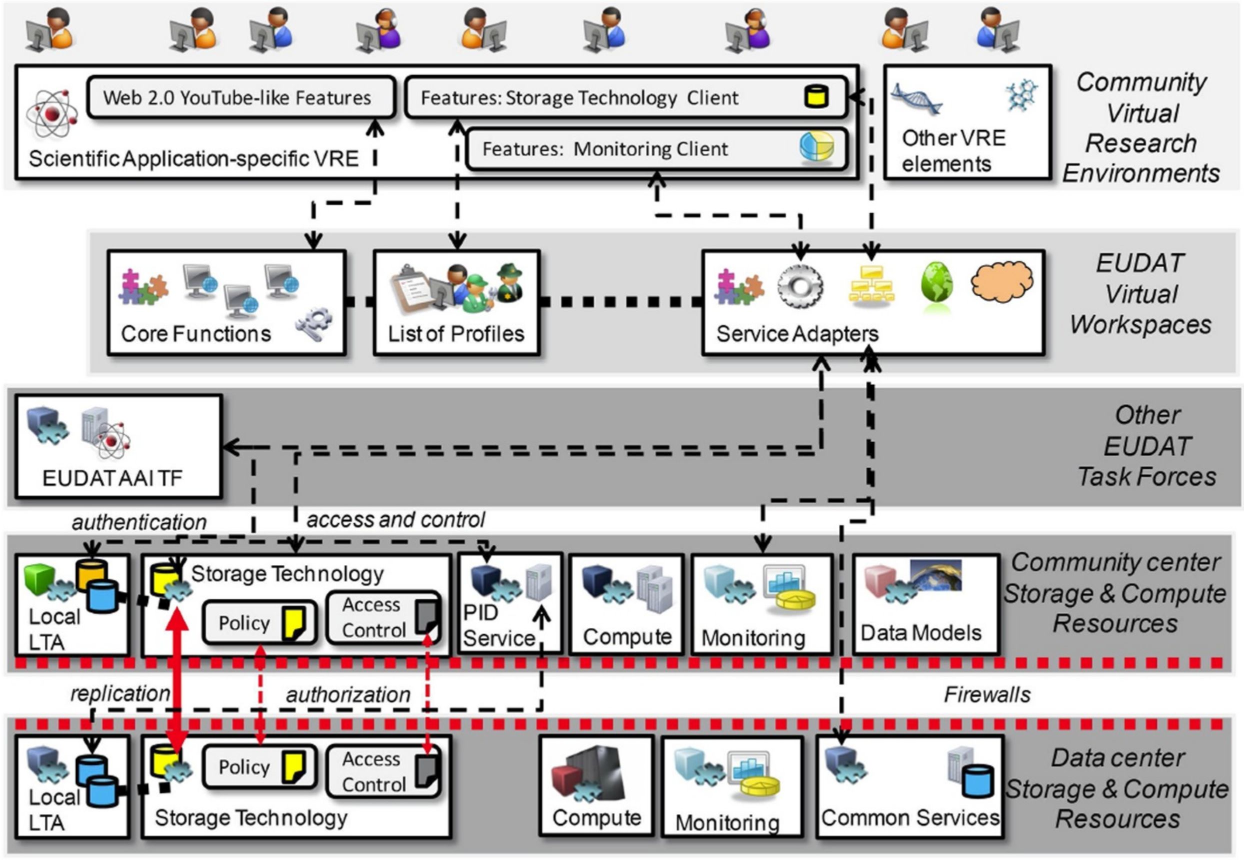


Figure 4 Conceptual view of the reference architecture with core building blocks derived from the use case.

technology of choice is JIRA mostly because it has been also already used as part of the EUDAT project internal tasks. Finally, the (RA7) implementation is most complex since it also affects all other core building blocks. Shibboleth and contrail have been analyzed and so far Shibboleth is the most promising candidate to realize a federated AAI service given its support by educational organizations in Europe.

5.3 Concrete architecture and implementation

Based on the aforementioned reference architecture core building blocks and the subsequent technology assessment, this section aims to describe one possible derived architecture implementation as illustrated in Figure 5. The architecture is described by addressing the concrete stakeholder requirements for the 'data replication service' listed in Table 4.

As Figure 5 reveals, the (RA3) core building block is located in the architecture on the community and data center side by using iRODS. By using this mature technology, the architecture is able to address requirement (R1) and the policy-based mechanisms of iRODS using 'micro-services in rules' [10] (see next paragraph for more details) addresses requirement (R2), (R3), and (R4). CCSMs are able to specify which data objects need replication and the enormous extensibility of iRODS with user-specific rules enable CCSMs to specify those objects that need M replicas for N years being automatically executed and thus enforced.

Table 5 'Reference architecture core building blocks'

Reference Model Entity	Reference Architecture Core Building Block
Virtual Workspaces	(RA1) Web-based Workbench
Archives / Repositories	(RA2) Data models with meta-data of scientific stakeholders
Storage	(RA3) Mature and Extensible Storage Technology (RA4) Local Long-Term Archiving
Generic and Common Services	(RA5) Persistent Identifier Service (RA6) Monitoring Service (RA7) Federated AAI Service (RA8) Ticketing service

Table 6 'Architecture core building blocks with potential technologies that have been assessed for their usage'

Reference Architecture Core Building Block	Potential Technologies
(RA1) Web-based Workbench	Requires development
(RA2) Data models with meta-data of scientific stakeholders	Existing at the community centers
(RA3) Mature and Extensible Storage Technology	DPM, dCache, iRODS
(RA4) Local Long-Term Archiving	Already available at the centers
(RA5) Persistent Identifier Service	DataCite/DOI, EPIC/Handle
(RA6) Monitoring Service	NAGIOS
(RA7) Federated AAI Service	Shibboleth, Contrail
(RA8) Ticketing service	JIRA

The aforementioned architecture element raise the requirement for security functionality in terms of enabling authentication and authorization and keeping identities while replicating data from community centers to data centers. Figure 5 shows that simple Access Control Lists (ACLs) can be used as part of iRODS, but that a use of a federated AAI solution is much more convenient. The related (RA7) is currently defined by the EUDAT AAI task force and the details are kept out of this paper to remain the focus on the 'data replication service'. Using the solution of the EUDAT AAI task force (e.g. a Shibboleth service) addresses requirement (R8) so that replicas are accessible by users while the access permissions as defined by the orginating center is kept. Getting iRODS to work with this security approach is one of the *'infrastructure integration issues'* (I1) we found during the implementation of the data infrastructure architecture. Those issues raise the demand for smaller integration developments to get several pieces of the architecture together and are described in more detail in the subsequent paragraph.

The (RA4) architecture core building block in Figure 5 addresses the requirement stated in (R9) in the sense that the iRODS system works seamlessly together with the

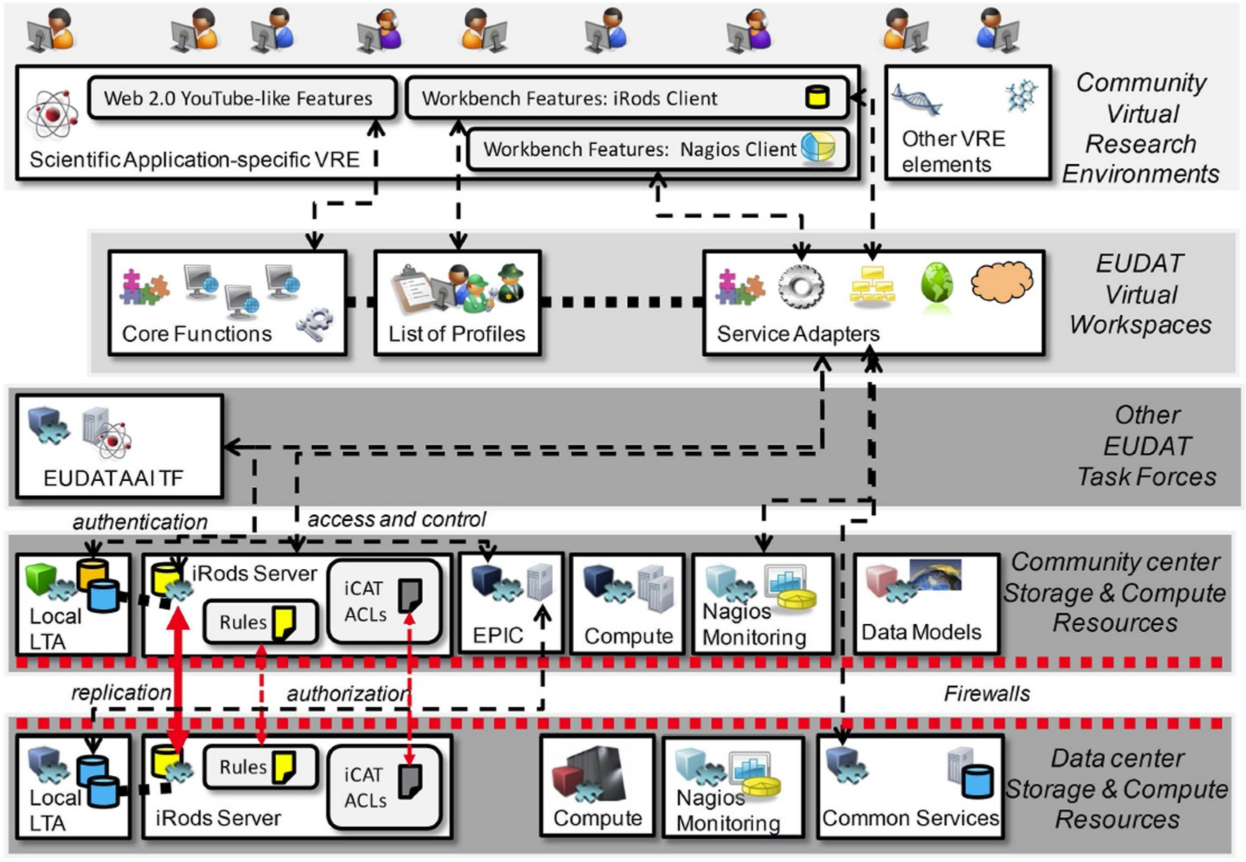


Figure 5 Conceptual view of the concrete architecture after the technology assessment satisfying stakeholder requirements.

local long-term archiving technologies at both the community centers and data centers. The scientific data and their models (RA2) are stored in the local LTAs (e.g. file systems) and are made available in the community centers via iRODS to the data infrastructure. Then, the replication rules automatically replicate this data to iRODS servers at the data center side in order to get stored in their local LTAs (e.g. storage disks, tape robots, etc.).

Another core building block is (RA5) implemented by the EPIC PID service technology as illustrated in Figure 5. It enables the basic functionality to address (R5) by providing the functionality to release PIDs, but still there must be an inter-working with the iRODS server to really get each of the replicated data objects assigned with a PID. Hence, this reveals another infratructure integration issue (I2) raising the demand to have iRODS working together with the EPIC service. In order to address (R6), the EPIC released PIDs also need to have a mechanism to update or add locations during the process of replication. In both cases, the access of EPIC to have an initial PID for a data object and the update of potentially multiple locations, require a link to the federated AAI setup in order to ensure that the use of PIDs is protected. This points to another issue (I3) again pointing out that production technologies are available, but still need integration development to make it work with the data infrastucture.

The architecture implementation in Figure 5 reveals another core building block on both the community and data center side that is (RA6). The NAGIOS system can be used here to monitor the health status of the underlying storage backend servers, as well as the iRODS service, and the EPIC service. This addresses (R7) and enables that the data infrastructure services are monitored in order to obtain information about their production status at any time. But each of the different NAGIOS probes need a small amount of developments in order to work with the different services that raises another infrastructure integration issue (I4).

The requirement (R10) is addressed by using the (RA8) core building block and JIRA as a ticketing service that is illustrated in Figure 5 as part of the common services hosted by the data centers. This collectively also stands for a help desk support method This service is mature and expected to work well with federated AAI solutions not necessarily requiring security integration-related developments, but in order to emphasize that the security is related also to this service it is part of the infrastructure architecture too.

In contrast, the realization of the core building (RA1) requires more integration development work in realizing partly community-specific Web-based workbenches that we collectively note as infrastructure integration issue (I5).

In order to address requirement (R11), this core building block has numerous parts as illustrated in Figure 5 that go beyond the aforementioned core functions around the Web 2.0 elements. It firstly needs to work well with the federated AAI solution emerging from the EUDAT AAI task force and secondly needs to offer convenient access to iRODS functionality (e.g. easy writing and monitoring of iRODS rules). The former is specifically important to realize the support for different roles of scientists that need to be configured as different profiles in the virtual workspace. The latter raise the demand for service adapters that are able to work with the backbone services of the data infrastructure (e.g. iRODS, NAGIOS, etc.) and that provides APIs to be used by the used Web 2.0 front-end and lightweight security frameworks.

It is also required to integrate information about the availability of the data services into the virtual workspace that can be realized by using a lightweight NAGIOS client.

5.4 Infrastructure integration challenges

The technology assessment and concrete derived technologies suggests that every technology is already available with a rich set of features and just can be used as production services as is. The implementation of the data architecture, however, reveals several *'infrastructure integration issues'* that require smaller developments in order to get the different building blocks of the architecture to work together as one ecosystem. The previous architecture description identified these issues and they are summarized in Table 7 together with potential approaches.

Because of the page limit of this contribution we can not provide detailed information about all of our taken approaches in order to overcome the issues. Instead, we focus here on the issue (I2), because it highlights best how particular key features of iRODS are used and how

Table 7 'Infratructure integration issues overview'

Data Infratructure Integration Issue	Potential Approach
(I1) Federated AAI-enabled iRODS servers	Shibboleth-enabled iRODS server
(I2) iRODS and EPIC inter-working	iRODS rules that make callouts to EPIC
(I3) Federatd AAI-enabled EPIC service	iRODS rules that make callouts to EPIC
(I4) Monitoring information obtained from services and underlying storage resources	NAGIOS probes for EPIC and iRODS and hardware storages
(I5) Web-based workbench that makes infrastructure services seamlessly available in a secure manner	Web 2.0 and light security frameworks to integrate NAGIOS and iRODS with federated AAI

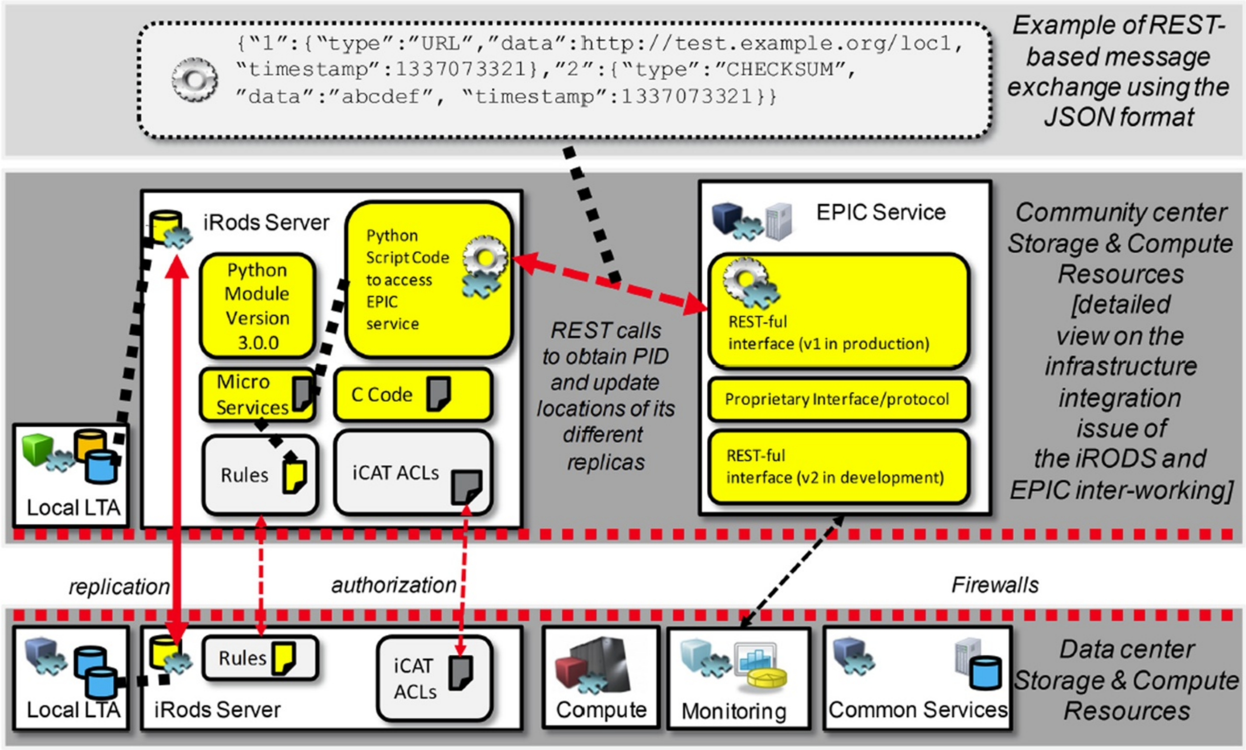
in general infrastructure integration issues are connected to development efforts. A more detailed overview of issue (I2) is therefore illustrated in Figure 6.

One of the expected functionalities of the *'data replication service'* is to ensure that the digital objects stored within the iRODS service get a persistent identifier known as PIDs. As Figure 6 illustrates, iRODS does not provide a solution for that and development work is needed to achieve an integration of EPIC with iRODS. While there are many solutions for PID, previous sections revealed that in EUDAT it was decided to use PIDs provided by the EPIC service. This service in turn is based on Handle System[r] that defines the format of the PIDs and the protocol used.

Although it might sound trivial to integrate the two services with each other, the development efforts are quite complex and require deep knowledge of both the EPIC APIs and the existing iRODS extension mechanisms around micro-services as part of rules. Hence, for the sake of simplicity we will present a less-detailed overview of our solution. In short, in our case the registration and manipulation of the PIDs boiled down to exchanging Java Script Object Notation (JSON) [20] representations of a PID handle via the HTTP protocol. An example of a handle in JSON format is illustrated in the upper part of Figure 6. The particular challenge is to integrate this message exchange with the Representational State Transfer (REST)-based service [21] API of EPIC within iRODS. We have written a Python-based client software to interact with the EPIC service.

The software firstly serializes a python hash table keyValues into JSON format with help of the SimpleJson library[s]. This library is a simple and fast JSON encoder/decoder for Python. Afterwards, a http client is prepared and we use http authentication with username and password for simplicity while in parallel the federated AAI task force in EUDAT is exploring better options. It is also necessary to set content-type header to inform the server that the PID will be sent in the JSON format. After these preliminary steps, it is possible to issue a http PUT request on a given URI with JSON representation of the new handle in the request body. The URI is composed of service address, handle prefix and handle suffix. It is also possible to delete a PID from the handle system by invoking a HTTP DELETE request.

The aforementioned python-based EPIC client can be easily integrated into data management policies in iRODS as illustrated in Figure 6. This is done by using PyRods and EmbededPython modules provided by irodspython project[t]. PyRods provides a Python API for iRODS (e.g. to access and manipulate data objects in iRODS) and EmbededPython allows to write iRODS micro-services in



```
{“1”:{“type”:”URL”,”data”:http://test.example.org/loc1,
“timestamp”:1337073321},”2”:{“type”:”CHECKSUM”,
”data”:”abcdef”, “timestamp”:1337073321}}
```

Example of REST-based message exchange using the JSON format
iRods Server
Python Module Version 3.0.0
Python Script Code to access EPIC service
Micro Services
C Code
Rules
iCAT ACLs
Local LTA
REST calls to obtain PID and update locations of its different replicas
EPIC Service
REST-ful interface (v1 in production)
Proprietary Interface/protocol
REST-ful interface (v2 in development)
Community center Storage & Compute Resources [detailed view on the infrastructure integration issue of the iRODS and EPIC inter-working]
replication
authorization
Firewalls
Local LTA
Rules
iCAT ACLs
iRods Server
Compute
Monitoring
Common Services
Data center Storage & Compute Resources

Figure 6 **Technical view to make iRODS and the EPIC service work together using rules and REST call-outs.**

Python. Our EPIC client presented above provides a set of micro-services for manipulating PIDs that we need to make available in iRODS. In this context, EmbededPython allowed to call Python functions directly from iRODS rules. An example of an iRODS rule that uses createHandle function defined above to register a "testHandle" PID in the PID service currently provided by a data center named SARA is given in Figure 7(a).

We assume that the Python code is stored in the location pointed by the *pyScript variable. As the next step, it is possible to integrate this EPIC call into data management policies of iRODS. An example for a rule which is called each time a new file is ingested in a given collection (i.e. "/zone/monitored/") and then creates a PID for the new data object is given in in Figure 7(b). It is a simplified version which uses fileName of the ingested data object as a PID suffix in order to better undertand the idea and not dive too deeply into the partly complex part of PID prefixes, suffixes, etc. For more information about these details we refer to the information available via the EPIC Website[u].

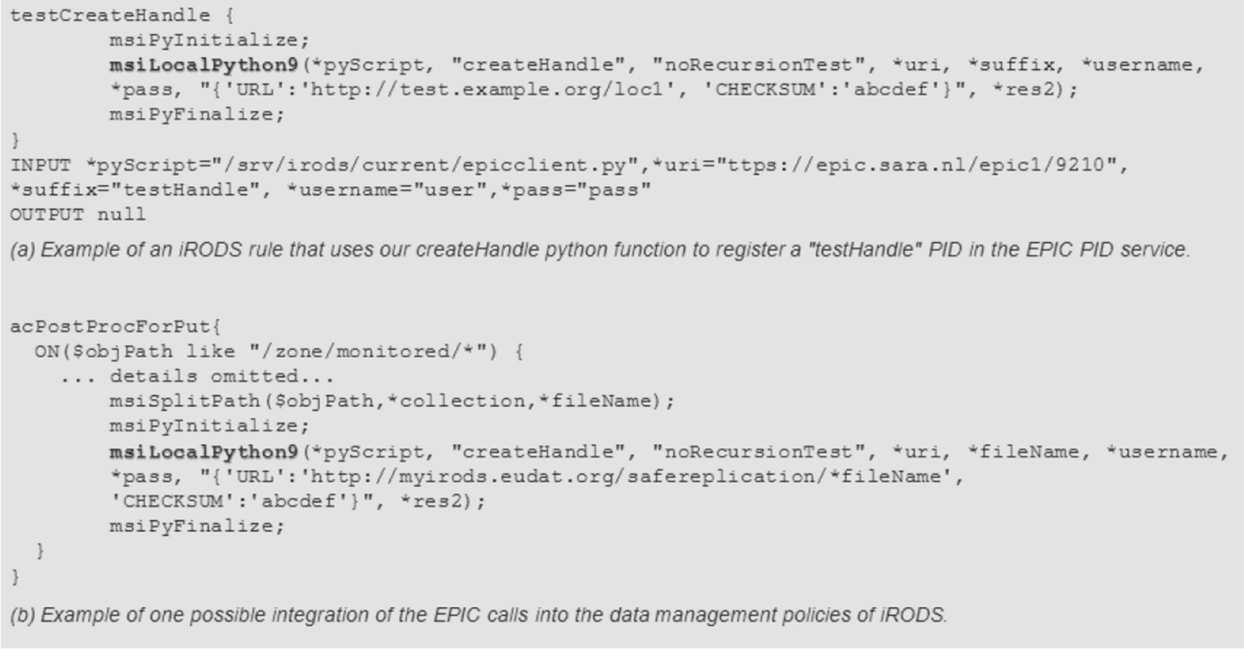
To sum up, we have presented only a part of our solution but it should be clear to the reader that the solution can be easily extended to define more sophisticated PID manipulation functions in the Python script on one hand, and more sophisticated policies in iRODS on the other. In fact the version used in production in EUDAT is much more complicated but based on the same basic concepts as presented above. Nevertheless, we believe that this already provides insights into the infrastructure integration issues that are commonly overlooked since production technologies are there, but how they seamlessly work together is in many cases still a matter of integration developments.

6 Related work

There is a wide variety of related work in the context of data infrastructures and possible architectural approaches. We set a focus on one particular data infrastructure activity that is relatively close to the ideas of EUDAT but for a smaller set of communities. In contrast to our here described approach that is much based on abstract entities and derived architecture core building blocks, the following data infrastructure activities follow a more standard-based approach within the project.

EUDAT members work on a more broader idea of scientific data infrastructure standardization to order to go beyond one single project that culminated in an Europewide activity named as the Research Data Alliance (RDA) [12]. This activity aims to create an open forum for discussing and agreeing on data-related standards, APIs, policy rules, data interoperability covering also those infrastructure integration issues raised earlier in this contribution (AAI, PIDs, etc.) to the semantic and regulatory levels.

One of the most known distributed data infrastructure was established by the High Energy Physics (HEP) community leading to the World-wide Large Hadron Collider Computing Grid (WLCG) [22].



```
testCreateHandle {
        msiPyInitialize;
        msiLocalPython9(*pyScript, "createHandle", "noRecursionTest", *uri, *suffix, *username,
        *pass, "{'URL':'http://test.example.org/loc1', 'CHECKSUM':'abcdef'}", *res2);
        msiPyFinalize;
}
INPUT *pyScript="/srv/irods/current/epicclient.py",*uri="ttps://epic.sara.nl/epic1/9210",
*suffix="testHandle", *username="user",*pass="pass"
OUTPUT null
```

(a) Example of an iRODS rule that uses our createHandle python function to register a "testHandle" PID in the EPIC PID service.

```
acPostProcForPut{
  ON($objPath like "/zone/monitored/*") {
    ... details omitted...
        msiSplitPath($objPath,*collection,*fileName);
        msiPyInitialize;
        msiLocalPython9(*pyScript, "createHandle", "noRecursionTest", *uri, *fileName, *username,
        *pass, "{'URL':'http://myirods.eudat.org/safereplication/*fileName',
        'CHECKSUM':'abcdef'}", *res2);
        msiPyFinalize;
  }
}
```

(b) Example of one possible integration of the EPIC calls into the data management policies of iRODS.

Figure 7 **Examples of iRODS rule using EPIC (a) and its integration into iRODS data management policies (b).**

PaNdata is a research initiative to realize a pan-European information infrastructure supporting data oriented science. The first step of this effort was PaNdata Europe that aimed at formulating baseline requirements and standards for data policy, user information exchange, data formats, data analysis software, and integration and cross linking of research outputs. After this project has been concluded in 2011, the PaNdata community started the PanData ODI (Open Data Infrastructure) project funded by the EU. This project take the aforementioned requirements and standards as a foundation for providing data solutions, and constructing sustainable data infrastructures, but specifically serving European Neutron and Photon communities. Hence, in contrast to EUDAT, the amount of participating communities is smaller than in EUDAT that aims to offer federated solutions together with all thematic groups of the ESFRI projects. However, it is important to understand that the insights of the aforementioned communities also support a wide range of other disciplines such as physics, chemistry, biology, material sciences, medical technology, environmental, and social sciences. Important applications in context are crystallography that reveals the structure of viruses and proteins and neutron scattering that is used within engineering components, and tomography provides fine-grained 3D-structures of the brain.

The PaNdata ODI will provide their participating scientific facilities with an intuitive and parallel access to the data services which include the tracing of data provenance, preservation, and scalability. It aims to deploy and integrate generic data catalogues within their target communities. Interesting in context of this contribution is that they follow a reference model named as Open Archival Information System (OAIS) [23] that is in contrast to our approach much more focussed on long-term preservation. This also includes the usage of HDF5[v] and Nexus[w] standards. In order to allow access across different institutions, a federated identity management system will be used that is in-line with the contribution of this paper and the activities performed by the EUDAT AAI task force. In order to provide scalability, the parallelization techniques will be used by developing parallel Nexus API (pNexus) with pHDF5[x]. This will enable PaNdata to avoid sequential processing and to manage the simultaneous data inflow from various sources such as X-Ray free electron lasers for instance.

To sum up, the PaNdata activities are highly relevant in the context of EUDAT and provide enormous potential for exploring synergies (e.g. long-term archiving of scientific measurement data). But both projects have a very different focus that is serving the needs of the neutron and photon communities in PaNdata, while EUDAT specifically supports multi-disciplinary science federating approaches from a broader set of communities. Hence, we need to acknowledge that a common understanding of the projects eventually based on reference models must be first established in order to explore synergies better.

Also many other scientific infrastructures share the demand for data-related services and that is one of the reasons why we conducted the EUDAT User Forum [12]. Other architectural approaches and synergies have been discussed and that will lead to more closer collaboration and new EUDAT service cases beyond those introduced in Section "Data replication service use case". The following examples of related work are based on findings obtained from some of the participating projects to the forum while a complete survey is not possible due to page limits.

The European Clinical Research Infrastructure Network (ECRIN) [24] that works on a distributed infrastructure with services supporting multi-national clinical research. The architectural requirements for such an infrastructure in terms of the registration of clinical trials, the repository for clinical trial data (e.g. raw data and anonymized), and the demand of long-term archiving (i.e. 15 years) all bear the potential for federating these ideas with the ones described in this contribution.

The Integrated Carbon Observation System (ICOS) [25] also shares several of the requirements stated as part of our reference model vision and its derived architectural design. The particular interesting part in terms of ICOS would be the integration of this data infrastructure since it collects data of roughly 60 atmospheric and ecosystem measurement sites and about 10 ocean sites. Requirements include near real time data collection and processing and to achieve the data while ensuring the traceability (and metadata provisioning) of the data.

Another project participating in the EUDAT User Forum was the European Life Sciences Infrastructure for Biological Information (ELIXIR) [27]. Also ELIXIR shares several requirements and approaches with EUDAT such as a wide variety of already existing data collections (repositories, gene ontologies, proteomics data, archives of protein sequences, etc.). The planned biomedical e-Infrastructure services of ELIXIR also aim to follow a federated AAI approach integrating many of the aforementioned already existing data repositories and archives.

Finally, also the data infrastructure for chemical safety diXA and its data pipeline [26] is of interest to EUDAT offering also synergies.

7 Conclusions

This paper was motivated by the high-level recommendations, plans, and roadmaps introduced at the beginning of this contribution and that we formulated more clearly as one potential vision named as 'ScienceTube'. This is not a product but rather a vision of federating the various existing approaches to the most possible degree enabling data sharing across scientific communities. We aim to implement some of these high-level visions with approaches underpinned with bottom-up activities that are scientific user-driven in order to ensure that these solutions are really used in production data infrastructures.

We are able to conclude that this is a quite complex undertaking and this contribution only focuses on one particular shared demand of scientific user communities that is the 'data replication service'. The identified reference model entities, their associated architecture work and the arising infrastructure integration issues already give a glimpse of how difficult it is to create an infrastructure where data itself should be the focus and not underlying compute or storage technologies. Although many people and funding organizations claim that every puzzle piece required for a data infrastructure is out there, we still observe crucial requirements for developments (cf. infrastructure integration developments) in order to bring the individual and eventually already used puzzle pieces together. Further work in this context will be to refine the reference model and its associated architecture work around other use cases such as data-staging or simple store for example.

From the process perspective, we can conclude that we created a working system where the *'scientific end-users are in the driving seats'* involving them in architectural decisions whenever possible. We are able to claim that we created a federated data infrastructure part not only 'for' the end-users, but collaboratively 'with' end-users. This is the only chance to get trust and acceptance by user communities to really use the established data services.

The RDA work will be a promising aligned activitiy to increase the role of standards in our reference architecture design process, including emerging standards also from other standardization bodies such as the Organization for the Advancement of Structured Information Standards (OASIS) OData[y]. Also standards used by PaNdata or OpenAire+ [28] are relevant to consider as part of our activities within EUDAT in general and our federated reference model design in particular. It is important to mention that we started this work already by performing the EUDAT User Forum where many other ESFRI projects have been part in order to communicate their demands for a federated pan-European data infrastructure.

Endnotes

[a] http://www.e-irg.eu
[b] http://www.einfrastructure-forum.eu
[c] https://verc.enes.org
[d] http://www.epos-eu.org
[e] http://www.clarin.eu/external
[f] http://www.vph-noe.eu
[g] http://www.pidconsortium.eu
[h] http://www.dkrz.de
[i] http://www.fz-juelich.de/JSC
[j] http://www.csc.fi
[k] http://www.ingv.it
[l] http://www.cineca.it
[m] http://www.sara.nl
[n] http://www.mpi.nl
[o] http://www.rzg.mpg.de
[p] http://www.ucl.ac.uk
[q] http://www.man.poznan.pl
[r] http://www.handle.net/
[s] http://pypi.python.org/pypi/simplejson
[t] http://code.google.com/p/irodspython/
[u] http://www.pidconsortium.eu
[v] http://www.hdfgroup.org/HDF5
[w] http://www.nexusformat.org
[x] http://www.hdfgroup.org/HDF5/PHDF5/
[y] www.odata.org

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All co-authors worked together in establishing the safe replication service based on interviewed requirements from end users while MR and PW carried out the work of formulating and crafting the ScienceTube vision before the EUDAT project started. All authors read and approved the final manuscript.

Acknowledgements

This work is partially funded by the EUDAT project that is co-funded by the EC 7th Framework Programme (Grant Agreement no: 283304).

Author details

[1] Juelich Supercomputing Centre, Juelich, Germany. [2] Max-Planck-Institut für Meteorologie, Hamburg, Germany. [3] Rechenzentrum Garching, Munich, Germany. [4] Stichting Academisch Rekencentrum Amsterdam, Amsterdam, Netherlands. [5] CINECA, Bologna, Italy. [6] INGV, Rome, Italy. [7] Deutsches Klimarechenzentrum, Hamburg, Germany. [8] University College London, London, UK. [9] CSC - IT Center for Science, Espoo Finland, Finland.

Received: 30 November 2012 Accepted: 4 December 2012
Published: 30 January 2013

References

1. Wood J, *et al* (2010) Riding the wave - How Europe can gain from the rising tide of scientific data. European Union, Italy
2. e-IRG Data Management Task Force (2009) e-IRG Report on data management 2009
3. Jones B, Davies D, Lederer H, Aerts P, Eickerman Th, Newhouse S (2010) ESFRI Project requirements for pan-European e-Infrastructure resources and facilities. European e-Infratructure Forum, Amsterdam, Netherlands
4. Mallmann D, von St. ViethB, Riedel M, Rybicki J, Koski K, Lecarpentier D, Wittenburg P (2012) Towards a pan-European collaborative data infrastructure InSide Magazine. Inside Magazine - Innovatives Supercomputing in Deutschland 10: 84–85. http://inside.hlrs.de/htm/Edition_01_12/article_27.html
5. Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters Communications of the ACM 51(1): 107–113. doi:10.1145/1327452.1327492
6. Wittenburg P, *et al* (2010) Workshop on research metadata in context. Nijmegen, Netherlands. http://www.mpi.nl/mdws2010. Accessed July 2012
7. Reetz J, *et al* (2010) Repository and workspace workshop, garching. http://www.mpi.nl/research/research-projects/the-language-archive/events/RWW/program-abstracts. Accessed July 2012

8. MacKenzie CM, Laskey K, McCabe F, Brown PF, Metz R, Hamilton BA (2006) Reference model for service oriented architecture 1.0. Organization for the Advancement of Structured Information Standards: 25–29. Proceedings of the ACL 2010 System Demonstrations
9. Hinrichs E, Hinrichs M, Zastrow Th (2010) WebLicht: web-based LRT services for Germany. In: Proceedings of the 10th ACL system demonstration. Stroudsburg, USA
10. Moore RW, Rajasekar A, Marciano R (2010) Proceedings iRODS user group meeting 2010 - Policy-based data management, sharing, and preservation. CreateSpace. ISBN 1452813426
11. Bicarregui J, Lambert S, Matthews BM, Wilson MD (2011) PaNdata: a European data infrastructure for neutron and proton sources In: Proceedings of e-Science All Hands Meeting 2011 AHM'11, York, UK
12. Riedel M (2012) Secure, simple, sound: data replication on the CDI. http://www.eudat.eu/newsletter. Accessed July 2012
13. Malan R, Bredemeyer D (2011) Functional requirements and use cases. http://www.bredemeyer.com. Accessed July 2012
14. Stewart C, Knepper R, Grimshaw A, Foster I, Bachmann F, Lifka D, Riedel M (2012) XSEDE campus bridging use case descriptions. Year 1. http://www.xsede.org/web/guest/project-documents. Accessed July 2012
15. Alvarez A, Beche A, Furano F, Hellmich M, Keeble O, Rocha R (2012) DPM: future proof storage In: Proceedings of the computing in high energy and nuclear physics. USA, New York
16. Fuhrmann P, Guelzow V (2006) dCache - storage system for the future Proceedings of the Europar 2006, 1106–1113. Dresden, Germany
17. Ball A (2011) Overview of scientific metadata for data publishing, citation, and curation Eleventh International Conference on Dublin Core and Metadata Applications (DC-2011). 2011-09-21 - 2011-09-23, KB, The Hague, The Netherlands. http://opus.bath.ac.uk/26309. Accessed July 2012
18. REPLIX Project. http://tla.mpi.nl/tla-news/the-replix-project. Accessed July 2012
19. Barth W (2008) NAGIOS: System and network monitoring, open source press GmbH. ISBN 1-59327-179-4
20. Crockford D (2006) The application/json Media Type for JavaScript Object Notation (JSON) Internet Engineering Task Force (IETF). RFC4627
21. Fielding RT, Taylor RN (2002) Principled design of the modern Web architecture ACM Transactions on Internet Technology (TOIT). Vol 2(2): 115–150. doi:10.1145/514183.514185
22. World-wide Large Hadron Collider Computing Grid (WLCG). http://wlcg.web.cern.ch. Accessed July 2012
23. Lavoie BF (2004) The open archival information system reference model: introductory guide DPC technology watch report
24. Demotes-Mainard J (2004) Towards a European clinical research infrastructure network: the ECRIN programme Therapie. PubMed - US National Library of Medicine - National Institutes of Health 59(1): 151–153
25. Paris J, Ciais P, Rivier L, Chevallier F, Dolman H, Flaud J, Garrec C, Gerbig C, Grace J, Huertas E, Johannessen T, Jordan A, Levin I, Papale D, Valentini R, Watson A (2012) Integrated carbon observation system EGU general assembly, 2012. Vienna, Austria
26. DIXA Consortium (2012) diXa: A new data infrastructure for chemical safety. http://www.dixa-fp7.eu/home. Accessed July 2012
27. ELIXIR Consortium (2012) European life sciences infrastructure for biological information. http://www.elixir-europe.org/news. Accessed July 2012
28. Manghi P, Manola N, Horstmann W, Peters D (2010) An infrastructure for managing EC funded research output - the OpenAIRE project The Grey. Journal (TGJ): An Int. Journal on Grey Literature 6(1): 31–40

doi:10.1186/1869-0238-4-1
Cite this article as: Riedel *et al.*: **A data infrastructure reference model with applications: towards realization of a ScienceTube vision with a data replication service.** *Journal of Internet Services and Applications* 2013 **4**:1.