Funded under: 01UG1110A until M

Report to the Technical Advisory Board (TAB) DARIAH-DE and CLARIN-D

March 28th, 2013

Project title:
DARIAH-DE – Aufbau von Forschungsinfrastrukturen für die e-Humanities
DARIAH-DE – Digital Research Infrastructure for the Arts and Humanities

Project period:
01.03.2011–28.02.2014

Reporting period:
01.03.2011–31.03.2013

DARIAH-DE

Project coordination: Göttingen State and University Library

Funding: Federal Ministry of Education and Research (BMBF) project reference number 01UG1110A – M

Table of Contents

Introduction

DARIAH-DE supports digitally-enabled research and teaching in the arts and humanities. The project is developing a research infrastructure which will offer tools, core services, and access to research data as well as materials for research and education in the Digital Humanities (DH).

DARIAH-DE is the German national contribution to the European research infrastructure "DARIAH-EU Digital Research Infrastructure for the Arts and Humanities".

Currently, 17 partner institutions from the fields of information technology as well as the arts and humanities are involved in DARIAH-DE, including universities, data centers, disciplinary institutions, one academy, one commercial partner, and one library:

- Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)
- DAASI International GmbH (DAASI)
- Deutsches Archäologisches Institut (DAI)
- Technische Universität Darmstadt – Interdisziplinäre Arbeitsgruppe Digital humanities (Germanistische Computerphilologie / Philosophie / Ubiquitous Knowledge Processing) (TUD)
- Universität Detmold/Paderborn – Musikwissenschaftliches Seminar (DT/PB)
- Göttingen Centre for Digital Humanities (GCDH)
- Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
- Universität zu Köln – Historisch-Kulturwissenschaftliche Informationsverarbeitung (HKI)
- Leibniz-Institut für Europäische Geschichte (IEG)
- Jülich Supercomputing Centre (JSC)
- Karlsruher Institut für Technologie (KIT)
- Otto-Friedrich-Universität Bamberg – Fakultät für Wirtschaftsinformatik und Angewandte Informatik (MInf-BA)
- Max Planck Digital Library (MPDL)
- Rechenzentrum Garching der Max-Planck-Gesellschaft (RZG)
- Salomon Ludwig Steinheim Institut für deutsch-jüdische Studien (STI)
- Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB)
- Universität Würzburg – Institut für deutsche Philologie (UWÜ)

The central mission of DARIAH-DE is to enable the interoperability of tools and research data. Following internationally valid and accepted standards and policies, the project aims at ensuring their long-term preservation and future use.

DARIAH-DE also supports and advises researchers as well as research projects in planning humanities research initiatives within a digital environment.

Effective ways of handling digital resources, concepts, and methods in the Digital Humanities must be introduced into training and instruction for humanities researchers at all educational and career levels. In close consultation with disciplinary communities, existing study and training courses are being coordinated, made more visible, and, if

necessary, developed more fully. Moreover, DARIAH-DE is developing individual qualification modules, such as international workshops for experts dealing with specific themes.

In order to establish digital research in the arts and humanities, it is necessary to enhance knowledge of digital research methods and practices. The use and application of these processes is supported by special tools and services that will be designed, adapted, and made available as a basic infrastructure within the context of DARIAH-DE.

In order to emphasize the added value of Digital Humanities methods, services, and tools, discipline-specific requirements in the form of concrete research questions have been identified. Based on these specifications, individual solutions in the form of "demonstrators" have been developed to demonstrate both specific methods and the overall potential for research in the Digital Humanities.

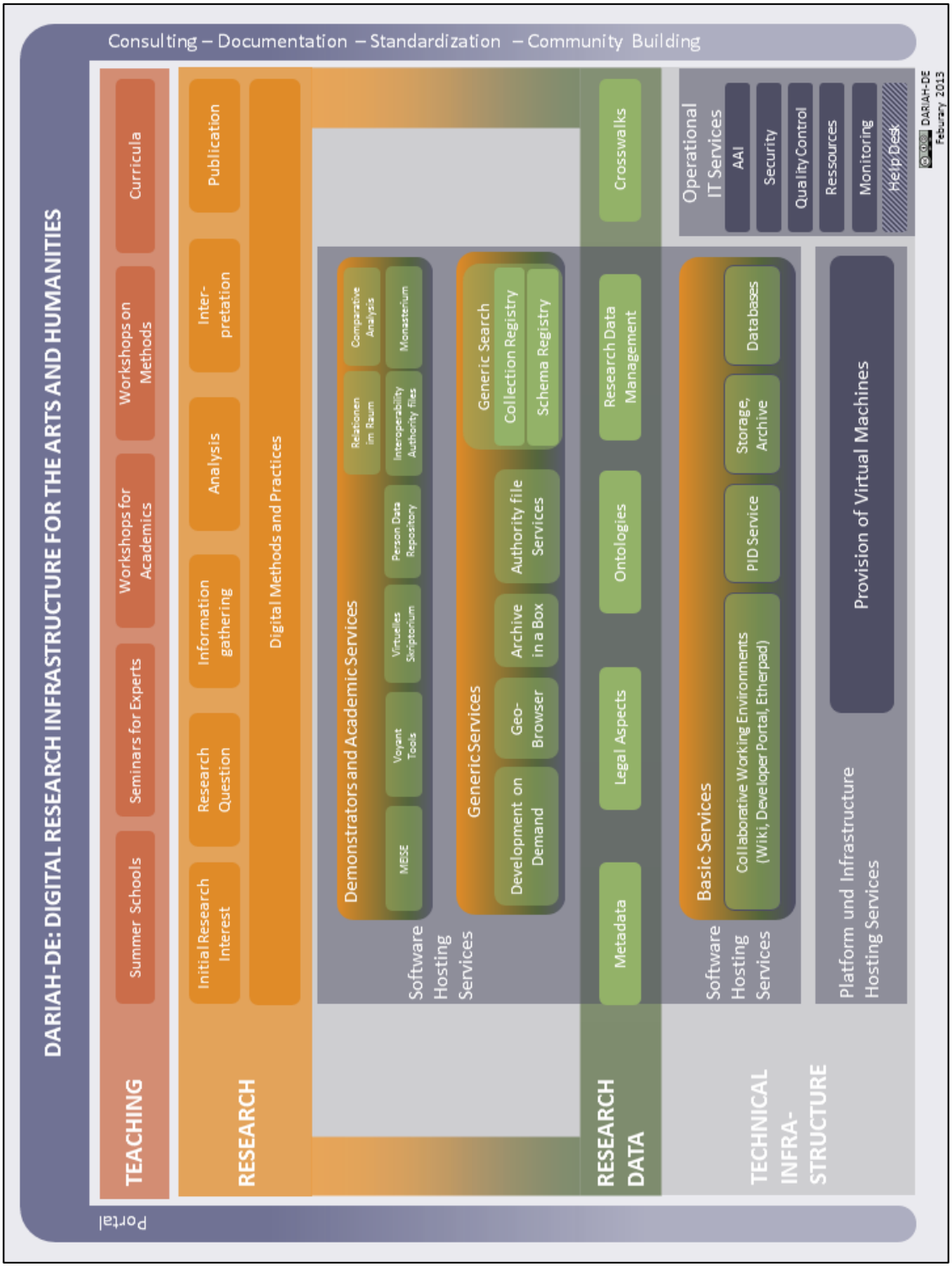
Research data form an essential basis for scientific work during the entire research process, from collecting and analyzing data to publication and the subsequent use by third parties. Unrestricted access is one requirement that is of central importance for working with research data. DARIAH-DE evaluates, discusses, and recommends standards for data, metadata, licensing, and tools as well as for procedures and organizational structures.

This infrastructure will enable researchers to carry out research in an increasingly digital environment, across disciplines and institutions in collaborative ways and towards sustainable results.

Below is an organizational diagram that is reflected in the structure of this report of the specialized technical infrastructure and the interlacing of the four DARIAH-DE core working fields: research, training & education, research data and infrastructure.

Visualization of the DARIAH-DE research infrastructure

Author: DARIAH-DE consortium (Version: February 2013)



Authentication and Authorization Infrastructure (AAI)

Author: Peter Gietz (DAASI)

Due to the very nature of the research process and due to the scarcity of service- and storage resources, authentication and authorization infrastructure (AAI) is a necessity for every technical research infrastructure. Such an AAI also provides for the re-usage of user accounts of user management systems of the research institutions and it thus prevents that every service has to manage its own user base which would lead to very many usernames and passwords that have to be memorized by the user. While simple central authentication services have been established in the social networks based on technologies such as OpenID, more reliable and secure technologies based on the Standard SAML (Security Assertion Markup Language) have been established in the academia. One of the advantages of SAML is the provision of the single sign-on feature (SSO), by which a user only needs to authenticate at one service and will be authenticated for any service in the system. Based on SAML technologies so called federations have been set up by many national research networks in Europe, such as the DFN-AAI in Germany. The eduGain project seeks to create a European-wide Interfederation of such federations, so that every researcher within Europe can authenticate with the log-in name and password of their home organization, the so called identity provider.

DARIAH-DE has developed a concept for deploying SAML based technologies, especially the Open Source Software Shibboleth to smoothly integrate in the existing (inter)federations so that all DARIAH services (software services as well as storage services) could be used by authenticated users. For the implementation the project faced several challenges:

- Not every researcher belongs to an organization that is part of a federation
- Not every Identity Provider is willing to participate in eduGain
- Not every Identity Provider is willing to provide (personal) data to the DARIAH service provider
- The campus identity provider have no information about the privileges of a user within the DARIAH virtual organisations, e.g. memberships in projects like CENDARI
- Shibboleth only supports WebSSO, i.e. browser based services, but not autonomous web services
- Some projects want to re-use OpenID based identities of social networks

The now productive DARIAH AAI can already cope with most of these challenges. For researchers who cannot authenticate via the credentials of a home organization can get an account in the central DARIAH user management based on LDAP technology. The privileges of these users, as well as of the users authenticating via their home organisation can be managed by an intuitive web based administration tool, which provides for distributed management of privilege-groups. (see figure 1) A DARIAH service provider can retrieve data from the identity provider as well as from the DARIAH privilege administration. Users whose identity provider hesitate to send data to the DARIAH service provider, can provide DARIAH their data via a registration form. A number of DARIAH service have been integrated into the DARIAH-AAI.

Web service support has been successfully tested by using OAuth2 technologies. OpenID integration has also been tested successfully.

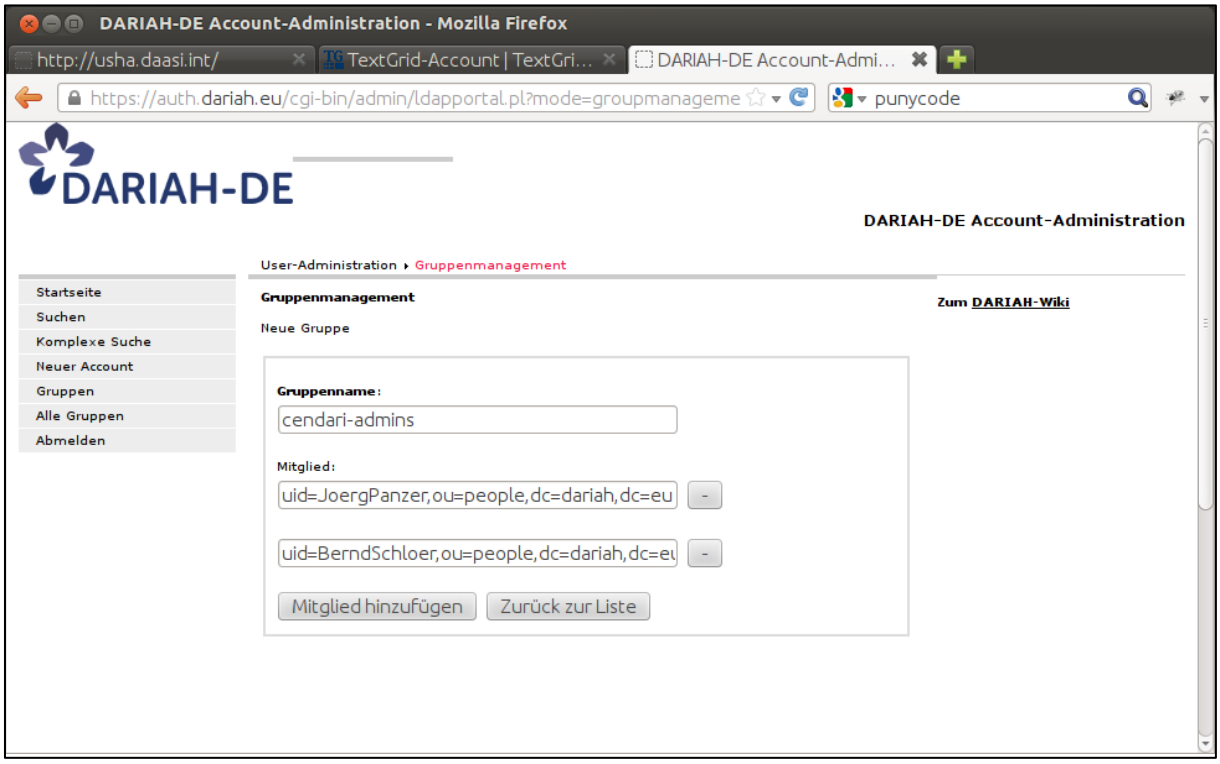


Figure 1: DARIAH user administration

Besides the technical developments, DARIAH was also active in a number of organizational AAI activities:

- DARIAH-EU AAI workshop
  - Agreement on a Role model for a DARIAH-EU privilege management
- Work with CLARIN:
  - Discussion of interoperability
  - Call for action together with CLARIN to stress the importance of releasing personal attributes to trustworthy academic SPs
  - First concepts for a and a common eHumanities federation, that could become part of the eduGain interfederation
- Active in European initiatives:
  - Federated Identity Management workshops
  - EUDAT AAI workshops
  - TERENA AAA study for the European Commission

Bit Preservation

Author: Rainer Stotzka (KIT)

The sustainable management of large amounts of research data is gaining importance for research projects all over the world. The DARIAH Bit Preservation, as a part of an archiving system for the arts and humanities, allows for a high performance, sustainable, and distributed storage of research data as the basis of virtual research environments. A great challenge in designing such a service is to provide a standardized, consistent yet easy-to-use API for accessing the data that remains stable even if backend technology changes over time. We developed a RESTful API for the DARIAH Bit Preservation which includes an administrative extension, and which is secured by an Authentication and Authorization Infrastructure (AAI) based on SAML. The implementation of the API offers distributed access by usage of the HTTP protocol and is able to handle a high number of files. Data transfer rates of up to 45 MB/s were achieved for uploading large files in the local network.

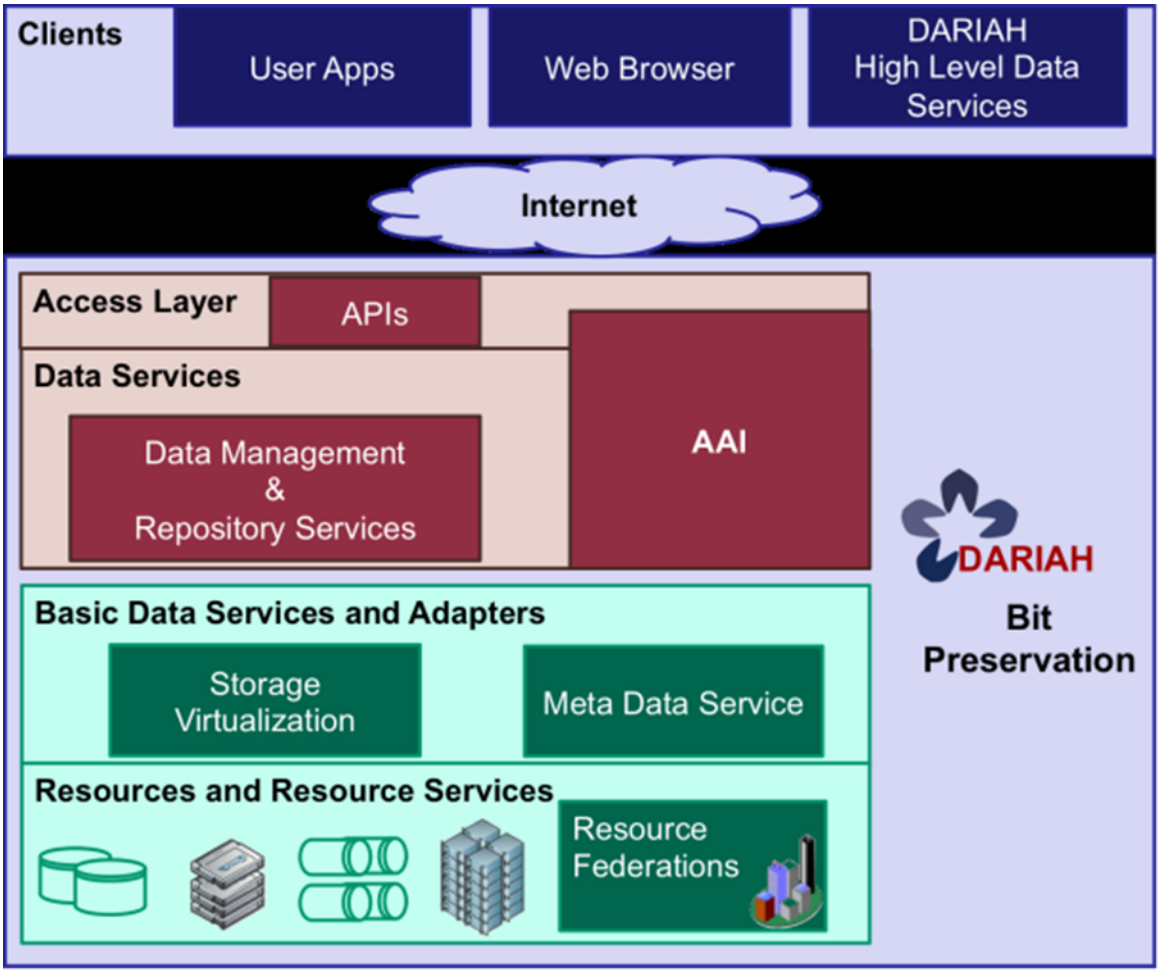


Figure 2: Architectural overview of the DARIAH Bit Preservation service

Confluence Wiki

Author: Stefan Schmunk (SUB)

DARIAH-DE uses the Confluence Wiki system from Atlassian as a central research management system for internal project management, more specifically for tasks such as the creation of subject-specific content and the collaborative preparation of reports and publications. Additionally, the Confluence Wiki is made available to external DH projects on request. The production of texts is a central element in contributing to knowledge in the arts and humanities and is therefore of significant importance as a fundamental instrument for the Digital Humanities.

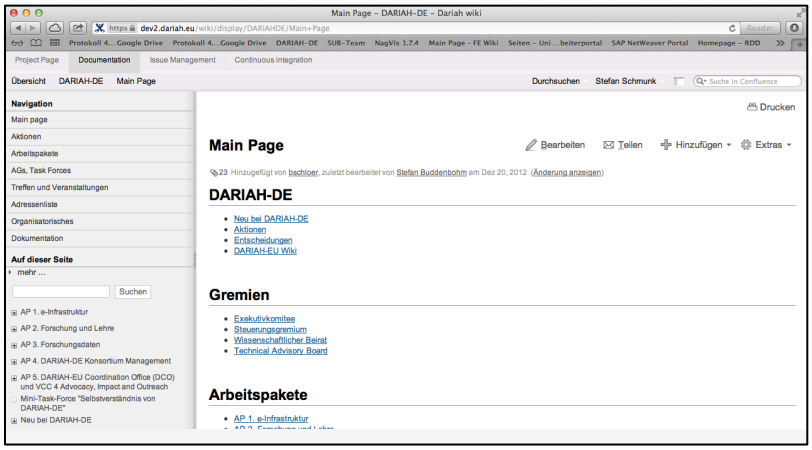


Figure 3: Main page of the DARIAH-De Confluence space

From a technical and administrative perspective, Confluence also offers an easy handle (including JIRA integration) via the provision of spaces in the installed DARIAH-DE instance and is already shibbolethized. Currently the DARIAH-DE Confluence Wiki is already being used by 18 external DH projects with a total of more than 350 users. The following screenshot shows a visualization of typical use patterns.

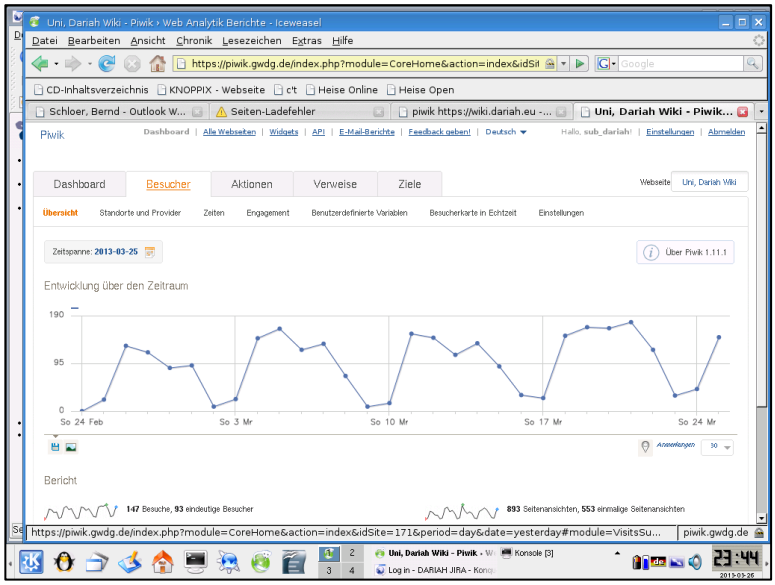


Figure 4: Typical use pattern of the DARIAH-DE space

DARIAH-DE Portal

Author: Martina Kerzel (SUB)

The DARIAH-DE portal serves as the central access to the infrastructure of DARIAH-DE. The user accesses scientific and technical services and data. Also research and education in the Digital Humanities are supplied with various materials.

A one-stop shop is envisioned that enables the user to select in a time- and cost-efficient and location-independent way IT components and other services.

The portal will be implemented with Liferay, an open-source solution, that displays data, applications and information within a consistent user interface through a conventional web browser. Liferay is based on Java and offers in addition to a server-oriented content management system (CMS) the opportunity to enhance the portal through the integration of Liferay Portlets or self-developed portlets.[1] This integration has already been successful with the Zotero-based DARIAH bibliography.

This modular approach brings the benefit of being able to integrate gradually various applications of the DH community into the portal. 1. Tier: cursory integration via linking, 2. Tier: Embedding of the service through iFrame and 3. Tier: deep integration through a self-developed portlet.

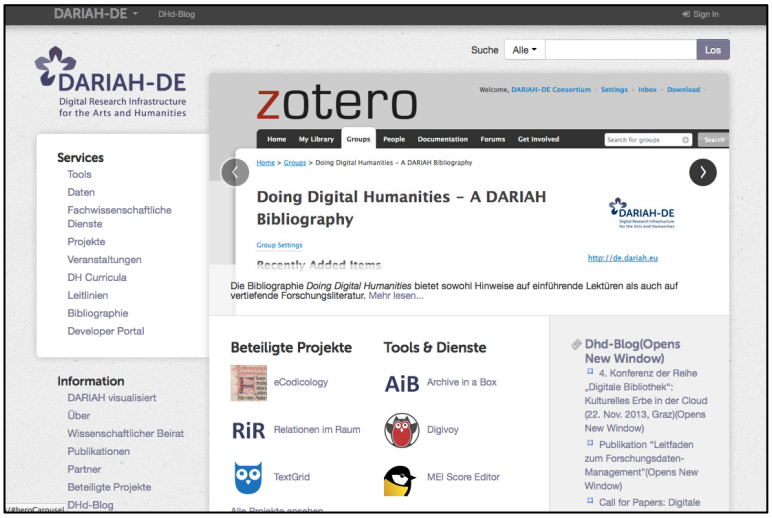


Figure 5: The DARIAH-DE portal at an early stage

For the creation of portlets the developer can access not only the technical infrastructure for the programming of applications (developer portal) but also the guidance and support of the experienced DARIAH colleagues.

The cross-linking of services, data and applications is an important issue. Through modeling of data to methods, tools, services, projects and research questions the community-based initiatives gain a new level of semantics, which can be of use not only for DH newcomers but also for the experienced DH researchers.

[1] Vgl. http://www.liferay.com/de bzw. http://de.wikipedia.org/w/index.php?title=Liferay_Portal&oldid=114842110 (24.03.2013).

Databases

Author: Benedikt von St. Vieth (JSC), Jedrzej Rybicki (JSC), Ulrich Schwardmann (GWDG)

Motivation

The efficient storage and sustainable deployment of information is of essential importance:

- the growing amount of data implies challenges on the extraction and accessibility of data.
- for the participation of third parties a simple and useful access is necessary.
- Relational and Non-SQL databases are established instruments to describe relationships between data, to display them and to allow a most easy and performant access.

Problem

At least the following points have to be considered for the operation of a database system (DBS, Datenbanksystem):

- Hardware resources (operation, maintenance).
- Operating systems (installation, maintenance).
- Database system (installation, maintenance, security, backup).

At all three levels critical situations can appear, that endanger the security of data. Be it a blackout, malicious intents of third parties or errors in the software.

Solution

Database systems are administered and maintained by datacenters (Rechenzentrum) and allocate for usage:

- The demanding party receives host/ port and user account/ password for direct access to the database.
- The maintenance of the server resides with the datacenter.
- The backup of the database is ensured.

➔ Clearly structured responsibilities.

➔ No administrative overhead for the customers.

The offer includes at first PostgreSQL, but can be extended continually and is dependent on the datacenter.

Data Registries and Generic Search Framework

Authors: Tobias Gradl (MInf-BA), Christof Plutte (BBAW)

Data Registries and Generic Search Framework – Integrating Heterogeneous Research Data Collections

Access to research data is one of the most important aspects in digital research but often very difficult on a larger scale due to the diversity of digital data sources and the heterogeneity of the information they contain.

The DARIAH federation infrastructure aims to address these problems by building a comprehensive framework of registries and generic services: The *Collection Registry* serves as an online registry to hold and publish descriptions of research data collections and their machine-readable access points (e.g. OAI-PMH). Metadata schemas used within the collections are registered and semantically enriched in the *Schema Registry*, which also allows scholars to define disciplinary- or collection-specific mappings and transformation rules to other schemas (e.g. DC, MODS). The registered collection can then be harvested by the *Generic Search* via the registered access points using the schema and crosswalk information to map data and to facilitate a faceted, federated search that is dynamically tailored to relevant collections and data structures - all from a single user-oriented portal.

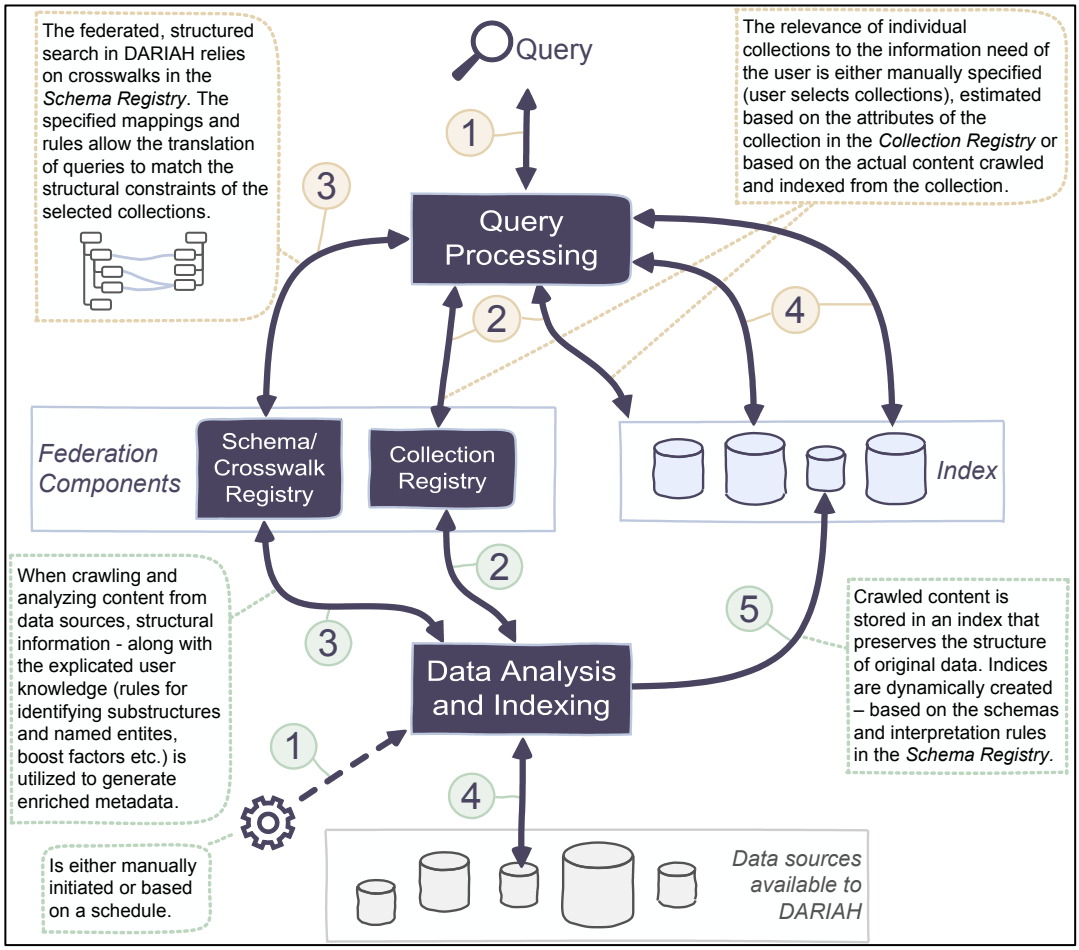


Figure 6: The DARIAH federation infrastructure

The user interface of the *DARIAH Generic Search* provides capabilities for a quick full-text search as well as extended options allowing the specification of facets on both content- and collection-level.

Aside from the typical content search, users can choose to retrieve collections relevant to their information need and navigate into the results.

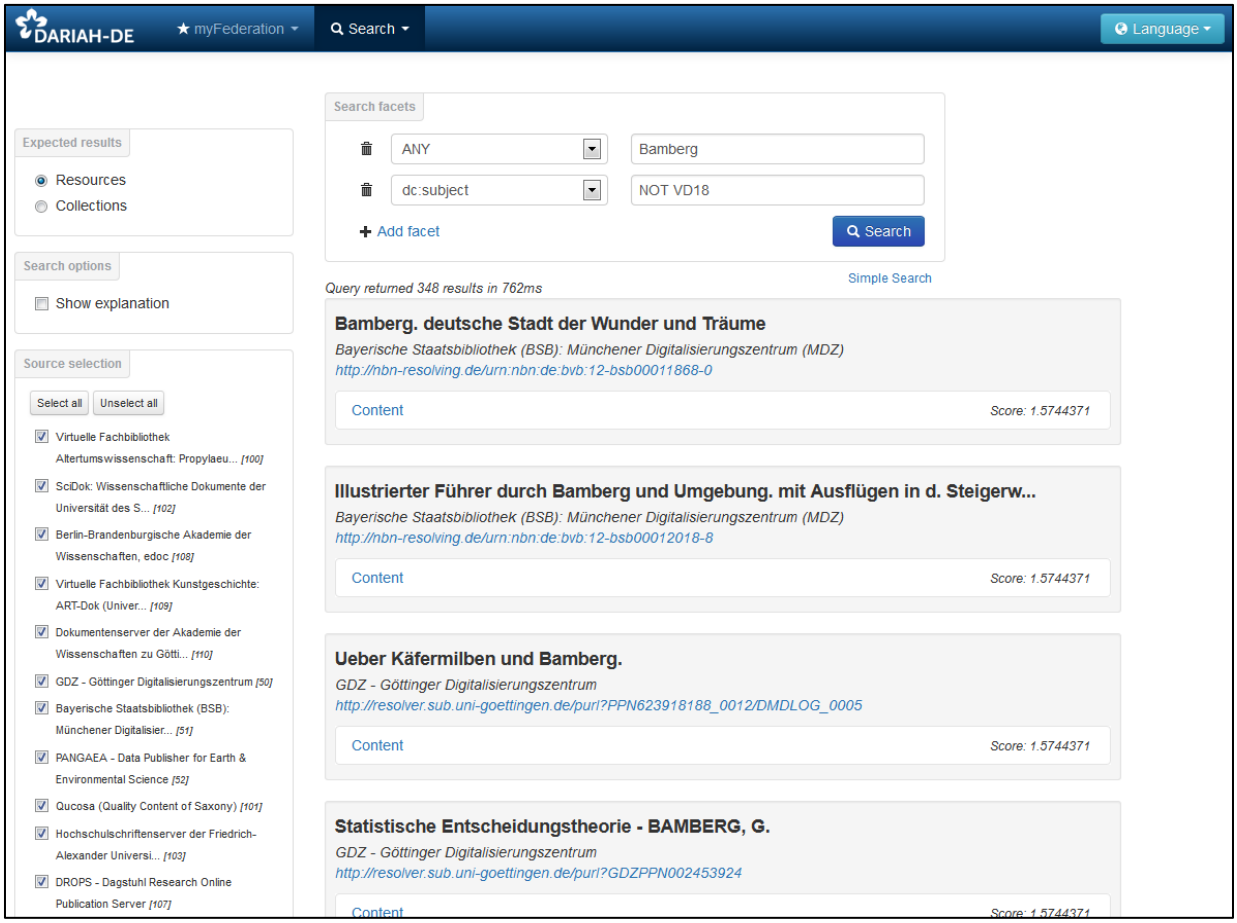


Figure 7: Search results of the DARIAH Generic Search

Developer Portal

Author: Bastien Saquet (MPDL)

The DARIAH developer portal has been developed during the DARIAH development phase and has been immediately used as a central component to achieve the other tasks of the DARIAH project. Its goal is to support software development within the DARIAH community. This is done by:

- Making software development tools available.
- Helping collaboration in distributed teams and between the different actors in software development (developers, scientists, librarians, etc.).
- Improving software quality with guidelines and services of the DARIAH infrastructure.

The list of the supported tools and services can be found in the figure below.

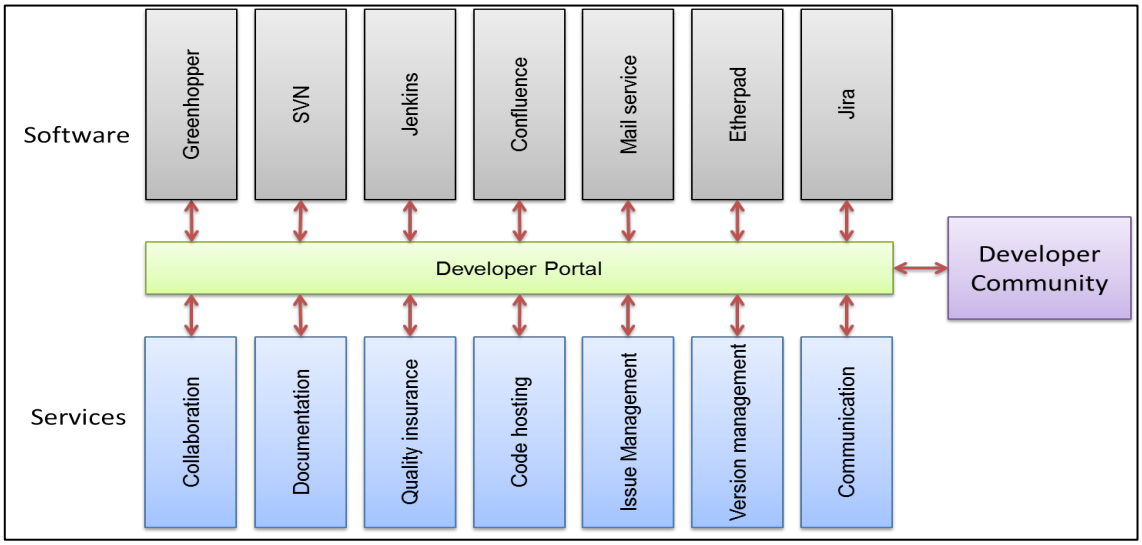


Figure 8: Developer Portal: Tools and Services

All this tools and services are available for productive usage to all DARIAH partners and associated projects like EHRI,[2] CENDARI,[3] etc.

[2] EHRI: European Holocaust Research Infrastructure: http://www.ehri-project.eu/

[3] CENDARI: Collaborative European Archive Infrastructure: http://www.cendari.eu/

Hosting Environment

Authors: Jedrzej Rybicki (JSC), Tibor Kálmán (GWDG)

Technical Factsheet

Synopsis

This document provides some information to users interested in using the DARIAH hosting facilities. It gives a general overview of its features, in particular:

- what kind of resources and hosting services DARIAH is able to provide-
- which requirements need to be fulfilled to get access to the DARIAH Hosting Environment
- description of a process of application for VM or Service in the hosting environment

Introduction

Digital methods become an important item in the everyday arsenal of researchers from the arts and humanities. Hence, the provision of a means for accessing compute resources is an important goal of the DARIAH-DE project. Resources offered within the DARIAH e-Infrastructure can be primarily used for either hosting services for the humanities or performing computations. In this regard DARIAH-DE offers state-of-the-art solutions resulting from the long-year experiences of the computer and data centers participating in the project:

- Juelich Supercomputing Center (JSC)
- Rechenzentrum Garching (RZG)
- Gesellschaft für wissenschaftliche Datenverarbeitung mbh Göttingen (GWDG)
- Karlsruhe Institute of Technology (KIT)

The DARIAH hosting offer aims to strike a balance between the maximal flexibility for advanced users and seamless access to typical scientific services for less experienced users. Therefore, different kinds of hosting services are offered. Maximal flexibility is covered by the VM hosting option. In this case, users get access to a virtual machine running on the resources provided by the data centers and can install and configure the services inside as they prefer to, or use the VM for data processing. For less advanced users demanding an instance of existing, well-established services, the option of assisted hosting is more suitable. In this scenario a request for an existing, popular service is made (e.g. an instance of Fedora Commons), and the installation and initial configuration is done by the experts at a given computer center. The requester is then provided with an URL to an up-and-running instance. Here is an overview of the common features of these two options:

- access to virtual machines with required operating system
- for testing (on-demand: use and throw away) or production purposes
- high-end resources (compute, storage, network)

- basic system configuration provided (firewall, domain, security updates)
- computer center services (monitoring, backup if needed)
- integration with other DARIAH services (distributed monitoring, AAI, bitstream preservation)

There is a growing repository of well-known DARIAH services and service containers (Apache, Tomcat) which our computer centers are able to pre-configure automatically. In these cases the user only needs to deploy the service. Such applications will be called embedded applications (as opposed to stand-alone applications) throughout the rest of this document.

It should be stressed at this point that the DARIAH Hosting Environment is an object of research itself. Computer scientists are seeking the best ways of providing resources and assisting the users. It is still in the pilot phase and thus is undergoing a number of changes; nevertheless the provided services are of high quality and can be used for both testing purposes and production-ready hosting.

Requirements

Before a service for humanities can be transferred into the DARIAH Hosting Environment, it has to be justified that the service is important for DARIAH and its users. Depending on the type of service (short time, testing purposes or long term usage), this clarification has to be done by different boards. Furthermore, every hosted service must fulfill some general and site-specific technical requirements. The users of the DARIAH hosting infrastructure and the services hosted within must accept the DARIAH Terms of Use (reference).

General technical requirements

Each service request should be justified and sustainability must be proven by the requester. This is something that only the requester can do and that the DARIAH board should evaluate, e.g., by proving that the application is in fact of interest for DARIAH users, providing the expected number of users, and load and usage/maintenance time horizon. Furthermore:

- for each service a short description of its functionality needs to be provided, including the requested version, a link to development history (releases) and a future roadmap
- the services should be open-sourced and have a permissive license (Apache, BSD, etc), and the data stored in the machines (i.e. content managed by the services) must also have a permissive license.
- a set of start/stop scripts with documentation must be available, and a VM should survive a restart without any problems.
- the name and email address of a person responsible for the service needs to be provided, and this person should be able to answer questions regarding configuration, etc.
- documentation of the service should be available e.g., on the internal DARIAH wiki

An exact specification of the runtime environment must be provided:

- for stand-alone applications, only open-source operating system will be supported (Debian, Ubuntu LTS, SLES)
- an estimation of resources needed by the application is needed (CPU, RAM, HDD size). Those estimations will be taken into account but there is no guarantee that all the resources will be provided from the day one; instead, dynamic reallocation of resources to account for increasing load will be applied
- estimated load should be provided (the expected number of users)

A specification of the runtime environment must include requested dependencies:

- for the service dependencies, the same criteria hold with regard to open-source, documentation and licenses,
- for system libraries dependencies: current, supported, and secure versions available in generic open-source operating systems must be used
- for embedded applications, current versions of application servers (Tomcat, Apache) will be supported
- for services running in application servers, necessary changes in the general config files should be provided (if applicable)

Security requirements

All services should fulfill basic security requirements (e.g. passwords must not be stored nor exchanged in open text, etc). After a vulnerability is detected in an application, updates have to be provided and applied. Depending on the type of vulnerability, it is possible that the service will not be available until a fix is provided.
In case of the VM, the hosting person responsible for the machine should take care of operating system patches when they have to be applied.

Network requirements

For stand-alone services, exact specification of port (ranges) needs to be provided (necessary to configure firewall rules). Embedded services hosted within Apache or Tomcat will get the port as configured by the computer center during the server installation. The network traffic of the applications will be monitored. Typically only passive applications (responding to external requests and not initiating own network connections) will be supported.

Site-specific requirements

There are differences between sites offering hosting facilities, so that some services can be offered only by particular computer centers. In particular the assisted hosting option is only offered by the Juelich Supercomputing Center. Although all involved parties try to keep most of the technical details hidden and make the process of gaining access to resources as opaque as possible, there are a few differences in their offerings which should be accounted for.

Juelich Supercomputing Center (JSC)

The operating system can be selected from Debian, Ubuntu LTS, and SuSE Linux Enterprise Server. Due to some internal regulations and external obligations towards the DFN, Juelich Supercomputing Center can only host services offered for “limited”

group of scientific users (project partners). A simple way to guarantee that is to offer services secured by DARIAH AAI or filtered DFN AAI. All services accessible to the public must undergo a security audit before respective ports are open in the firewall. For computation jobs and short-living instances, experimental access to the OpenStack-based private cloud can be provided.

Rechenzentrum Garching (RZG)[4]

The RZG VM hosting cluster is based on a XEN virtualization environment running in a blade center. Entitled admins may request VMs for deploying their specific services. The preferred operating system is Suse Linux (SLES). VMs from different projects are protected against network isolation and IP spoofing with Iptables rules on the host. The purpose of the hosting environment is to provide VMs for test environments as well as for production services. Due to security and administration reasons, administration of the VMs requires the application for an RZG account. For safe data storage, VMs can connect using data management software to our HPSS system.

Gesellschaft für wissenschaftliche Datenverarbeitung mbh Göttingen (GWDG)

The DARIAH Hosting Environment at the GWDG contains general hosting services offered to research infrastructures and special hosting services offered to DARIAH. The general hosting portfolio of the GWDG is currently being redesigned and the general hosting services will be based on a new cloud environment of the GWDG. The general services include templated virtual machines with current Ubuntu and SuSE Linux Enterprise Server (SLES) operating systems, monitoring of virtual machines, backup of virtual machines, and setting up HPC clusters on demand. Access is provided to GWDG users and project partners.

For the DARIAH project special hosting services are also provided. These services include managed hosting of virtual machines on DARIAH hardware, and running virtual machine images prepared not by the GWDG (external images). The services are provided within the preparation phase of the DARIAH-DE project, but might also be extended for later phases.

Karlsruhe Institute of Technology (KIT)

KIT does not offer compute resources at this time. They might be, however, incorporated in the later phases of the project.

Process of requesting hosting and compute resources

The requester specifies its request by filling a web form (reference) and submitting such a request. Request from the web form are posted on an internal DARIAH mailing list. Representatives of all computer centers (hosting officers) should be included in this list. The monitoring of the request processing is done via Jira tickets. One of the hosting officers is selected to be hosting manager. He is responsible for creating tickets for each incoming request. The request should also be attributed to a hosting officer. This

[4] German designations appear as long as they are common in the project context and make the attribution for the reader more easy.

attribution should be done in such a way that either the best suited computer center (e.g. all requests for assisted hosting will be attributed to an officer from Juelich) or the center with most available resources will be selected.

Before the preparations for hosting are started, it must be proven that the requested Service is needed by DARIAH. Here we need help from outside. The DARIAH Board should be on the mailing list too. Once the hosting manager creates a ticket, the ticket should be assigned to the DARIAH board manager and the computer centers have to wait until the ticket returns to them with a „go“ or „no go“ commitment.

A requesting user should be able to view and comment on the request ticket. For requests for VMs, a SSH public key should be provided to the responsible center.

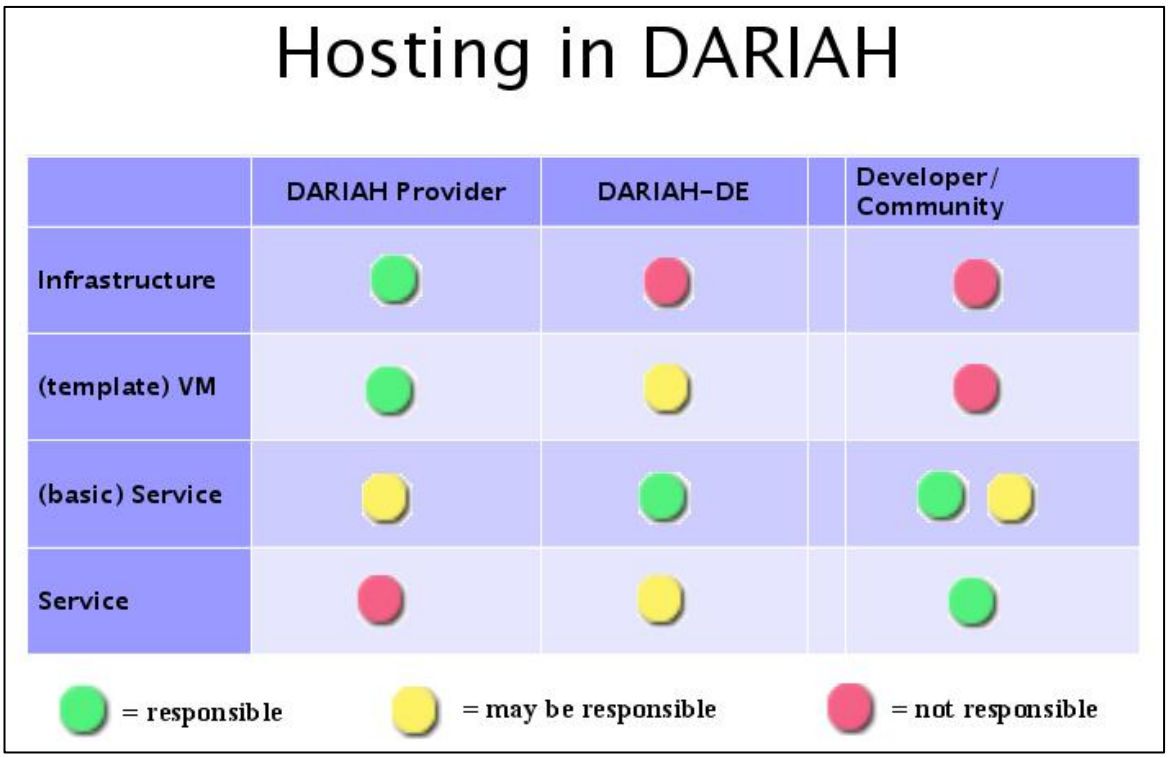


Figure 9: Responsibilities at different DARIAH infrastructure levels

Monitoring

Author: Benedikt von St. Vieth (JSC)

The DARIAH-DE infrastructure and service monitoring is a system based on well-known building blocks and it integrates the existing monitoring systems of the different participating computing centres.

Nagios is used to check the availability of hosts and services and to collect the information of the distributed monitoring systems. NagVis is used to visualize the data collected by Nagios to create a view that shows service dependencies and to create an infrastructure overview.

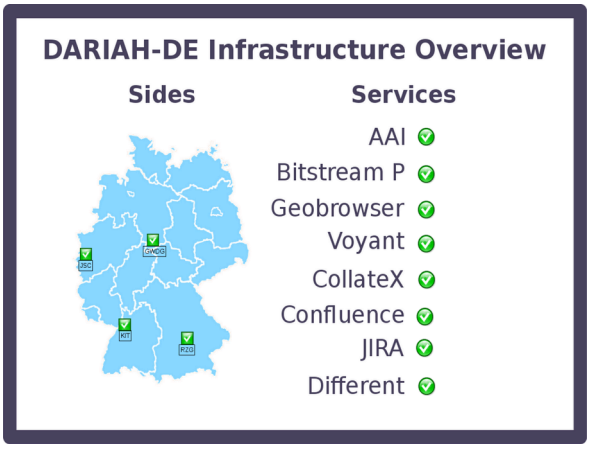


Figure 10: Infrastructure overview

To monitor the availability of all pieces of the infrastructure, we use different methods:

- Direct checks: Especially DARIAH demonstrators are available for everybody, therefore we are able to start function checks at the central monitoring server that use http requests or other protocols to make sure the service is running and working properly. These function checks can use existing Nagios Probes but some of them are proprietary scripts/programs that service developers provide.
- NSCA (Nagios Service Check Acceptor): For computing centres that already have running monitoring systems the central monitoring server accepts their results. At the central monitoring server, a daemon is running that accepts the incoming results and redirect it to Nagios.
- NRPE (Nagios Remote Plugin Executor): Whenever a computing centre does not want to monitor parts of the DARIAH-DE infrastructure we are able to use a installed NRPE at the target machine to execute remote checks.

Because there is no central registry of hosts/services we scan the Nagios-log for incoming check-results that do not fit to existing hosts and services and create a configuration based on these scans. For demonstrators and other central services, we use information provided by the administrator of the service to create service checks and to get a contact that we can inform in case of problems.

PID-Service

Author: Daniel Kurzawe (GWDG), Tibor Kálmán (GWDG)

Not only the data, but also sustainable references are important to ensure the sustainability of digital objects. Therefore, persistent identifiers (PIDs) are used to build a stable layer between the reference to an object (URI, etc.) and the digital object itself. In DARIAH-DE, PIDs are used in different contexts, from data citation to referencing digital objects in long term archiving systems. One of the most common use cases is the identification of digital objects on the preservation layers in several data centres. Another important use case is keeping the digital content and its metadata together.

The PID resolution and PID management is not built up by DARIAH-DE, but DARIAH-DE relies on the PID-Service of the European Persistent Identifier Consortium (EPIC)[5]. The PID development within DARIAH-DE is coordinated with EPIC. Therefore DARIAH-DE hosts a PID test environment[6].

DARIAH-DE also identified some issues related to current PIDs. The possibility of converting an EPIC PID to another type of identifier (especially DOI PIDs) is an emerging issue, as well as referencing subsets of digital objects and giving more possibilities for the granularity. In DARIAH-DE, PID related issues are handled in the infrastructure work package.

[5] http://www.pidconsortium.eu/

[6] http://dariah-vm07.gwdg.de/

Quality assurance

Author: Rainer Stotzka (KIT)

Quality assurance in DARIAH is concerned with the quality of services offered to scholars in the academic disciplines. All basic services provided by the computing centers are monitored to guarantee continuous access. Higher level software services produced by the partners or by external groups are already closely guided within the development process.

In DARIAH-DE a "Service Life Cycle" state diagram has been developed in which the development steps as well as the transition conditions to the next steps are clearly defined. A team of scientific and technological mentors escorts the service and helps with advice to assure reasonable exploitation of existing DARIAH services and to check the usability and value of the service for the humanities scholars.

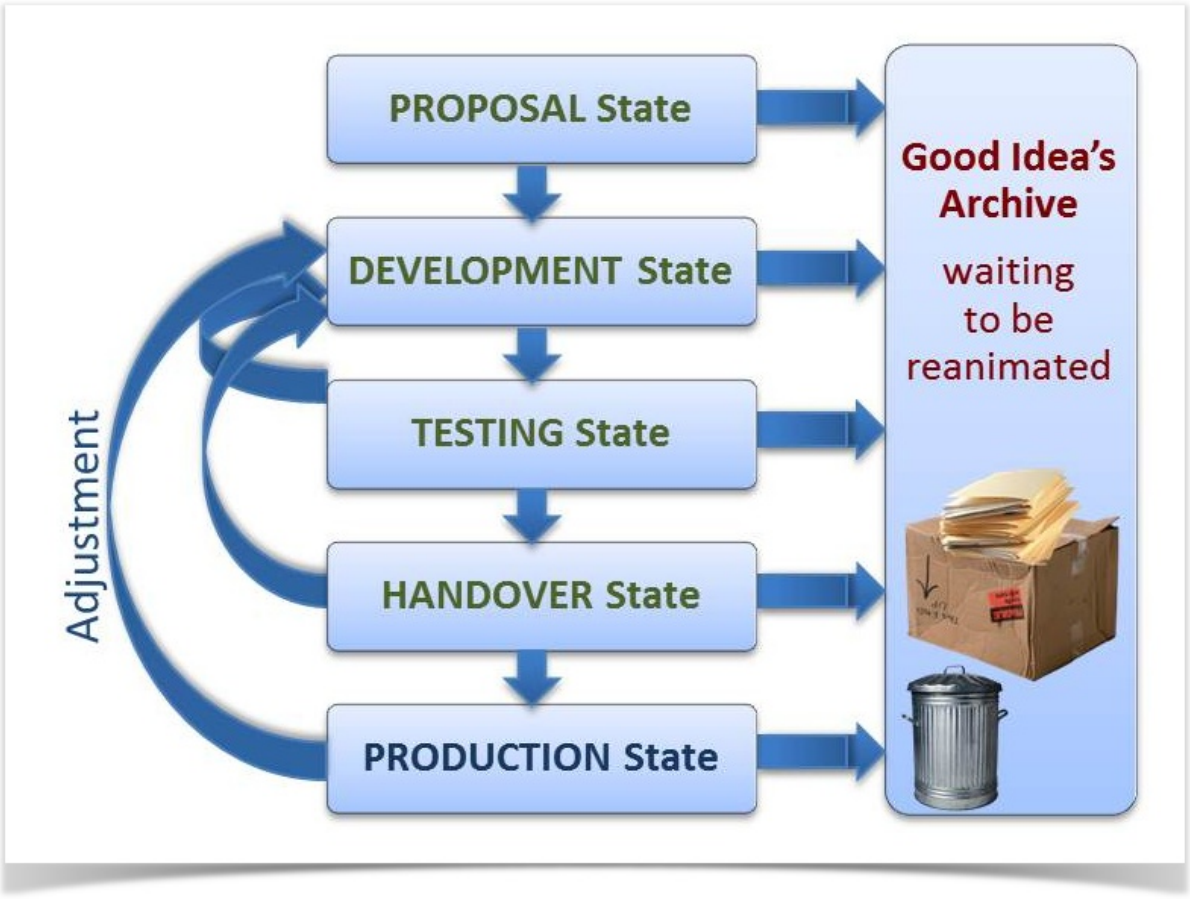


Figure 11: The DARIAH Service Life Cycle

Security

Author: Johannes Reetz (RZG)

The term security refers to confidentiality, integrity and availability of data, systems and services provided by or related to DARIAH-DE. Therefore, also critical functions, such as access controls and backup procedures, of sites participating in DARIAH-DE can be affected. Security is about how data, services and systems provided by DARIAH-DE are protected against risks according to best information security practices in accordance with national and EU laws. Security related assumptions of the DARIAH-DE service providers, users and other stakeholders are and will have to be agreed on common efforts to secure services and data related to DARIAH-DE during the course of the construction phase.

The computing centres in DARIAH-DE are engaged in national and international collaborations and e-infrastructures, and are operated in compliance to the best practices and guidelines for information security from sources such as the Trust Framework for Security Collaboration among Infrastructures (SCI), the security policy of EUDAT, policies of the European Grid Initiative or PRACE. Each of these centres has a security officer assigned who is responsible for incident coordination, risk assessments, continuous operational security and for developing and adapting security controls and guidelines at his site.

As one security-related task, the contact addresses of these site security officers which are of the form dariah-security@<site domain name> have been published on the DARIAH-DE wiki, together with further contact information which allow to direct any security notification to the computing centres in case of security incidents or discovered vulnerabilities.

The site security officers are responsible for ensuring secure systems configurations, vulnerability handling, incident handling, anomaly detection, monitoring and logging, providing security information within and between sites. This comprises also a secure authentication of machines and services using ssl-certificates from widely accepted Certification Authorities (DARIAH-DE uses ssl certificates of the DFN PKI).

As another security-related task of the computing centres, the common terms of use (ToU) have been specified during the reporting period. The ToU applies to all users of services provided by DARIAH-DE, and which makes them understand their security and privacy related rights and obligations.

Storage Architecture

Authors: Rainer Stotzka (KIT), Francesca Rindone (KIT), Johannes Reetz (RZG)

Storage Infrastructure: Scope

Establish ***open, federated, distributed, and dynamic*** infrastructure services to ensure data preservation and interoperability across DARIAH (and more)

"Low-barrier" entrance for new scientific disciplines and infrastructure providers to foster the attractiveness of scientific networks

Infrastructure and software development:

- Support the creation and enhancement of data infrastructures for the management of research data
- Establish common (standardized) storage and archive interfaces to guarantee technological sustainability for high level services
- Open idea: Development of a Preservation-in-a-Box (PiB) package

Archive Service

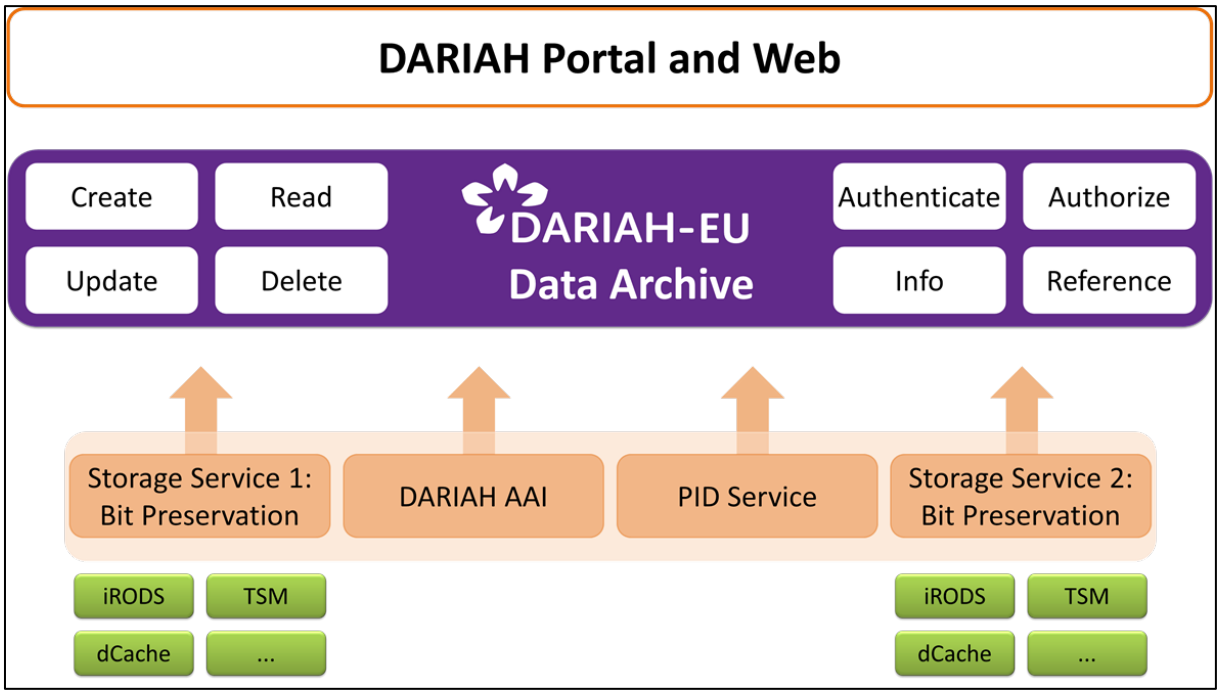


Figure 12: The DARIAH Archive Service

Potential Integration

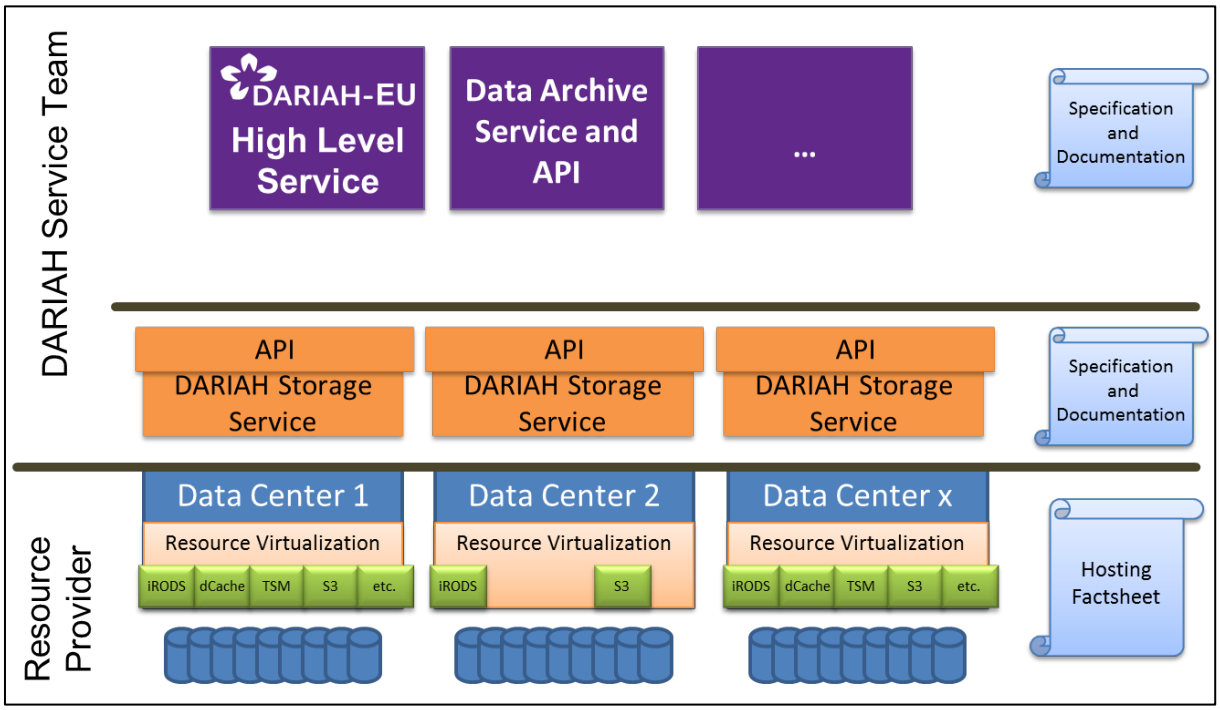


Figure 13: Potential Integration

At the moment several projects (March 2013: 17 projects) coming from different disciplines and different funding use the DARIAH-DE storage service, e.g. the Steinheim Institut für deutsch-jüdische Studien to archive Kalonymos, the BMBF-project Relationen im Raum etc.

dawa – dATA aRCHIVE wEB aPPLICATION

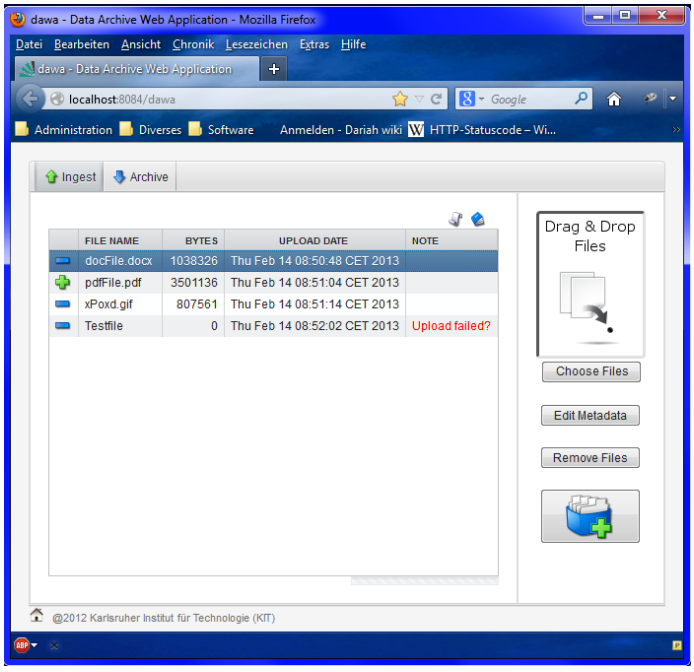


Figure 14: dawa – Data Archive Web Application

Data Life Cycle

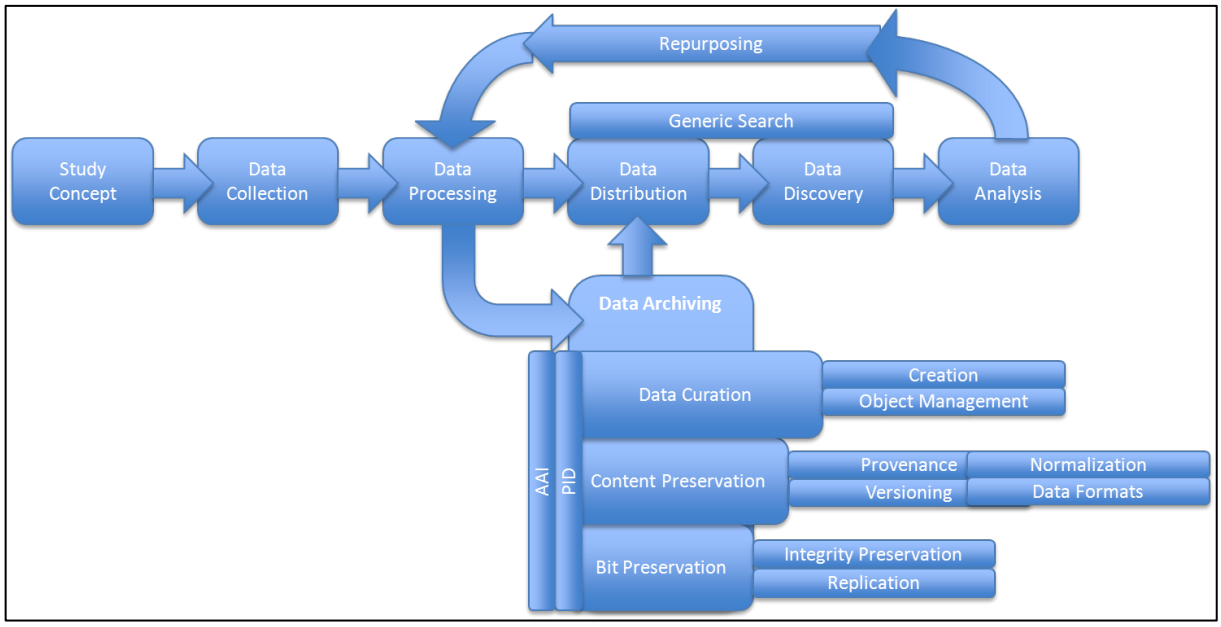


Figure 15: data Life Cycle

Source: DDI Structural Reform Group. "DDI Version 3.0 Conceptual Model." DDI Alliance. 2004. Accessed on 11 August 2008. <http://www.icpsr.umich.edu/DDI/committee-info/Concept-Model-WD.pdf>, adapted by Daniel Kurzawe and Rainer Stotzka, 13. Feb. 2013

Storage Federation Architecture

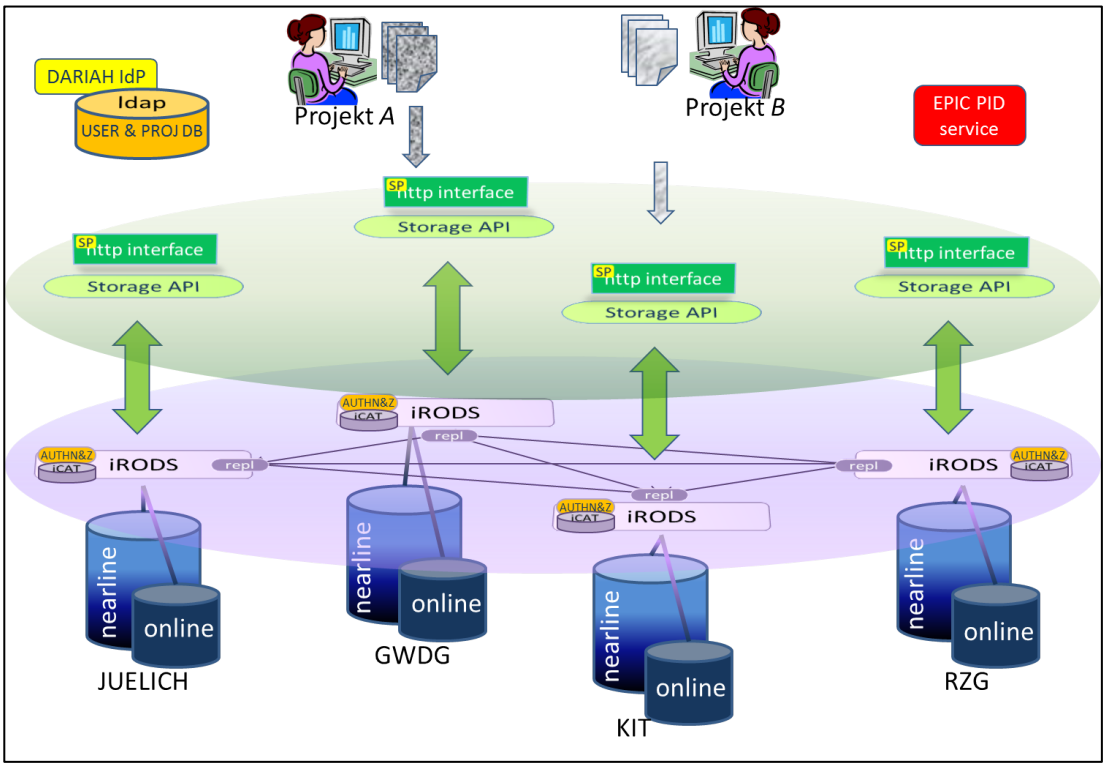


Figure 16: Storage Federation Architecture

Terms of Use

Author: Ulrich Schwardmann (GWDG)

DARIAH has the claim to open its services to other projects and communities of the digital humanities. This implies that it is necessary to integrate groups of users, which go far beyond the institutions of the DARIAH consortium itself. Therefore it is necessary to have a reliable legal relation between the users of DARIAH services and the service providers with their partially different existing terms of use, which fulfills a sufficient set of their requirements.

The service centers agreed therefore to a common Terms of Use, that has to be signed by each user of the DARIAH services, which goes back to a similar approach in the context of D-Grid and was adopted to the special needs in this context. A legal contract has always to be agreed between legal entities in order to be valid. As long as no common legal entity of Dariah (ERIC) exists, one current problem therefore is, that the Terms of Use usually must be agreed between all participants of the service infrastructure, users as well as providers. To avoid the overhead for users to sign a contract with each service provider, we agreed in DARIAH for the moment, to set up only one contract between the users and only one of the provider partners, that includes a hint, that this has to be fulfilled also at use of any of the services provided by the consortia partners of DARIAH.

The status at the moment is, that this contract is agreed by the users at an IBM-Lotus system of one of the partners, which provides a database of the valid contracts to the service providers. This will be integrated into the AAI structure of DARIAH soon, such that an appropriate attribute is defined and can be used by the services the ensure the agreement of the terms of use.[7]

[7] The current version of the terms of use can be found under:
https://s-lotus.gwdg.de/gwdgdb/age/terms-of-use.nsf/6777fe61d5a9b9fbc1257>9fb003b03b4/$FILE/Terms_of_Use_engl_v4.pdf
Please also note the appendix which contains the current version of the ToU of March 2013.

User Support

Author: Johannes Reetz (RZG)

The increasing number of DARIAH users and more services to be reliably provided via the technical infrastructure imply more support requests. The issue tracking functionality of JIRA was used and is still sufficient for developers. In order to be able to provide a consistent point of contact for technical and operational support requests, DARIAH-DE is currently evaluating a helpdesk platform with a flexible ticket queuing system and which is capable to connect the normal technical support channels at JUELICH, GWDG, KIT, RZG and other sites when needed. As a start, the computing centers already defined and published dedicated support addresses that serve concrete technical or operational support requests. The helpdesk systems under evaluation can also provide the platform for the more sophisticated helpdesk approach which comprises specialized and domain-specific support for the Digital Humanities that is envisaged in the next phase of the project.

Conclusion

During the first two years of the DARIAH-DE project, the computing centers and software partners participating in the consortium laid the preliminary groundwork for the establishment of a sustainable technical infrastructure based on the requirements and needs of researchers in the arts and humanities. While much of this initial work has already been productively realized, there are still some requirements to be addressed in the coming years, together with representatives from the computing centers, the information specialists and the software developers, and partners from the various disciplines in the humanities. As a research-driven infrastructure project, DARIAH-DE confirmed that research projects and national and international collaborations in particular need sustainable research infrastructures. Currently there are **more than 350 researchers** from a wide spectrum of disciplines in the humanities using the options and services provided by DARIAH-DE, such as the Confluence Wiki system, the Developer Portal, VMs, storage, or field-specific services such as the Geobrowser and DigiVoy.

The competences, skills, and expert knowledge accumulated over the course of the last two years and collected by the DARIAH-DE consortium form an excellent foundation to continue this research and community-driven process in the humanities. The number of users, which has rapidly grown in recent months, the diverse requests, and the enormous need for advising and support services from various DH projects—from thematic and methodological in addition to disciplinary perspectives—demonstrates that digital research infrastructures for the humanities can only be effectively established in the context of close interaction and connections between the areas of instruction, research, research data, and basis infrastructures, and that they can only be successful when the entire spectrum of arts and humanities disciplines are involved. The DARIAH-DE consortium and its partners look forward to supporting these activities.

Appendixes

Terms of Use (March 2013)

GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

Declaration of Consent to the Terms of Use

in the Context of Dariah-DE

DARIAH-DE

Gefördert vom
Bundesministerium für Bildung und Forschung

The use of the software, the resources and the infrastructure of the GWDG and that of all providers cooperating with the GWDG in the context of the project DARIAH-DE takes place at the user's own risk and responsibility. The realization of economic profit from the use of these resources is prohibited.

The laws of the Federal Republic of Germany apply to the use of this infrastructure, software, and associated resources. You commit yourself, that you will not use this infrastructure to distribute content of an insulting, vulgar, obscene, pornographic or libelous nature, or content glorifying violence, in any form. In addition, the conscious or intentional introduction or spreading of viruses or other malicious or harmful programs is strictly prohibited. You concede to the operators of this infrastructure, software and associated resources the right to relocate and/or remove content at their discretion. In particular, user accounts which have been created, including stored data, may be deleted after the termination date of a project, unless arrangements have been agreed for a continuation of service. You are not permitted to use your access to resources and services to obtain personal data in the sense of the Federal Data Protection Act (Bundesdatenschutzgesetz) or to store or process personal data obtained by other means. You are not permitted to copy copyrighted data or programs from one computer system to another. The use of illegally obtained copies of copyrighted data or programs on the resources of the GWDG or providers cooperating with the GWDG in the context of the project DARIAH-DE is strictly prohibited.

Breach, violation or infringement of these rules will result in immediate and permanent account closure. The operators of the infrastructure and associated resources reserve the right to take legal action and pass on network traffic data or similar data to the law enforcement agencies.

You hereby consent to the automated storage, processing and transfer of the personal data you provide at the time of registration and during operation to the institutions or other entities, whose resources are being used, as far as this is necessary to ensure proper operation of the resources. This information will be used for no other purpose.

The GWDG as well as providers of resources cooperating with the GWDG in the context of the project DARIAH-DE only assume responsibility for damage caused by hacker attacks or computer viruses if such damage resulted from intent or gross negligence on their own part.

By registering, you agree to these Terms of Use.

Abbreviations

BMBF: Bundesministerium für Bildung und Forschung, Federal Ministry of Education and Research

CEI: Charters Encoding Initiative

DARIAH-DE: Digital Research Infrastructure for the Arts and Humanities

DARIAH-EU: Digital Research Infrastructure for the Arts and Humanities

DH: Digital Humanities

DHd-Blog: Digital Humanities in the Germanspeaking area

EpiDoc: Epigraphic Documents

ESFRI: European Strategy Forum on Research Infrastructures

ERIC: European Research Infrastructure Consortium

GND: Common authority files

MEI: Music Encoding Initiative

MEISE: Music Encoding Initiative Score Editor

OAI-PMH: Open Archives Initiative - Protocol for Metadata Harvesting

TEI: Text Encoding Initiative

VRE: Virtual Research Environment

URLs

BMBF: www.bmbf.de

Brochure on DH Curricula: www.cceh.uni-koeln.de/dh-degrees-2011

DARIAH Bibliography: www.zotero.org/groups/dariah_bibliography

DARIAH-DE: www.de.dariah.eu

DARIAH-EU: www.dariah.eu

DHd-Blog: www.dhd-blog.org

ESFRI: ec.europa.eu/research/esfri/

List of DH Curricula: www.dig-hum.de/studienstandorte

TextGrid: www.textgrid.de

VRE Guide of the Allianz-Initiative:
www.allianzinitiative.de/fileadmin/user_upload/Leitfaden_VRE_de.pdf

Zotero Support: www.zotero.org/support

Email addresses

dariah-sub@sub.uni-goettingen.de
dhdblog@mpiwg-berlin.de
empfehlungen@dariah.eu
forschung@dariah.eu
forschungsdaten@dariah.eu
Info-cceh@uni-koeln.de
infrastruktur@dariah.eu
lehre@dariah.eu
register@dariah.eu
support@de.dariah.eu