DARIAH

Digital Research Infrastructure
for the Arts and Humanities

DARIAH Technical Report – overview summary

Project acronym:	**Preparing DARIAH**
Project full title:	**Preparing for the construction of the Digital Research Infrastructure for the Arts and Humanities**
Grant agreement no.:	**211583**
Document status:	**In progress, public version**
Version Number and date:	**November 2010**
Author(s):	Stavros Angelis (DCU), Andreas Aschenbrenner (UGOE), Agiatis Benardou (DCU), Tobias Blanke (CeRch), Natasha Bulatovic (MPDL), Lou Burnard (CNRS), Panos Constantopoulos (DCU), Costis Dallas (DCU), Seth Denbo (CeRch), Malte Dreyer (MPDL), Matthew Driscoll (NFI), Senka Drobac (RBI), Michael Franke (MPDL), Christiane Fritze (UGOE), Dimitris Gavrillis (DCU), Andreas Gros (MPDL), Ivan Grubišic (RBI), Eric Andrew Haswell (NFI), Maarten Hoogerwerf (DANS), Pierre-Yves Jallud (CNRS), Stuart Jeffrey (ADS), Thomas Kachelhoffer (CNRS), Neven Kmetic (RBI), Katerina Kouriati (AA), Barbara Levergood (UGOE), Wolfgang Pempe (UGOE), Mike Priddy (CeRch), Julian Richards (ADS), Laurent Romary (UGOE), Rutger Kramer (DANS), Karolj Skala (RBI), Eva Wedervang-Jensen (NFI)

0. about this document

This deliverable provides an abstract of DARIAH[1] technical work accomplished during the initiative's preparatory phase. It builds on extensive surveys of user requirements, architecture modelling and technical prototyping. The raw materials to those preparatory activities are captured in the full report, as well as code repositories and other documentation.

The technical architecture aims to deal with the rapid technical and organisational evolution in the field, through modularity and decentralisation. Continual change also had impacts on the documentation of the DARIAH technical architecture: the concepts described in this overview report are on a high level of abstraction to capture the more stable bits of the architecture; nevertheless, they can be expected to evolve and grow over time.

Throughout this document, there are several brief reference boxes to (a) real users and their requirements [in blue], (b) exemplary systems created and hosted by DARIAH partners with hundreds of users [in green], and (c) experiments and demonstrators that informed this work [in orange]. In addition to the abstract and theoretical reasoning displayed by the body of the text, these "reality checks" illustrate the choice, implementation and application of the future DARIAH technical infrastructure.

Index

[1] DARIAH – Digital Research Infrastructure for the Arts and Humanities. www.dariah.eu

1. Using DARIAH

DARIAH is designed to be a means for linking people, services and data for research in the arts and humanities. Most likely, DARIAH will not be one technical solution, but many, according to community activities and willingness to collaborate.

It is of paramount importance to ensure that the DARIAH infrastructure addresses the variety of “information behaviours” displayed by the widely diverse target user community of DARIAH. Two of them are demonstrated in the user stories shown in the blue boxes hereafter:

User Story “Material Culture”

D., holding a first degree in Classical Archaeology, is a PhD candidate in Material Culture at a Greek University; his postgraduate studies were on Anthropology. D. works mostly in the library of his university using his laptop.

He is currently researching material culture and the perception of space and landscape in an area of northern Greece. His topic lies within the areas of archaeology and cultural anthropology, therefore D.'s sources are both artefacts and interviews with local people as well as extensive visual material.

He is actively scouring relevant online sources in order to find previous work on the perception of space and space as artefact. When connected through the academic network, he is recognised by JSTOR and Ingenta and is given free access to papers stored by these services. But he also works a lot with grey literature and web-accessible materials in general: personally archived papers, as well as YouTube, and similar projects carried on elsewhere in the world. In addition, he checks on the availability and details of particular papers using an online library catalogue.

Apart from the browser, D. keeps a note-taking application open in another window. Every time he finds something of interest he copies the excerpt and adds it to his own notes. Apart from access to scholarly literature, D. manages continuously information about particular sources he works with, both digital and non-digital. He keeps this data in a personally constructed database where he categorises all his findings and adds personal comments and ideas.

As his work is largely interdisciplinary in nature, there is a propensity for making serendipitous discoveries while looking for something unrelated. This whole setup means that he needs to login to different services at the same time—even login multiple times a day if disconnected—and keep multiple open windows in his browser.

Needs a single sign-on system

User Story “textual sciences”

F. is Professor in Literature and Literary Computing. He works mostly in his office at the university where he has a desktop and is always connected to the internet but also sometimes on the road travelling to conferences, and even at home.

F. is currently involved in working on a highly interdisciplinary project of a hybrid edition of a literary text, encoding generic processes of drama and collaborating with specialists on the visualisation of the results. He is using text collections of text corpora, especially from his narratology work, or his work about ideas of history.

F. uses other online collections as well, such as the library of his university, which also has huge collections in digitised form. One of the problems he considers important is that online databases and metadata collections are in one language only, which makes them inaccessible to many scholars working in the same field as he.

F. uses quite a few IT tools, such as some XML editor like *<oXygen/> XML Editor* do most of his encoding and the XSLT processing and post processing. Interestingly, Information Management appears to be crucial in his literary and historical work. He envisages an environment in which you

> can collect data in such a way that allows you to pose questions to a text corpus and then if a thing has been useful you publish it, or, otherwise you just get rid of it.
>
> One of the key issues F. stresses is that of the perception of metadata, especially when it comes to interdisciplinary work, where people disagree on the meaning of a concept or word.

It was therefore a fundamental task to establish a framework, at an appropriate level of abstraction, for capturing ways of working, representing information behaviours, specifying user requirements and classifying corresponding services and tools. This framework was developed on the basis of an empirical, interview-based research, and adopting an activity theory approach. The major outcomes are a) a process model of scholarly activity and b) an analysis of user requirements correspondent to generic scholarly activities called “scholarly primitives”. An infrastructure based on such primitives might at best follow the concept of a marketplace of services.

1.1. Research Processes in the Arts and Humanities

In the course of our empirical research, we attempted to cover the best part of the conventional arts and humanities disciplines, from classical and prehistoric archaeology to ethnomusicology and art history. Researchers come from across the fields of the arts and humanities, covering a wide scope of disciplines. They can be at various stages of their careers, ranging from postgraduate researchers to full professors.

In order to develop a conceptual model for scholarly research activity that we checked for relevance on the basis of an initial analysis of empirical research. The model (Figure 1) complies with the CIDOC Conceptual Reference Model, an established and stable international standard (ISO 21127) for cultural information (Crofts et al. 2010).

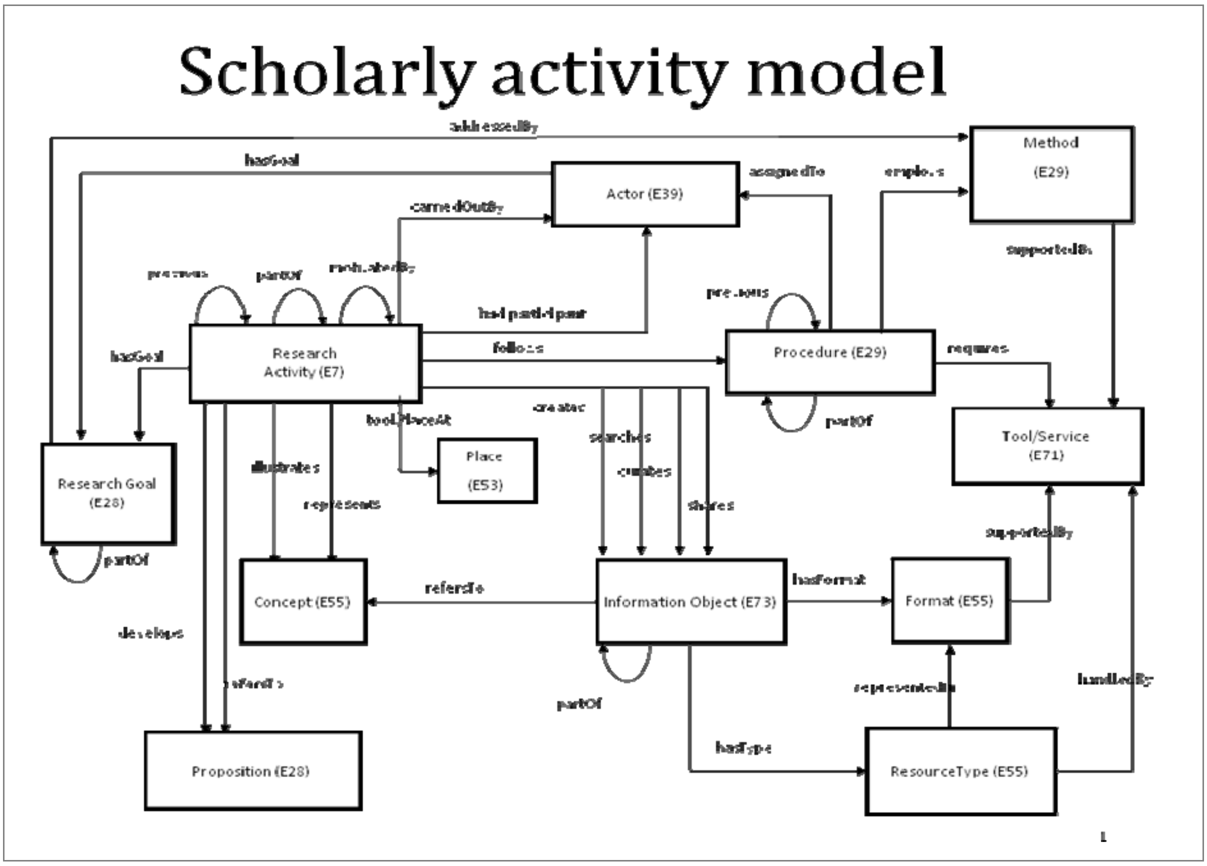


Fig. 1: Scholarly Activity Model.

Regarding the components of the scholarly activity model, scholarly research is understood as a *purposeful process*, carried out by *actors,* individuals or groups, according to specific *methods,* following a cultural historical activity theory approach.

Research processes usually are complex, consisting of simpler *tasks,* which may be carried out in parallel or in series. The detailed structure of the research process, and way of working for each step, are specified by a corresponding *procedure*.

The DARIAH model of the research process is intended to facilitate the design and development of information repositories and services in digital infrastructures that support research in the arts and humanities. To this end the model should allow to represent the details of a particular research activity both at the level of planning (*how it should be done*) and of actual execution. This dictates the distinction between process and procedure while maintaining corresponding (though not necessarily isomorphic) descriptions of the two.

Services thus become an important mediator between methods, procedures and information repositories. From a functional perspective, digital scholarship is embodied in services available. From a methodological perspective, services evolve to better meet requirements.

1.2. User Requirements

The literature (cf. Palmer et a. 2009) defines five “scholarly activities”: *searching, collecting, reading, writing, collaborating.* These are further refined to a more detailed list of twenty granular “scholarly primitives”; of these, *browsing, collecting, re-reading, assembling, consulting* and *notetaking* were found to be particularly common in the humanities. (Ross, 2010) pointed out that research primitives need to be differentiated from information seeking processes, and perhaps systematically related with them.

Primitives are one proven way of communicating infrastructures with stakeholders. They help develop infrastructure functions as common abstractions on research processes and help shape future infrastructure decisions. They are less useful to understand gaps in existing technologies and services to support every single step within research activities. Here, research life cycle models are shaping the way we study digital research processes.

The table below displays how user requirements emerging from DARIAH conceptual modelling work relate to generic “scholarly activities” mapped out by Palmer et al. 2009.

User requirement group	User Requirement	Searching	Collecting	Reading	Writing	Collabo-rating
Information management	Categorization of digital content	x	x	x	x	x
	Integration of digital content	x	x	x		x
	Personalised and preference-aware data management	x	x	x	x	
	Single, personal workspace holding integrated material from various sources	x	x	x	x	
	Context-aware data management	x	x	x	x	
	Tools supporting archival digital format (similar to PDF/A)		x	x		x
	Data annotation tools, supporting storage of comments, thoughts and	x		x	x	x

User requirement group	User Requirement	Searching	Collecting	Reading	Writing	Collabo-rating
	responses					
Information access	Tools enabling composite-term ("keyphrase") rather than single-term ("keyword") search	x	x			
	Search across multiple, distributed data sets, texts and images	x	x			
	Personalised and preference-aware data retrieval	x	x			
	Effective viewing and use of digitised visual and audio material			x	x	x
Collaboration	Tools and services supporting collaboration					x
	Shared workspaces					x
	Real time communication					x

Table 1: User Requirements.

User requirements established by DARIAH on the basis of evidence from scholarly practice fall within five key themes*: scope of information objects, access conditions, semantic interoperability, collaborative work,* and *information use versus analytical tools* (cf. Benardou et al 2009).

Scope of information objects

All strands of our research—review of literature, interviews with researchers and expert forum—point to the fact that scholarly activity engages simultaneously with different kinds of information objects, from those typically defined as "primary data" to scholarly objects that are the outcome of scholarship (such as publications); in fact, the clear distinction between primary data and scholarly objects is blurred in actual practice. On the basis of this fact, it is imperative that arts and humanities infrastructures provide effective means for identification, reference, access, representation and management mechanisms for this continuum of information objects, ranging from primary data and resources to complex scholarly objects.

Access conditions

Regarding access conditions and rights, it should be noted that the humanities research environment is characterised by a complex layering of different access regimes for different kinds of resources. These range from open access for some, to heavily restricted access for other kinds of resources due to either intellectual property or heritage protection legislation. It is important that this broad diversity of access regimes cuts across different kinds of information objects, from primary resources to the outcomes of scholarly work. A credible system that allows consensual sharing and reuse of such socially generated information (as in annotations on particular information objects) needs to provide adequate mechanisms of sharing and trust.

Semantic interoperability

Semantic interoperability emerges as a key need for scholarly research, especially when such research is cross-disciplinary, multilingual, or based on distributed resources. This is exactly the general case for the DARIAH infrastructure on account of its multidisciplinary and European scope, and the fact that it will need to provide access to resources that are located in different primary repositories, expressed in different languages, and providing content organised on the basis of different disciplines.

Collaborative work

While the production of research outcomes in the arts and humanities still is typically solitary, findings in our empirical work and further evidence suggests that there are important social dimensions in the way research is supported by specific information-laden activities. In particular, it emerges that even “primary data” in humanities research are typically the product of “thick description”, i.e. the interaction of scholars with primary cultural objects on the basis of specific disciplinary knowledge and methods, and thus they have the imprint of specific scholars.

A second trend is the emergence of tightly-knit, active communities of practice between researchers in particular sub-disciplines and research areas, that is built around a collective blog, an online archival resource, or other digital tool or service. As digital humanists become participants of multiple communities, they share, interact and co-create on the basis of their research interests, scholarly competence, and activities in which they participate. Supporting fruitful interaction in these research communities may involve approaches, methodologies and tools developed in the field of knowledge management, especially where institutional meet social and personal aspects of sharing and managing information.

Tool-supported use of information

Understanding the nature and information requirements of scholarly research, notwithstanding differences between disciplines, research fields and methodological approaches, is an important motivation and a prerequisite for the definition of infrastructures, services and tools fit for the purpose of current and future scholarly research.

1.3. Digital Humanities Research Life Cycle

Going beyond user requirements analysis and linking them with the existing technical environment, we use the research life cycle to communicate tools and services we have worked with to support the various stages of arts and humanities research. How does a Digital Humanities life cycle model of research look like? A large part of its research will be linked to search and discovery of data and complementary information, which is afterwards gathered and repurposed for analysis purposes. The activities of Digital Humanities involve processes within these stages as well as connections among them. So a research life cycle approach has two aims:

- It enables the identification of coverage, where research processes are adequately covered by infrastructure, where they do not need to be covered, and gaps where infrastructure is missing to improve and enhance research processes.
- It informs the DARIAH consortium about partner technologies as well as other tools and services that might be useful to fulfil DARIAH’s aim to match the whole life cycle of research. We have concentrated on those technologies that members of DARIAH have either direct experience with or have easy access to.
- It gives an overview of a wide variety of technologies in the arts and humanities communities as well as those communities working together with DARIAH stakeholders, and it helps to make priorities on which tools and services to work on and support.

The figure below shows that with these technologies we can in fact cover a generic research lifecycle for arts and humanities.

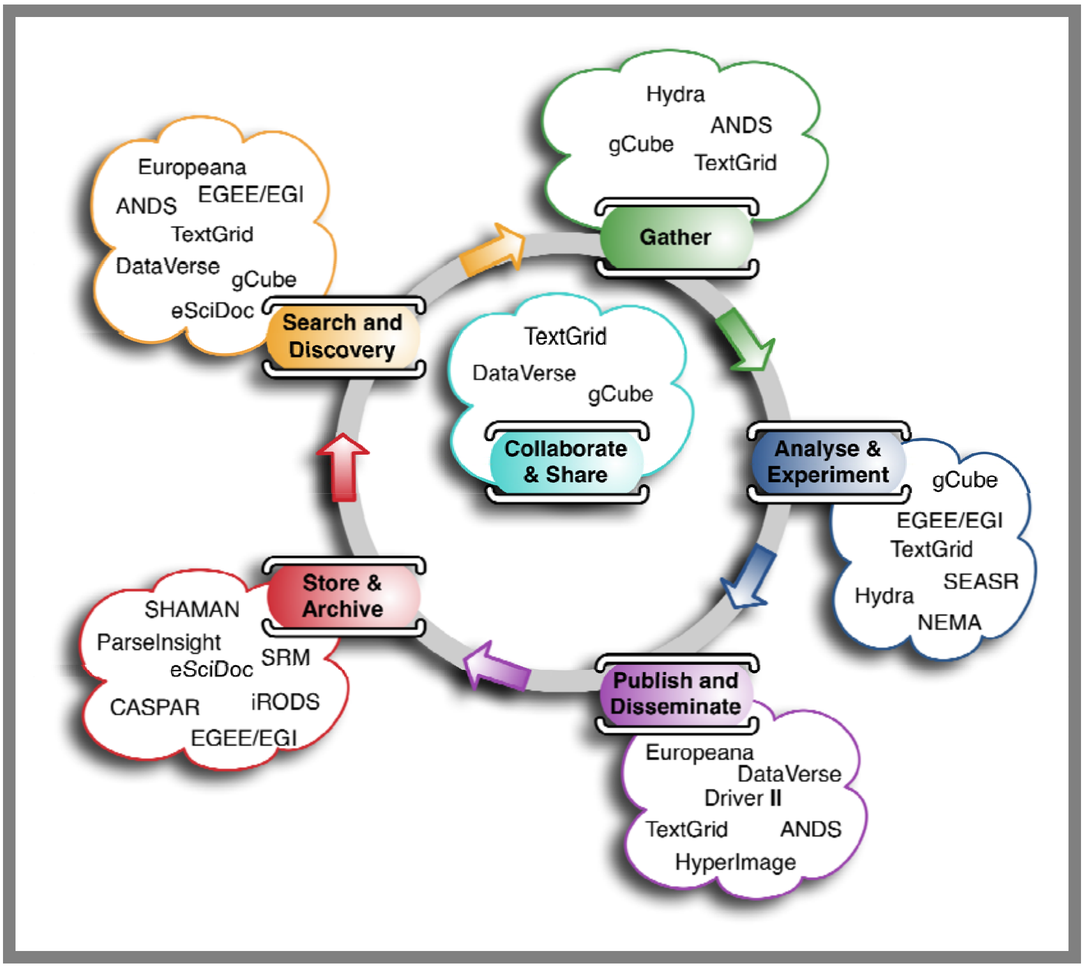


Fig. 2: Case studies mapped to the generic research life cycle for the digital arts and humanities.

2. Architecture of Participation

The DARIAH research infrastructure is an open, collaborative environment that enables research in the A+H by linking data, functionalities, and people. Its "architecture of participation" accommodates A+H data centres, research networks and researchers that are widely independent, stem from multiple backgrounds, interact with DARIAH following diverse goals, and employ various entry-points into DARIAH. Linking this diversity, DARIAH aims for a very light-weight and decentralised infrastructure that can be fit to each stakeholder's situation. Rather than a single technical solution, DARIAH may be many, according to community activities and willingness to collaborate.

The DARIAH technical architecture is built of three horizontal tiers, as well as vertical interoperability frameworks for both data and services. In each of those aspects and for every component, DARIAH seeks a broad interest base and collaborations. In particular, core infrastructure services may be created in close interaction with affiliated initiatives (including CLARIN and other ESFRI initiatives).

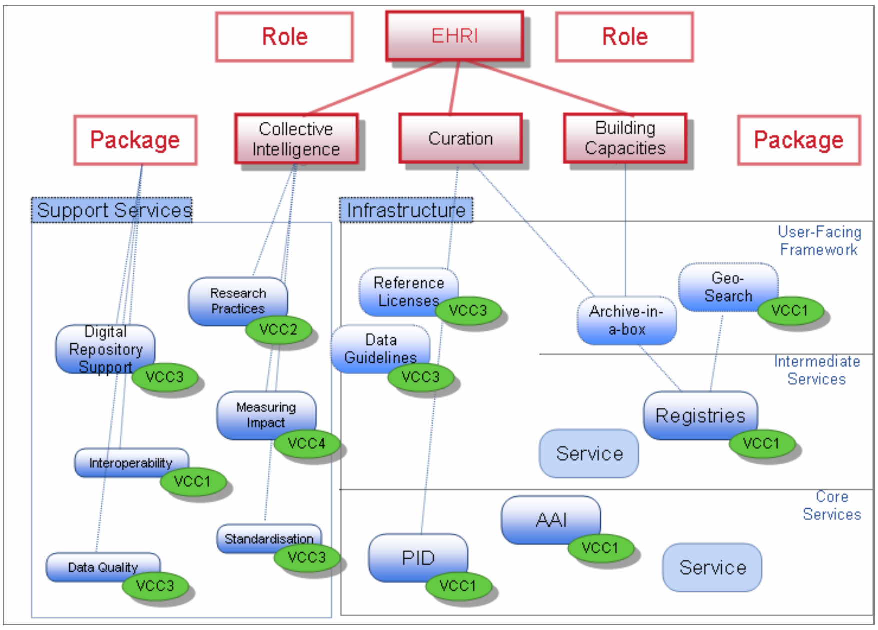


Fig. 3: Composing various services and infrastructure components to establish a tailored research environment, in this example for EHRI, the European Holocaust Research Infrastructure.

2.1. Three-Tier Architecture Model

The DARIAH technical infrastructure is built as a loosely-coupled service-oriented architecture with three structural tiers in its architecture model: (a) the user-facing framework, (b) infrastructure service environment, and (c) core infrastructure. It also describes how services can move up and down these horizontal tiers, to enable an architecture of participation that is open to contributions and evolves over time.

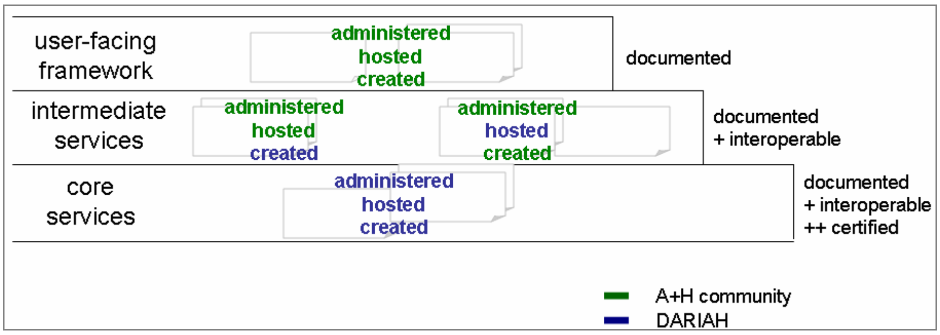


Fig. 4: Each tier may open up different organisational contexts for managing service components. Infrastructure services are created, hosted and administered by DARIAH ensuring reliability and scalability, whereas the A+H community is encouraged to contribute their own, potentially transient services towards the higher-level tiers.

1. User-Facing Framework (UFF)

The UFF accommodates a collection of end-user tools contributed by research projects or third parties. At a minimum, components in the DARIAH UFF tier need to be well-documented to facilitate reuse. Beyond mere documentation, tools and services ideally comply with the DARIAH service framework to foster interoperability with other DARIAH components. Other than that, there is no central control of development efforts in the UFF tier; collaborations are encouraged within the open DARIAH developer community.

2. Infrastructure Service Environment

Reference services fill the infrastructure service environment with life by offering actual research-relevant content for reference and reuse. For example, authority data on authors and other persons, thesauri, dictionaries from various epochs, and other reference data are often essential for research initiatives, yet are outside their scope. Shared reference services that offer data for reuse and perhaps mechanisms to contribute new data are hence infrastructure components for ensuring quality and efficiency in A+H research, as well as focus points for collaboration.

3. Core Infrastructure

The core layer includes services that serve to sustain the DARIAH infrastructure and establish coherent operation across the open DARIAH environment. Services such as a Persistent Identifier (PID) resolver and Authentication and Authorisation Infrastructure (AAI) are essential for enabling interoperability across the heterogeneous data sources and decentralised services in the DARIAH ecosystem. Other components in the core tier offer statistics and monitoring for ensuring stability and evolution in the DARIAH infrastructure despite its decentralised and open nature.

X. 'In-a-box' Services

These are currently two special DARIAH-created solutions aimed at A+H institutions who wish to create their own new digital archives or wish to build a digital research environment for their institution's research community. Both 'In-a-box' solutions combine software that is installed and administered at the institution and 'connects' to the DARIAH central infrastructure services.

2.2. Interoperability Frameworks for Data and Services

Linking diversity is at the core of DARIAH's philosophy. Disciplines in the humanities differ greatly with regard to their resources – their data, tools and methodologies. Moreover, innovation is sometimes associated with introducing variations into their data, tools, or methodologies, thereby reinforcing heterogeneity even within a single discipline. Through linking this diversity DARIAH aims to build bridges, and enable researchers from different disciplines or cultural backgrounds to collaborate on the same material, and to share their diverse perspectives and methodologies. A prerequisite to benefit from this opportunity, however, is interoperability between the diverse resources in DARIAH without enforcing specific formats. In other words, DARIAH aims to mediate between heterogeneous resources, and even though interoperability guidelines are optional, their implementation opens up additional opportunities such as increased visibility, collaboration, and the applicability of advanced techniques.

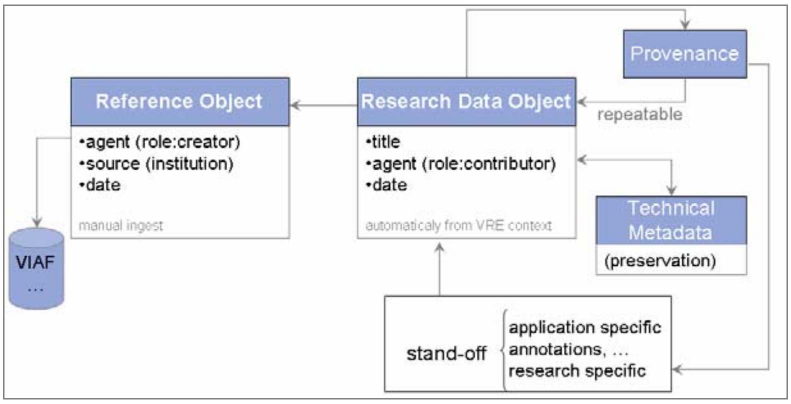


Fig. 5: Schematics of the DARIAH content model. To facilitate the role as a mediator between decentralised sources with potentially differing object semantics, the DARIAH metadata approach follows, among others, the conceptual foundations developed by Dublin Core[2].

Among the interoperability channels in DARIAH are digital objects and the data sources that contain them, as well as services and research environments.

- **Research objects:** Content models in DARIAH distinguish between the reference object (e.g. a sculpture by Michelangelo, a paper manuscript by the archaeologist XY, the born digital BBC Domesday[3]) and the digital research data object used for scholarly work. An object's content model may have modules that are specialised to specific domains or applications, and it may be distributed to various systems
- **Data sources:** Potential sources may include large institutional archives as well as homepages of individuals. Both the technical protocols for federating these data sources and their policies and concepts need to be shared to ensure effective interoperability
- **Services and tools:** To achieve service interoperability across tiers, service providers, and scholarly domains, issues need to be resolved including the underlying technical paradigm (e.g. SOAP, REST), passing on user identification, and ensuring provenance
- **Research environments:** Ideally, research environments are tailored to the specific needs of a domain or a research question, and they potentially combine various services and tools that may be reused between different research environments

[2] Stuart Weibel. The Evolving Metadata Architecture for the World Wide Web: Bringing Together the Semantics, Structure and Syntax of Resource Description. 1997 (http://www.dl.slis.tsukuba.ac.jp/ISDL97/proceedings/weibe.html) and Thomas Baker: A Grammar of Dublin Core. D-Lib Magazine, October 2000 (http://www.dlib.org/dlib/october00/baker/10baker.html)

[3] BBC Domesday (http://www.domesday1986.com/)

TEXTGRID

TextGrid is one of the first grid-based community projects in the humanities creating an infrastructure for the collaborative editing, annotation, analysis, and publication of specialist text resources. TextGrid represents the humanities in the German national grid initiative D-Grid, and provides a digital infrastructure, a collective network, and a comprehensive and extensible toolset for text scholars.

Technologically, its combination of grid and repositories, as well as services and tools with graphical interfaces establish an open environment that can be adapted to many use cases. In its core functionality, TextGrid focuses on (annotated) text as a data type since there is considerable demand in the community for processing text data.

TextGrid is particularly interesting for its openness. It avoids swamping the user with rules and requirements, yet still fosters participation and collaboration. For example interoperability can be achieved in a stepwise process and following an incentive system:

- Any data format can be uploaded, TextGrid ensures bit-preservation;
- Metadata facilitates data management and retrieval;
- By uploading XML-based texts, a series of services can be used on the data including streaming tools, an XML-editor, and other functionalities;
- For TEI encoded documents, TextGrid offers graphical editing, metadata extraction, and other functionalities;
- Defining a mapping to the TextGrid recommendation for a TEI core encoding allows interoperability on a semantic level.

One of the core goals of TextGrid lies in enhancing the re-usability of existing scholarly texts and services. For both areas—data and services—TextGrid offers various levels of integration: the lowest level offering a minimum barrier to participate, and the highest level offering maximum interoperability on a semantic level. Getting from lowest to highest is a stepwise process, and users are motivated and assisted for taking each step. In other words, interoperability is not given by design, but it is encouraged.

TextGrid is an example of a technology that supports core Digital Humanities of XML annotation in standard formats. It allows for collaborating and sharing of TEI resources as well as their dissemination.

MIXED– Migration to Intermediate XML for Electronic Data

MIXED is a digital preservation project. It uses a strategy of converting data to intermediate XML, and specifically for tabular data.

One of the obstacles to preserving data is software obsolescence. This stumbling block to preservation is usually tackled by either continuously migrating the data or by emulating the software tools. MIXED achieves preservation by migrating data to standard formats.

The strategy used with MIXED converts all datasets, upon ingest to the archive, into an intermediate, generic format. Upon dissemination of a dataset, it is converted from this generic format into a current vendor format of choice. It is likely that the intermediate format will also change, but at a much slower rate. The optimisation is that conversions are split into many contemporary conversions and a few time-bridging conversions. This is a much more manageable situation, and the complexity of bridging time can be dealt with by means of one well-defined format.

MIXED concentrates upon tabular data because the lack of standardisation is most keenly felt here, and there are several reasons for the acuteness of this feeling.

An XML schema, called M-XML, is used by MIXED in such a way that database and spreadsheets are expressed as valid M-XML documents. This format is a non-proprietary representation of the data. MIXED is an open source framework, which accepts converters as plug-ins. The converters allow conversion from existing vendor application formats to M-XML, and vice versa.

On ingest of tabular material into an archive, the data is converted to M-XML, and on dissemination, the data from M-XML is converted to any spreadsheet or database format the end user requires. Therefore it is possible to convert from one application format to another.

eSciDoc

eSciDoc is an e-Research environment developed specifically for use by scientific and scholarly communities to collaborate globally and interdisciplinary. It is an infrastructure encapsulating a Fedora Commons repository and implementing a broad range of services.

Its service-oriented architecture fosters the creation of autonomous services, which can be re-used independently from the rest of the infrastructure. eSciDoc provides a generic infrastructure and specialised solutions within the context of research questions. It integrates existing solutions and implements new ones.

The target audience of eSciDoc are research organisations, universities, institutes, and companies interested in e-Science-aware knowledge and information management. eSciDoc enables the user to publish, visualise, manage, and work with data artefacts or objects. Objects include both publication data and research data across disciplines. eSciDoc addresses aspects of data reliability, data quality, data curation and long-term preservation. It covers the whole lifecycle of objects, and supports semantic relations between objects.

The eSciDoc system is designed as a service-oriented architecture (SOA) implementing a scalable, reusable, and extensible service infrastructure. Application- and discipline-specific solutions can then be built on top of this infrastructure. Data resource access is provided via Fedora’s own REST and SOAP interfaces.

3. Service Catalogue and Roadmap

This section provides an inventory and a timeline for the technical infrastructure services to be created during the DARIAH construction phase.[4] It presents the core building blocks of DARIAH and their role in the infrastructure, although not all of them are currently covered or fully specified. With the core infrastructure in place, further services will be added and adapted as new countries join DARIAH and novel technology developments emerge. For example, partner projects like the EHRI Holocaust Research Infrastructure add new requirements and opportunities to constructing DARIAH. More such dedicated research environments that build upon DARIAH are expected.

User-Facing Framework (UFF)

- **DARIAH Discover** - search across A+H collections, potentially enabling analysis and visualisation (e.g. geo-browsing, relation networks) for subsets of the objects

DARIAH in-a-box services

- **Archive-in-a-box** - reference list to relevant repository technologies
- **Research-Environment-in-a-box** - reference list to relevant research environment technologies

Infrastructure Service Environment

[4] More detailed specifications, cost estimations, evaluations of existing technologies, etc can be found in DARIAH reports D8.1-D8.2

- **Collections Registry** - machine-readable registry for object sources
- **(Ad Hoc) Resource Registry** - machine-readable registry for orphaned objects
- **Metadata Registry** - management and versioning of metadata schemata to enable e.g. metadata schema mappings
- **Services Registry** - references and descriptions for relevant tools and services
- **DARIAH Reference User Registry** - interoperability for user references across DARIAH partner systems with local user management and authorisation; mainly used to enable trustworthy and meaningful provenance and linked to AAI
- **DARIAH Authority Mediation Service** - a framework for integrating divers authority databases, for example for "creators" (e.g. authors, artists), dictionaries, controlled vocabularies and other reference data

Core Infrastructure

- **Persistent Identifier Services (PID)** - PID service and meta-resolver across distinct PID schemata
- **Authentication, Authorization Infrastructure (AAI)** - single-sign-on across Europe
- **Infrastructure Management & Information Services** - statistics and monitoring of (infrastructure) services and data
- **Provenance Tools** - a framework for tracking and visualising provenance in digital objects as part of ensuring the integrity and authenticity of digital objects

2011			
Q1	**Q2**	**Q3**	**Q4**
(startup phase)		**Developer Portal (DARIAH-internal)**	**Collection Registry v1**

2012			
Q1	**Q2**	**Q3**	**Q4**
Metadata Registry v1 **Archive-in-a-Box**	**Discover v1 (plain)** **Authority Med v1** **AAI framework (to be adopted by countries)**		**Collection Registry v2** **(Ad Hoc) Resource Registry**

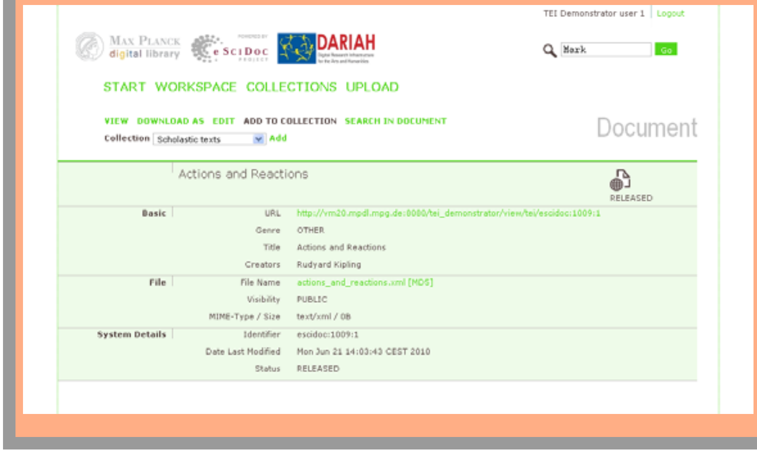
2013			
Q1	**Q2**	**Q3**	**Q4**
Metadata Registry v2	**Discover v1 (geo, graph, trust)** **Authority Med v2** **VRE-in-a-Box**		**User Registry** **Provenance** **Service Registry** **Infrastructure Monitoring**

TEI Demonstrator

The purpose of the DARIAH TEI Demonstrator is to demonstrate the practical benefits of using TEI for the representation of digital resources of all kinds, but primarily of original source collections within the arts and humanities. As a community-focussed project, the demonstrator also aims to make it easy for humanities researchers to share TEI-encoded texts with others, and to compare their encoding practice with that of others in the TEI community.

The functions it provides are aimed primarily at humanities researchers with the following requirements. Registered users

- can upload and publish a TEI resource and associated materials to an online repository;
- can validate the TEI resources against the TEI-ALL Schema;
- can integrate the metadata provided with their resource into the repository database to facilitate cross-searching;
- can carry our a simple free-text and metadata search or combine it with more sophisticated XML-aware searching possibilities, and extract subsets ("collections") of documents;
- can generate an publishable XHTML version created dynamically from the TEI.



Collection: Scholastic texts

Actions and Reactions

Basic	URL	http://vm20.mpdl.mpg.de:8080/tei_demonstrator/view/tei/escidoc:1009:1
	Genre	OTHER
	Title	Actions and Reactions
	Creators	Rudyard Kipling
File	File Name	actions_and_reactions.xml [MD5]
	Visibility	PUBLIC
	MIME-Type / Size	text/xml / 0B
System Details	Identifier	escidoc:1009:1
	Date Last Modified	Mon Jun 21 14:03:43 CEST 2010
	Status	RELEASED

The initial implementation of the TEI Demonstrator uses the eSciDoc platform. It also uses the full spectrum of eSciDoc services to support service development and deployment. As everything is open source, new services can be easily added and existing ones can be amended.

ARENA2 Demonstrator

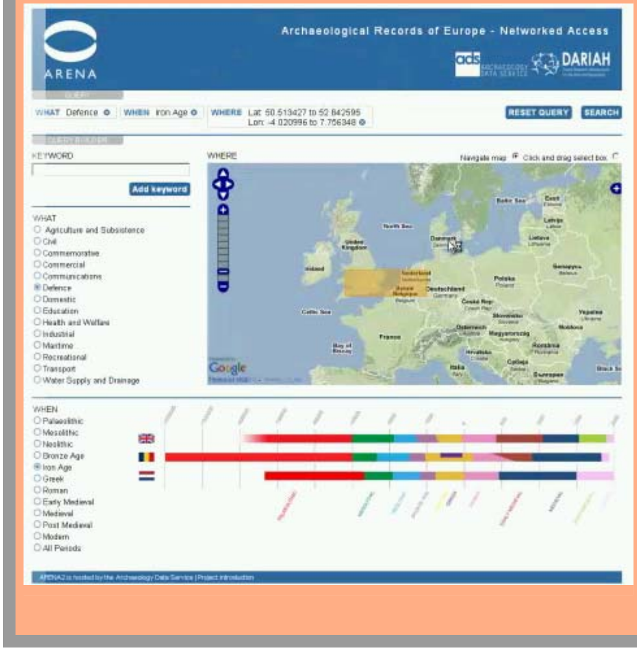
The objective of the ARENA2 demonstrator is to migrate ARENA into a sustainable environment by adding service logic and exposing its resources as autonomous services in a Service Oriented Architecture (SAO) over selected partner data centres.

This demonstrator will exhibit added value in terms of the ability to sustain applications from cultural heritage and arts and humanities research beyond the lifespan of this particular project.

The basic architecture adopted follows the ‘Publish-Find-Bind’ approach. Services publish themselves to a registry as being in accordance with a web service specification. These services are then found and bound to by a client. Key to creating a SOA implementation of the ARENA2 service is the specification of how services should communicate. This required the creation of the ARENA Gateway Service Specification document (AGSS). The Arena Gateway Service Specification itself consists of a WSDL (Web Services Description Language) document.

In this case the specification is the ARENA Gateway Service Specification, the client is the ARENA2 portal, the services are either compliant monument inventory services or ‘wrapped’ services based on legacy protocols such as z39.50 or OAI PMH and the registry is an instance of a Universal Description Discovery and Integration registry, the ARENA UDDI registry.

The existing ARENA2 prototype uses a very simple query building interface allowing the user to set values for the ‘Where, What, When’ elements of the service specification to build up a query. During the process of creating the service specification for each of the ‘Where, What, When’ elements a universally agreed and available controlled list would have to be selected to of a schema to which each of the diverse data sources could be mapped.



WHAT Defence

WHEN Iron Age

WHERE Lat: 50.513427 to 52.842595
Lon: -4.020996 to 7.756348

ARENA2 is hosted by the Archaeology Data Service | Project introduction

- WHERE – the search interface will have an open layers based geospatial selection interface;
- WHAT – mapping of the MI records to the top-level terms of the poly-hierarchical English Heritage Thesaurus of Monument Types (TMT);
- WHEN – the Forum for Information Standards in Heritage (FISH, UK) maintain the Manual and Data Standard for Monument Inventories (MIDAS) controlled vocabularies.

The service version of the ARENA2 portal integrates a number of design features that allow more intuitive and meaningful searching in comparison to the existing online prototype. The interface styling itself has benefited from extensive user testing of a similar design paradigm, taking place in relation to a separate project undertaken at the ADS.

General Purpose VRE Demonstrator

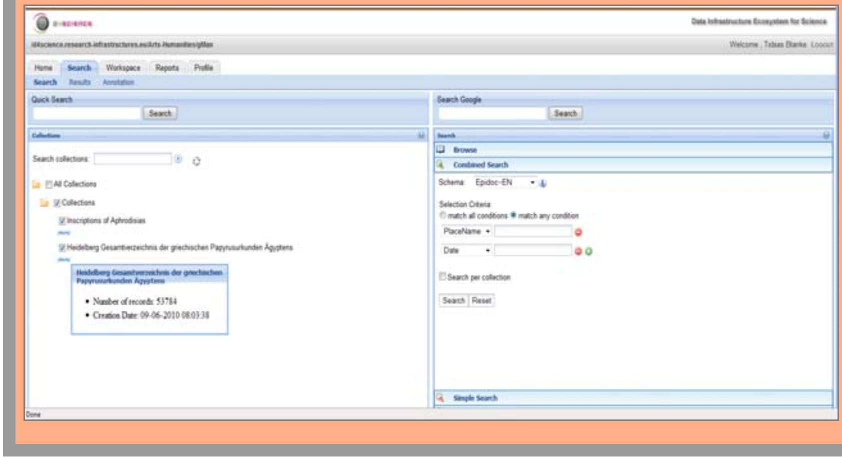
A VRE is considered a collaborative digital environment that facilitates the integration of information resources and tools for supporting research activities. The Centre for e-Research-based TEXTvre using the TextGrid environment project is concerned with the institutional integration of VREs in the specialised domain of Digital Humanities, specifically the creation of XML-based resources.

This VRE demonstrator aims to find new ways of integrating and organising the heterogeneous and often unstructured digital resources used in humanities research, including advanced search and browse services. Standardisation is unlikely to solve all issues raised in linking up humanities data, for several reasons:

- There is a great deal of legacy data in diverse and often obsolete formats;
- Training users in the application of a standard may incur a significant investment of time and money, which is not always available;
- Standards are generally developed within particular disciplines or domains, such as inscriptions, whereas research is often inter-disciplinary, making use of varied materials, and incorporating data conforming to different standards.

This demonstrator is based on use cases that were identified during the earlier research activities of the JISC ENGAGE project LaQuAT. LaQuAT investigated how to integrate scattered, heterogeneous and autonomous data resources relating to ancient texts, mainly databases but also including XML documents.

The starting point for this demonstrator was D4Science, a production-level infrastructure serving mainly scientific communities, which is not biased towards any particular discipline. gCube, on which the infrastructure is based, is a distributed, extensible system designed to support the full life-cycle of modern research, with particular emphasis on application-level requirements for information and knowledge.



Schema: Epidoc-EN

○ match all conditions ● match any condition

PlaceName

Date

Heidelberg Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens

- Number of records: 53784
- Creation Date: 09-06-2010 08:03:38

The demonstrator will address a research scenario involving the following stages:

- Document-centric and text-centric search; creation of virtual collections;
- Creation of annotations and links;

Generation of research reports.

PID Experiments

Permanent storage and access to digital material requires a more durable referencing method than currently employed by the Internet.

One mechanism that is widely used to deal with this problem of resource location changing is called Persistent Identifiers (PIDs). In short, PIDs are given to any resource or object that needs to be permanently identifiable. Once a PID is minted for a resource, it is tied to this resource for an indefinite period, and any reference using this PID will always refer to the resource it is tied to.

When a researcher cites an article or dataset in his (hardcopy) thesis, he needs to be assured that the citation itself will always lead to the original resource he has used.

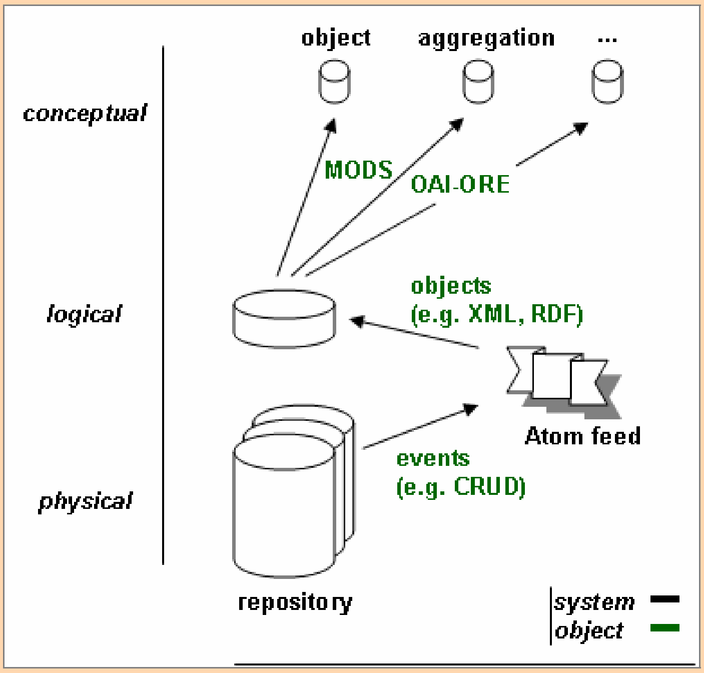
The PID experiment has been carried out to investigate in what way PIDs can help DARIAH design and construct its research infrastructure. To this end, a prototype PID system has been implemented that demonstrates the following scenarios:

- Forward clients to the actual location of the resource, based on the HTTP protocol, regardless of the PID's origin;
- Enable users to refer to parts of resources;
- Enable users to refer to particular representations of resources.

The architecture of the system accounts for the fact that, although this implementation will be based on HTTP, it should also be capable of working with other protocols. In other words, the architecture has been made flexible enough to adapt to different protocols and delivery mechanisms.

OAI-ORE / ATOM

In this infrastructure experiment, which directly targets the interoperability layer, we explored building a highly flexible repository federation for research data on the basis of loosely-coupled services and formats. We examined a prototype for the federation of grid-based TextGrid repository with an iRODS/Fedora repository, which caters for data analysis (i.e. XQuery capabilities on XML/TEI objects) across repositories, and other conceivable applications.



Research questions:

- Modelling in OAI-ORE and related standards. Transfer Fedora / OAI-ORE, automatic vs. manual, internal vs. external, etc.
- Disaggregating object models. Linking with external agents (Europeana, DRIVER).
- Federation of Repositories through OAI-ORE (virtualisation on the semantic/repository layer, as opposed to the storage layer—we look at both layers separately).

A Cloud API to the Grid

This experiment analysed patterns for mixing and merging infrastructures. In particular it looked at "repositories" as they overarch scientific infrastructure and interactive applications. In an analysis covering a series of experiments, we find an optimal setup in the combination of grid and web technologies through a REST-based interface, which opens up a variety of novel architectural patterns. This combines two contexts and usage patterns: infrastructure for large-scale scientific applications on the one hand, and open environments for interactivity and user-generated content and services on the other.

We connected grid and web environments and developed a RESTful abstraction upon the Storage Resource Manager (SRM). SRM is a versatile, pivotal grid standard, and it is highly interlinked with its environment on an operating system level. As an interface between a grid node (SRM) and a web server (the repository), we hence looked for a lightweight interface that is capable of translating between the two worlds. Existing cloud services are a premier model for translating between infrastructure and the web. Despite their simplicity, REST-based protocols satisfy all the needs of the web community.

Our experimental implementation therefore re-engineered the REST API of the Amazon S3 storage service using Python WSGI (Web Server Gateway Interface). The advantages of such a loosely-coupled, HTTP/REST-based architecture are manifold: The interface between the repository and the cloud-like service is obviously very light-weight. Due to the loosely-coupled architectural paradigm, the interdependencies between infrastructure and application (in this case: the repository) are minimised and the two can evolve separately.

A generic repository storage API and a decoupled architecture pattern like this enables other services to tie into the system environment. Multiple repositories can build on a single storage, and even specialised services e.g. for format conversion or other administrative tasks are conceivable to work directly at the level of the S3 API. Administrative workflows triggered by the repository, yet executed on the storage level may boost overall scalability of the system environment considerably. Moreover, this loosely-coupled approach may trigger the creation of low- level repository services and hence a variety of agents interacting in an open repository ecosystem.

4. Conclusions: Context and Outlook

The DARIAH[5] technical analyses build on extensive surveys of user requirements, architecture modelling and technical prototyping. The raw materials to those preparatory activities are captured in the full report, as well as code repositories and other documentation. The concepts captured in this summary report are on a sufficiently high level of abstraction to remain comparably stable over time. During the DARIAH Construction Phase, the VCC1 e-Infrastructure will implement this technical infrastructure and evolve/extend it where needed. To achieve this it will build extensively on experiences from the partners, reuse existing technologies and connect running systems.

The DARIAH technical architecture is designed in a modular and decentralised way, such that it is capable of growing and evolving with the pace of technology. In this sense, DARIAH will not be a single technical solution, but many, tailored to the research domain, as well as the technical and organisational context. Key design criteria include openness, user- and research orientation, trustworthiness and a framework to ensure the durability of its resources. While system integration across decentralised DARIAH hosts, scalability and elasticity, as well as other objectives were part of the design, they may sometimes be second to some of the key design criteria. For example, the independence of national A+H data centers and other system components may impact on the overall integration from a user-

[5] DARIAH - Digital Research Infrastructure for the Arts and Humanities. www.dariah.eu

perspective[6]; and the scalability of ingesting valuable resources into trusted repositories may suffer, when resources are validated and perhaps converted to standard formats to ensure durability.

The DARIAH technical architecture is embedded in an organisational framework defined through the VCC structure (Virtual Competency Centers)[7]. Some of the requirements for DARIAH can be realised through technical infrastructure, through organisational infrastructure, or through both. Hence, the interaction of the DARIAH technical architecture (to be developed by VCC1) with the organisational structure needs to be ensured over time. For example, the development of ontologies is essentially a standardisation process to be discussed in the community; while the technical integration of an ontology will be covered by the technical infrastructure in VCC1 (DARIAH Virtual Competency Center e-Infrastructure), the development of the ontology will be covered by VCC3 (Scholarly Content Management).

In other words, "infrastructure" in the arts and humanities goes beyond hard- and software and includes e.g. data standards, guidelines for selection and description of data, reference data licenses, and other issues. While focusing on the technical aspects, this document recognises the interaction with non-technical issues.[8]

Eventually, DARIAH will in many ways be fueled by a diverse and active community that contributes its data, metadata, and tools. To build and nurture this community, DARIAH is devising various technical and organisational mechanisms. This includes a technical developer platform for distributed software engineering (by VCC1), as well as e.g. training/education (VCC2), best practices for data management and interoperability (VCC3), and community engagement projects (VCC4). From this perspective, nurturing an active community is essential not only for the usage, but also for sustaining and extending the infrastructure. In other words, the community is the infrastructure.

5. Bibliography

Benardou et al. 2009
Benardou, A (2009) “Understanding the information requirements of Arts and Humanities scholarship”, Online at: http://www.ijdc.net/index.php/ijdc/article/viewFile/144/206

Ross 2010
Ross, S. (2010) Contribution to "Scholarly primitives" session discussion, *Expert forum on scholarly activity and information process, Athens,* 10–11 June 2010.

Links

TextGrid: http://www.textgrid.de/en/ueber-textgrid.html

DGrid: http://www.d-grid.de

MIXED: http://mixed.dans.knaw.nl

eSciDoc: https://www.escidoc.org

SRM: https://sdm.lbl.gov/srm-wg

[6] e.g. while single-sign-on are common practice, authorization and security across decentralised infrastructure components may be more patchy.

[7] cf. DARIAH information booklet. www.dariah.eu

[8] For a description of the DARIAH VCC structure, policies, etc see www.dariah.eu (e.g. the Information Booklet)