PREPARING DARIAH

Panos Constantopoulos [a], Costis Dallas [a], Petern Doorn [d], Dimitris Gavrilis [a], Andreas Gros [c], Georgios Stylianou [d]

[a] Digital Curation Unit, Athena Research Centre, Greece - (p.constantopoulos, c.dallas, d.gavrilis)@dcu.gr
[b] Computer Graphics and Visualization lab, European University, Cyprus – g.stylianou@euc.ac.cy
[c] Max Planck Digital Library (MPDL), Munich, Germany – gros@mpdl.mpg.de
[d] Data Archiving and Networked Services, The Hague, The Netherlands - Peter.Doorn@dans.knaw.nl

KEY WORDS: Arts, Humanities, e-infrastructure

ABSTRACT:

In this paper, a preparatory project for an integrated European research infrastructure in the humanities is presented. This project, *Preparing for the construction of the Digital Research Infrastructure for the Arts and Humanities - or Preparing DARIAH for short*, is part of the ESFRI e-infrastructures programme and supports the emergence of a new collaborative framework in which researchers are able to maximise the impact of their work on the international stage and aims at providing the foundations for the timely construction of the infrastructure requisite for the arts, humanities and cultural heritage communities in the digital age. DARIAH uses an interdisciplinary approach and involves tackling a number of interrelated issues such as strategic, organisational, financial, technical and conceptual in order to facilitate long-term access to and use of all European humanities and cultural heritage information for the purposes of enhancing and expanding research, thereby increasing our knowledge and understanding of our histories, heritage, languages and cultures. The DARIAH network will act as a place where the incubation of new ideas and ways of working can be facilitated and developed, and then transitioned into established organisations thus ensuring long-term sustainability and stability and the integration of these methods and techniques into everyday research practice. DARIAH will support research practitioners at all stages in the research process, and at differing levels of sophistication, from beginners through to those employing advanced techniques and methodologies.

1. INTRODUCTION

The project, *Preparing for the construction of the Digital Research Infrastructure for the Arts and Humanities* - or *Preparing DARIAH* for short, aims at providing the foundations for the timely construction of the infrastructure requisite for the arts, humanities and cultural heritage communities in the digital age. The 'grand vision' for the DARIAH Research Infrastructure is to ensure the long-term availability and access to the cultural heritage related information along all European partners followed by its augmentation and expansion. It's long term benefits include the enhancement of existing knowledge, research expansion and the better understanding of our heritage, history, languages and culture. This vision is characterised by innovation: new ideas and ways of working will be incubated, facilitated, developed and turned into established organizations. DARIAH will make the integration of these new ideas and practises into everyday life easier and will also support researchers in all stages in the research process and at different levels of sophistication, from simple to more advanced techniques and methodologies.

The organization structure will consist of a central office based on one or more of the partners. The central office will act as a coordinator of the network. DARIAH intends to be inclusive and will welcome new partners who wish to contribute and learn. The guiding principle behind DARIAH is the bringing together of the different stakeholders into a federated knowledge network: digital archives, libraries, data centres, research practitioners, research groups, technologists, computer and information scientists and other supporting services such as legal and advisory services. This mix of all these stakeholders exists already.

2. VISION AND IMPACT OF DARIAH

Preparing DARIAH aims at preparing for the construction of DARIAH by 2010. This preparatory phase will address coordination, strategic, financial, governance, logistic, legal and technical issues. The project will deliver the following:

1. a business plan for the construction and operation of DARIAH
2. a consortium that will be committed to the construction and operation of DARIAH
3. all legal documents regarding the rights and obligations of DARIAH partners, allowing the inclusion of new partners.
4. financial support capable of sustaining the initial development and operation of DARIAH

DARIAH will make researchers aware of the data available in the EU community and provide them with the means to locate and access this data. Research practice in the arts and humanities is often about meaning, interpretation and re-interpretation. This meaning is usually extracted from the data available from a wide range of primary and secondary sources. DARIAH intends to exploit the widespread broadband connectivity and the power of the web-based tools available for the analysis of digital information in order to provide support for the changing nature of the research practice in the arts and humanities. With the proper knowledge creation and

information sharing, scholars will be able to explore vast amounts of information, answer new questions and perform their work more efficiently.

3. DESCRIPTION OF DARIAH

3.1 Strategic work and coordination

In order to ensure project viability, DARIAH must first identify its goals and objectives so that to ensure maximum impact. A clear and comprehensive view of the state of the art must be provided during the strategic work development. This work will provide an insight on the needs (data and tools) from users across Europe with emphasis on the partner countries. Products and services already provided by data organizations situated in partner countries will be identified and be included in the technological architecture. Finally, the subset of standards currently used in the arts and humanities in Europe will be indentified.

Although DARIAH starts with a core of partners and data providers, aims to expand to include partners from all Europe thus increasing its power as each partner joins the infrastructure. As a consequence, investigation must be carried out among the various EU member states in order to identify potential partners, organizations, material owners, researchers. A user requirement analysis must also be carried out especially because in DARIAH one can find many user communities (archaeology, history, classics and arts) with possibly different requirements. Based on the above studies, a non-technical standards framework will be created that will encompass the services that should be included in the technical framework. In this phase, a set of policies will also be developed for the following areas:

- Digitization
- Collection development
- Collection ingestion and management
- Preservation
- Compliance as a trusted digital repository

3.2 Financial work

To be able to provide long-term services to the European research community in the humanities, a sustainable business model for DARIAH has to be defined. The business model also has to ensure the project's adaptability to new partners' needs. Furthermore, it has to be based on a cost model that captures fixed costs, e.g. for core personnel, as well as variable costs related to services offered to individual scientists, research communities or wider national and international projects. Based on the cost model a corresponding funding scheme has to be developed. One crucial part of establishing the funding scheme will be to gain support from national funding agencies for services used by national research projects. Therefore, several national funding agencies will be included into the process of formulating a funding model from the very beginning. In addition, participating institutions support DARIAH through their own data centres. Furthermore, as soon as DARIAH can offer services and support to third-party institutions, partnership services can be included into the business model.

A well-designed business model will attract new partners and stakeholders (research institutions, universities, cultural heritage sector, publishers, funding agencies, government agencies and the European Commission).
Related to the financial work is also the preparation of exit strategies for long-term preservation of data in case the DARIAH infrastructure comes to an end. Such strategies have to be defined in close collaboration with cultural heritage institutions such as national libraries and archives.

3.3 Legal work

The exchange of cultural heritage data among many partners in different countries requires agreements that will lay down the rights and obligations of each partner.
Some of the legal issues that will be addressed by DARIAH are: data depositing, data preservation, data dissemination, and corresponding products and services to be provided.
Legal work is of major importance since DARIAH is inherently heterogeneous. Firstly, the partners are from different countries with different legislations. Secondly, the addressed organizations are different, e.g. universities, companies, public sector, and, thirdly, the distributed repositories hold data with different classification status, e.g. publicly available, partially available, or commercial.

An analysis of EU legislation – mostly regarding intellectual property rights and privacy regulations – must be carried out along with relevant EU directives concerning preservation and access to data. Special attention will be paid to the Open Access Movement on data and publications. Guidelines, like the UNESCO Charter on the Preservation of the Digital Heritage, will also be considered.
Among the deliverables for DARIAH are a consortium agreement, an accession form for future partners, a depositor licence agreement, and a draft user licence agreement for products and services.

3.4 Technological reference and architecture

A robust technological architecture will act as a technical reference and will be used in the proof of concept prototypes. This architecture will take into consideration the distributed and heterogeneous nature of DARIAH. More specifically, it will:

- Embed existing centres
- Leverage the creation of new centres.

Because of the enormous amounts of data that are available, the technological framework will adopt grid-based solutions mainly for the storage of the digital content. The technological architecture will consist of the following layers:

1. Basic infrastructure: This layer will be designed so as to provide virtual infrastructure for data storage, support single sign-on services and manage the distributed repositories within this virtual layer.
2. Virtual repository layer: This layer addresses the highly diverse and complex data in the arts and humanities by allowing for seamless access for users and additional services through the use of repositories.
3. Interoperability layer: This layer implements the standards and guidelines that will act as a "semantic glue" between the different repositories and will allow to link their contents together.
4. Service layer: This layer provides tools that will interface with the DARIAH archive and/or other

tools. Users will be able to "plug" their own tools and thus expand DARIAH services by creating new services or by augmenting existing ones through an open and flexible SOA architecture.

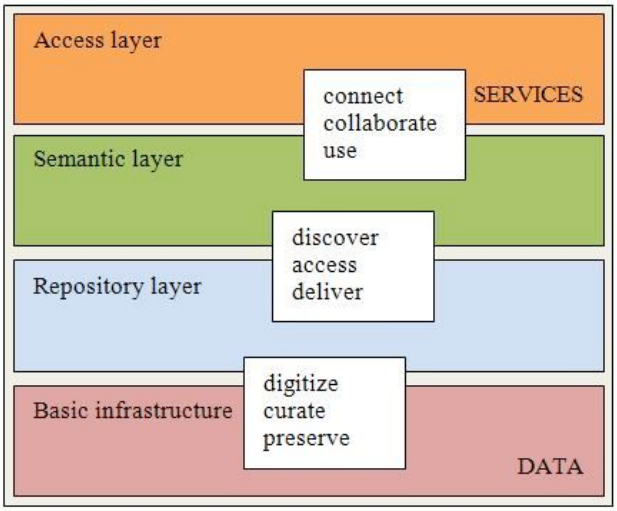

Figure 1: DARIAH technical layers

The technological infrastructure of DARIAH will be a service oriented architecture (SOA) and will be based on the Fedora repository architecture. It will adopt technologies like SOAP and REST, it will make data available for PMH harvesters and ontologies will be employed to make these data interoperable across different data centres. The adoption of an open architecture is imperative since in the virtual repository layer, there exist providers with a plethora of repositories such as Fedora, eSciDoc, EASY and other national repository systems. The proposed architecture must be able to virtually link all these repositories together. The big advantage of a SOA architecture is that users can implement their own services as modules that can be used to expand the available services. This approach will ensure that new partners will be free to implement their own ideas into the DARIAH architecture. Furthermore, one can combine these services by building higher level services on top of existing ones thus expanding the available services exponentially.

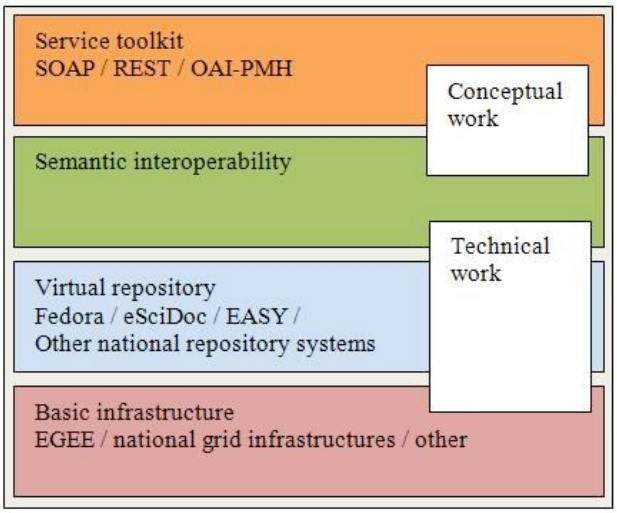

Figure 2: Architecture of the proposed DARIAH prototype

3.5 Conceptual modelling

Research in the arts and humanities is characterized by meaning extraction from often incomplete or "fuzzy" datasets. The interpretation and re-interpretation of the results constantly changes not only when new data is available but also when hypotheses change. Thus, this research practice is characterized by constant change. In order to design a system that will ease this process, data must be semantically annotated in such a way that researchers can extract knowledge more easily. This requires a mechanism for semantic annotation of the data along with an interoperability layer that will link different datasets together since they are distributed throughout Europe.

In order to provide a high quality interoperability layer that will allow a coherent virtual integration among the various repositories, an interoperability framework must be proposed that will act as a common reference across the various DARIAH members. Because of the diversity of formats and different metadata standards found in the arts and humanities, the need arises for linking those formats together and/or proposing new common standards for newly created data centres.

Conceptual modelling is especially important and complicated in the arts and humanities because knowledge is created through hypotheses that are based on facts and other hypotheses that constantly change. Thus, a proper knowledge representation is required to both preserve knowledge and aid researchers in their work.

The goals of the conceptual modelling will be three-fold:

- identify the different processes among the various scientific domains and map them to the existing data centres;
- assess digital assets at the various data centres and establish recommendations on their representation; and
- evaluate the process and object models bearing in mind the heterogeneous environment and the various user groups / scientific domains.

A robust and well-designed semantic layer will help better connect the different data centres within a common metadata framework. The use of mappings between metadata schemas will also be explored in order to enhance the work of harvesters. Furthermore, semantics will help address the issue of intellectual content preservation in addition to bitstream data preservation - a major problem in the area of the arts and humanities.

4. CONCLUSIONS

In conclusion, DARIAH is building a multinational research infrastructure for the arts and humanities intended to support researchers from all European countries. As more countries join DARIAH, more data and more tools will be available thus increasing its value. According to the current schedule, DARIAH will be ready for construction by 2010 (that is when the preparatory stage ends). By that time, new technology, data and human (partners) infrastructure will be ready.

References from websites:
http://www.dariah.eu