Partnerships, relationships and associated initiatives - Towards a strategic plan for DARIAH

Laurent Romary

DARIAH-EU

Partnerships, relationships and associated initiatives — Towards a strategic plan for DARIAH

R EU 4.3.1

Version - 08 December 2011
VCC – VCC4 Advocacy, Impact and Outreach
Responsible Partner – State and University Library Goettingen

DARIAH-DE
Aufbau von Forschungsinfrastrukturen für die e-Humanities

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

Project: DARIAH-DE: Forschungsinfrastrukturen für die e-Humanities
BMBF Fund number: 01UG1110A bis M
Duration: März 2011 bis Februar 2014

Autors: Laurent Romary

1. Overview

This report is a subjective attempt to provide an overview of the environmental factors that may impact on developing a strategy for the future of DARIAH. To this end, we revisit the core objectives of DARIAH and use this analysis to better understand how DARIAH may be further integrated in a rapidly evolving European context. Indeed, we have observed in the recent years two major trends that directly impact on the establishment of DARIAH as a (e-)Research infrastructure in the humanities:

- A remarkably growing interest for digital humanities in nearly all research domains in the humanities at large;

- The development of generic eScience initiatives, usually anchored on strong political activities at national and European levels.

In this context, DARIAH faces a double challenge of a) possibly having difficulties to focus on precise objectives in terms of service provision, because of the large number of communities to address at the same time and b) to spend most of its energy in liaising (or concerting) with other ongoing (maybe ephemeral) projects and/or political bodies which see their own activities as related to ours.

To circumvent these difficulties, we first define some core strategic orientations based on the one hand, on the essential steps in the digital scholarship workflow and, on the other hand, suggesting a strong data oriented perspective for DARIAH, which may help us identify where we have a real role to play and where we need to collaborate with others. Whereas we perfectly acknowledge that the technological context is also an important factor to consider, and will indeed appear at several places within our presentation of the institutional landscape, we do not provide here any corresponding background analysis[1].

The second half of the report focuses on reviewing ongoing initiatives at European level and making recommendations as to how DARIAH should, or should not, collaborate with them. Such collaborations first and foremost rely on the identification of real complementaries from which DARIAH can gain benefits.

We hope that this report will help both the board of directors and the DCO to take efficient decisions as to our priorities and expect this report to be a background document for a future DARIAH roadmap (resp. strategic plan).

2. A quick history of DARIAH and consequences on the strategy

The history of DARIAH started in January 2006 when representatives from four European institutions[2] met to identify how they could join efforts in providing services to the research

[1] This should indeed be a topic addressed in the strategic plan of DARIAH

[2] Sheila Anderson, director of AHDS; Peter Doorn, director of DANS; Laurent Romary, director for scientific information at CNRS; Ralf Schimmer, representing Harald Suckfuell, in charge of scientific information for the Max Planck Society.

communities they served, with a strong focus on the humanities. The idea behind this meeting was to go towards a consortium of institutions, which would ensure long-term sustainability of the underlying infrastructure and a strong political voice towards the EU. Each institution had in charge a national role in coordinating or developing digital services in the humanities and could thus speak with a national perspective.

DARIAH was thus put together as a top-down initiative of scientific information institutions, each having a duty to provide services to their respective research communities. In a way, one could say that DARIAH was conceived without the research communities themselves. But I would analyse this as a very beneficial factor since it gave us flexibility with regards the actual scholarly coverage of our activities.

Finally, we can also draw lessons from the personal history of the very institutions that contributed to the establishment of DARIAH, observing how it is to conceive sustainable infrastructures for research:

- The AHDS was disbanded just two years after the initial DARIAH meeting, on the assumption that the services it provided did not justify the cost, in a context where scholars had supposedly gained their digital autonomy;

- Changing Director General of the CNRS has resulted in dropping scientific information from the actual priorities of the institutions. It took several years before the national infrastructure could actually start offering real services to the community and has not yet reached its full maturity;

- The existence of DANS was questioned on several occasions by its mother institutions and it took a lot of energy to demonstrate its utility within the human and social sciences landscape in the Netherlands;

- The Max-Planck Society heavily reduced the initially high ambitions of the Max Planck Digital Library to concentrate its scope on basic services to research libraries of the institutes.

Beyond these prominent examples, it would not be difficult to gather up many other difficulties[3] that have marked the history of infrastructures in the humanities. Interestingly, in such a hazardous context where basically all actors involved underwent major difficulties, DARIAH has managed to go through all hurdles and actually moved forward to a stage where it is about to become one of the most stable components in the eHumanities landscape. In a way, this can be seen as a success and understanding why, namely it has gained from the average positive momentum in favour of digital humanities, may also make us confident for the future.

Still this should not prevent us from analysing the reasons why it is so complex to establish an infrastructure for the humanities, a problem that can be construed along the following lines of tension:

[3] Impossibility to find a sustainable structure for the DHO in Ireland, governance of Adonis in France, fading out of ELSNET, etc.

- A research infrastructure in the humanities should be able to provide concrete short-term services that may give it a scholarly recognition;

- It should have at the same time a clear vision about its general objectives that will guide the evolution of the infrastructure over the years;

- It should gain institutional support for both aspects and demonstrate that it matches the strategic objectives of its funders;

- It should elicit how much it complements local initiatives to provide technical support to researchers;

- It must show its value for money in the sense that scholars do not see the infrastructure as consuming budget that would otherwise go to research.

These elements potentially apply to all scientific domains. Still, the humanities represent an even more complex environment because, on the one hand, of its highly fragmented scholarly structure, and, on the other hand, its low technical literacy. Whereas DARIAH has managed to gain institutional recognition at European and national level, it is its capacity to relate to this complex community of users that will be a real measure of its success.

3. A user's oriented view on DARIAH

In the short term, DARIAH will have to provide simple services that correspond to the expectations of its users. By *users*, we actually mean here the now quite large community of scholars who have to deal with digital content, whether or not they master the technical background related to the creation or the management of these resources.

The adequacy of services to expectations will rely a great deal to the level of digital awareness that scholars actually have, which in turn may change rapidly in the coming period. We will thus have to face the difficult situation of answering changing needs, as well as having to deal with a very heterogeneous community ranging from early adopters of digital techniques to completely computer illiterate scholars.

In this context, simple services can be characterised by the fact that, on the one hand, they can easily be adapted to new usages and new demands, and, on the other hand, that they are closely anchored on the basic processes related to the scholarly research process, seen here from the point of view of working with digital data or sources.

In the remaining section we will briefly go through what we think are the essential aspects of the scholarly research process and identify the services that DARIAH should prioritize accordingly[4].

[4] see also: "Reinventing research? Information practices in the humanities", Research Information Network report, April 2011.

3.1. Finding and quoting digital sources

The most important step for introducing a virtuous digital circle in the humanities research process is to provide scholars with the means to identify and locate existing digital sources, which they can explore, study, and finally quote in their own research. To help achieving this, DARIAH must quickly deploy services along the following lines:

- Discovery portals that acts as single entry points to existing online resources;
- Recommendations on the optimal web searchability (cf. what to provide access to, which entry points, in the context of sitemaps, for instance) to be widely disseminated within the research communities in the humanities, but also to funding agency for them to integrate these in their call for projects;
- Interfacing in such portals of exemplary resources and archives in targeted scholarly domains (this could be based on the direct output of national and European initiatives such as EHRI or CENDARI) to foster the use of online resources;
- Recommendations concerning the citation of sources in the humanities, combining appropriate reference to the source as well as to its creator.

3.2. Creating and annotating digital content

The second important step in going digital is to be able to create one's own digital assets out of existing primary analogue sources, or annotate (resp. enrich) existing digital sources. In this domain, DARIAH should prioritize the provision of services that help scholars in becoming rapidly autonomous in working in a digital environment; in particular, we need to focus on the following core services:

- guidelines for the elementary creation of digital sources ("starter set") – together with appropriate reference examples[5];
- provision of editors in a box that point to a reduced set of environments that can be directly installed or used online to create scholarly relevant digital content;
- advertise and/or organize training workshops all over Europe so that scholars or newly hired students can be trained and gain quick autonomy.

3.3. Preserving and disseminating content

Once digital assets have been created, it is essential that researchers do not wonder on the way they can make them widely accessible, while being trustful in the way the resource will be used and cited. To this end we recommend to have the following priorities on the DARIAH short-term agenda:

[5] In the case of textual resources, we would for instance point to the TEI by example page (http://tbe.kantl.be/TBE/) and contribute to its maintenance

- Provide transparent services to facilitate the unique identification of participants in a research. In this domain, we should take an early part in the Orcid initiative;

- Provide an online service for research asset PIDs. We should in this context strengthen our relationship with EPIC[6] and DataCite;

- Provide recommendations concerning a core set of meta-data they have to associate to their resources to make it useful and citable for other researchers (identification and documentation of the source, sampling strategy, description of the digitization added-value, proper identification of responsibilities and affiliations)

- Provide recommendations on simple licensing schemes to be associated to digital assets. Basically, we should advocate a simple CC-BY license for all publicly funded projects to which no further constraints apply (cf. open access discussion below);

- Offer an early service for archiving and hosting generic digital resources (images, XML transcriptions). This should not only be implemented through an archive-in-a-box strategy, but also through the offering of real hosting services (e.g. exist farms)

3.4. Additional service related to publications

Although scholars may not request it right from the onset, DARIAH needs to provide the necessary expertise concerning the management of publications in the humanities. We thus recommend that the following aspects be pursued at an early stage of the creation phase of DARIAH:

- provide advice (even proselytise) on open access and in particular the early deposit of scholarly paper in a publication repository;

- recommend appropriate editorial platforms for the creation of new journals or the migration of existing ones towards scholarly respectful models;

- provide a critical study of existing scientific social networks and in particular identify their actual capacity to relate to publication archives.

3.5. Conclusion

In the short term, we thus recommend that DARIAH be efficient towards scholars by offering modest but targeted services. DARIAH should also be able to boast this modesty towards external actors (members, EU) and show how it is part of a long-term strategy of developing an infrastructure for the humanities.

We think that the adequate provision of a sound portfolio of such needs-oriented services will facilitate the development of more ambitious digital humanities environments. In particular, such basic services should be thought of as preliminary building blocks towards the creation

[6] EPIC – the European Persistent Identifier Consortium; http://www.pidconsortium.eu

of more elaborate virtual research spaces[7] based on a more data-oriented perspective as outlined in the next section.

4. A data oriented view for DARIAH

4.1. Towards a stable perspective for DARIAH

Contrary to the short-term strategy, the long-term vision of DARIAH should somehow take distance with a purely user-centric view. Indeed, and given the speed at which the technological awareness evolves at present, it is nearly impossible to anticipate what the scholars will actually request from a digital infrastructure in the humanities within even the next five years. In this context, our duty is to create a sound and solid background that is likely to ensure, on the one hand, the stability of digital assets in the long run, and, on the other hand, the development of a wide range of even unanticipated services to carry out research on these assets.

This data-centred strategy echoes various reports and statements that have been issued recently and in particular "Riding the wave", which has put the management of scientific data very high on the EU commission's agenda. This report stresses the importance of a long-term strategy concerning the management of scholarly data in all disciplines, which comprises both technical aspects (identification, preservation), editorial (curation, standards) and sociological (openness, scholarly recognition).

In this section, we go even further by considering that a *data-centred strategy* for DARIAH will secure a long-term vision both in terms of the deployment of future services, but also in the way we will organise our collaborations with other initiatives, in particular in the cultural heritage domain. To do so, we outline the role of *digital surrogates* in digital humanities as a core concept for data management and explore the actual consequences of such a vision.

Note: we will speak henceforth of *primary sources* as covering all types of documents or information sources that may be used as testimonial information to support a research. This wide notion typically covers objects such as: manuscripts, artefacts, sculptures, recordings, statistical data, observations, questionnaires, etc.

4.2. Surrogate – definition

We define here a surrogate as an information structure intended to identify, document or represent a primary source used in a scholarly work.

Surrogates can take a wide variety of forms ranging from metadata records, scanned image of a document, digital photographs, transcription of a textual source, or any kind of extract or transformation[8] on some existing data.

The notion of surrogate is at the core of digitally based scholarship since it is intended to act as a stable reference for further scholarly work, in replacement – or in complement – to the

[7] cf. Romary, "scientific information"
[8] E.g. the spectral analysis of a recorded speech signal

original physical source it represents or describes. By definition, it should always contain some minimal information to refer to the source(s) it is based upon.

In turn, a given surrogate can act as a primary source for the creation of further surrogates, for instance with the purpose of consolidating existing information or creating complex information structures out of different sources.

As a consequence, a network of digital surrogates will reflect the various steps of the scholarly workflow where sources are combined and enriched up to the point that the results can be further disseminated to a wider community. Indeed, we do not anticipate a flat space of digital surrogates, but a complex data space integrating the various evolutions that such surrogates may encounter.

In the remaining sub-sections, we will analyse the consequences of having surrogates at the centre of our perspective concerning digital humanities and contemplate the impact on our delivery of services.

4.3. Data management issues

A coherent vision on a unified data landscape for humanities research should be based upon a clear policy in the domain of standards and good practices. In particular, DARIAH should not only make strong recommendations as to which standards may optimize the sharing and use of digital surrogates in research activities, but it should also contribute to shaping the standardization landscape itself by supporting participation in the corresponding working groups and organisations.

Acknowledging the fact that other communities of practice (publishers, cultural heritage institutions, libraries) may have different agendas and practices in the domain of standards, we should also spend effort on defining interoperability conditions between heterogeneous worlds (e.g. EAD – TEI relationship).

Finally, we need to assess the consequences of an extremely distributed network of potential data sources, ranging from individual scholars to major national libraries. Providing guidance to individual users as to how one can navigate and use digital assets in such a heterogeneous data landscape will be a major challenge for DARIAH. To this end, the evolutionary surrogate model outlined above will be essential in defining conditions aggregating identifiers, versions and enrichments of digital assets.

4.4. Technical issues

Whereas the data landscape will heavily rely on third party providers (cf. political issues below), the development of a data-based strategy for DARIAH will impact on some of our technical priorities in the short term as well as the long-term. We can outline the three levels where DARIAH should put some specific efforts as follows:

- define a repository infrastructure for scholarly data where researchers can transparently and trustfully deposit their productions. Such an infrastructure should be in charge of maintaining permanent identification and access, targeted dissemination (private, restricted and public) and rights management. In this context we should

identify the optimal level of centralization that allows efficiency, reliability and evolution[9];

- spend meaningful effort in defining and implementing standardized interfaces for accessing data through such repositories, but also through third-party data sources. The objective of such interfaces must be to make it easy to derive simple services in the domains of threading, searching, selecting, visualising, importing data;

- experiment the development of agile virtual research spaces based on such services that allow specific research communities to adopt their own data-based research workflow while being seamlessly integrated in the DARIAH data infrastructure[10].

4.5. Licensing issues – open access strategy

The evolution of digital humanities towards a complex and interrelated data landscape will require having a strong policy as the legal conditions under which each data asset will actually be disseminated. To tackle such issues, there are indeed two different, but probably complementary, points of view:

- the ideological perspective brings the debate to identify that each scholarly production, financed by means of public funding, is by essence a public good[11]. This should lead us to defend a generalised open access strategy for all scholarly productions;

- a pragmatic view, informed for instance by the experience of the genomic domain, acknowledges that it is unpractical, even impossible, to do data based research within a data landscape bearing heterogeneous reuse constraints and/or licensing models.

All in all, the core reasons why we have no choice but work towards an open data space are well identified and boil down to the issues of[12]: more efficient scientific discovery and learning, allows other researchers—and the wide public—access to raw numbers, analyses, facts, ideas, and images that do not make it into published articles and registries, better understanding of research methods and results, more transparency about the quality of research, greater ability to confirm or refute research through replication.

To achieve this in the humanities, DARIAH should give guidance on two complementary aspects:

- advocate an early dissemination of digital assets, explaining that the fear of compromising academic primacy should be put in perspective with the potential gain in extra citation to the data itself;
- encourage the systematic use of a Creative Commons license CC-BY, that basically supports systematic attribution (and thus citation) of the source.

[9] cf. Romary, Laurent and Chris Armbruster (2010), "Beyond institutional repositories", International Journal of Digital Library Systems 1, 1 (2010) 44-61 — http://hal.archives-ouvertes.fr/hal-00399881; for a discussion of possible models.

[10] See Romary (tbp), "Scientific information"

[11] which, in the humanities strongly overlap with the notion of "scientific good" (as opposed to the case of bio-medical research for example)

[12] Freely adapted from a personal communication from Trish Groves, Deputy editor, BMJ (British Medical Journal). Note here that although the words used are clearly referring to hard sciences, they seem to perfectly fit what we could dream of in the human sciences.

Taking again example from what happened in the genomic field, CC-BY should be preferred to less restrictive (e.g. CC-0) licenses since attribution lies at the centre of the academic process, and of course to more restrictive ones, which are either inapplicable ('share-alike') or preventing a wide use of the digital asset ('non-commercial').

Besides, DARIAH should apply this scheme to itself in such a way that all documents and data produced specifically within DARIAH (or DARIAH affiliated projects) should be associated with a CC-BY licence.

DARIAH should also contribute to large scale negotiations with cultural heritage partners (libraries, museums, archives, or representatives thereof) to ensure global agreements through which as light licensing schemes as possible are applied to the data made available to scholars.[13]

Finally, it makes no sense to have strong views on the free dissemination of scholarly data if no reference is made to publications proper. We thus suggest that DARIAH should indeed have a policy with regards publications, which could be easily implemented through some of the following actions:

- Only refer on the DARIAH web site to publications which are available through a publication archive
- Provide a weekly newsfeed on the DARIAH web site pointing to some exemplary papers available online
- Facilitate the move towards open access publications with the support of technical platforms such as revues.org
- Participate in initiatives facilitating access to publications by content and better linkage between publications and research data
- In the long run, facilitate the creation of a network of HS publication repositories

4.6. Political issues

The global strategy put forward above concerning the management of digital assets/surrogates in the humanities is by far too complex to be dealt with alone within DARIAH. It is of strategic importance that we articulate our activities in this domain in strong collaboration with the various actors of the data continuum we have identified. In particular, we need to consider how much data potential providers (cultural heritage entities, libraries or even private sector stakeholders such as Google) could become partners in creating the seamless data landscape we are all dreaming of. Such partnerships should be articulated along the following lines:

[13] To cite here the final conclusions of the High Level Expert Group on Digital Libraries, under the auspices of commissioner Reding: "public domain content in the analogue world should remain in the public domain in the digital environment."

- General reuse agreements[14] that would systematically apply when scholars require access to sources available from data providers, comprising usage in publications, presentation on web sites, integration (or referencing) in digital editions, etc.;

- Definition of standardized formats and APIs that could make the access to one or the other data provider more transparent;

- Identification of possible scenarios through which the archival location of version of records are clearly identified and, by the same token, enrichment mechanisms are contemplated[15].

4.7. Conclusion

DARIAH should contribute to excellence in research by being seminal in the establishment of a large coverage, coherent and accessible data space for the humanities. Whether acting at the level of standards, education or core IT services, we should keep this vision in mind when putting priorities as to what will impact the sustainability of the future digital ecology of scholars. Above all, such a strategy should directly influence the way we will advocate DARIAH towards funding or supporting institutions, and also how we will manage our collaboration schemes with other initiatives in Europe and worldwide. These last two aspects will be the main focus of the remaining sections of this report.

5. Strategy

Independently of its initiators in 2006, DARIAH is at the service of the political forces that have brought it to life. Indeed, DARIAH would have never existed if the initial ideas that lead to the proposal as infrastructure on the ESFRI roadmap had not been shaped step by step to match the expectations of various institutional actors. In the following sections we thus quickly outline the opportunities and challenges resulting from our positioning towards various stakeholders from countries to European organisations.

5.1. Countries

Because of the ERIC organisation as essentially a consortium of member states, the ultimate decisional entities for DARIAH will always be the countries themselves. It is thus essential to have a good understanding on how DARIAH should actually articulate its policies with regards the national roadmaps (whether explicit or implicit) in e-Infrastructures for the humanities.

In this context, the following sections provide a quick overview of the situation within the various DARIAH members and try to identify the main challenges that we may have to face.

[14] We should take as backgournd document the "The Europeana Licensing Framework", issued in 2011, see http://creativecommons.org/weblog/entry/30609

[15] For example, TEI transcriptions made by scholars could be archived back in the library where the primary source is actually situated

Note: in this initial version of the report, only the core DARIAH partners are beings mentioned. The next iteration of the report in April will contain a more comprehensive overview of the European partnership.

5.1.1. France

Background: In 2004, the CNRS established Adonis as a national infrastructure in the Humanities, which, despite the continuous governance problems, has been steadily the contact point for DARIAH in France.

Analysis: The current landscape in France maybe a little complex to understand. We indicate here the main contributors to the establishment of DARIAH:

- Adonis (100% CNRS) will host DARIAH in France and provide services associated to its own activities (e.g. the Isidore portal), in particular in co-leading (with the Netherlands VCC3)

- Cleo (joint structure between CNRS, EHESS, and three universities) is a national institution providing editorial services (hosting journal, scholarly blogs, etc.)

- Corpus is the national coordination of digital humanities communities in France; it will also serve as contact point for CLARIN

- ABES is the national institution providing library services to Universities. It offers a central catalogue and several authority lists

- Inria is the national research institution in computer science

Challenges: In the context of a strong support from the French Ministry of Research in favour of DARIAH, the French contribution remains to be more coordinated, with a clear governance on the French level.

5.1.2. Germany

Background: Digital Humanities activities in Germany have gained a special momentum with the funding of the TextGrid project in 2006.

Analysis: The DARIAH-DE project, funded by the BMBF has been the first national roadmap to be actually implemented in February 2011. Organized as a strong multi-disciplinary partnership, its work programme closely matches the DARIAH-EU general organisation in VCCs.

Challenges: the biggest challenge for DARIAH will be a) to bring more EU partner in phase with the progress made in Germany and b) to take up early results from DARIAH-DE and make these concrete services at EU level.

5.1.3. Denmark

Background: Denmark has been an early participant in the DARIAH initiative, with the Nordisk Forskningsinstitut bringing it its background philology and humanities computing at large.

Analysis: In the context of establishing its national roadmap, Denmark has put together a single national infrastructure acting for both CLARIN and DARIAH. In this context, Denmark will lead VCC2 together with Ireland.

Challenges: The national centre in digital humanities will only be effective in early 2012 and it is as of now not completely clear as to how the country will organize its own activities in relation to the various actors (Universities, but also the Royal Library). There is a very high potential for contribution which we need to help being elicited.

5.1.4. The Netherlands

Background: The Netherlands has had the most stable national infrastructure in humanities and (partly) social sciences with DANS, as well as a traditionally strong presence in CLARIN.

Analysis: The various stakeholders of the CLARIN and DARIAH initiatives in the Netherlands have submitted a joint project ("CLARIAH"), requesting a fair amount of funding for the construction phases of both infrastructures, putting forward several common building blocks.

Challenges: There is a clear risk that going towards an integrated projects with CLARIN, where the latter is actually hosted in the country may lead to difficulties in identifying the specific contributions to DARIAH. On the other hand, it can be a very important experiment for shaping the future European landscape from an ESFRI/ERIC perspective.

5.1.5. Ireland

Background: In the context of its national funding program PRTL4, Ireland funded the establishment of a national infrastructure in digital humanities: DHO (Digital Humanities Observatory). This structure is lacking a sustainable funding scheme and will disappear in 2012.

Analysis: Ireland (through the IRCHSS[16]) has agreed to participate in DARIAH and co-lead VCC2, on the basis of the existence of several third party national (PRTL5 program) and European (CENDARI, CULTURA).

Challenges: The main challenge in the construction phase of DARIAH will be for Ireland to gain its own dynamics with a) a sustainable contribution and b) a proper integration of the various Irish partners.

[16] Irish Research Council for the Humanities and Social Sciences

5.1.6. Austria

Background: Austria has had since many years a good visibility in digital humanities, in particular through the developments carried out in the academy.

Analysis: Through its co-coordination of VCC1, Austria has positioned itself as strongly technology oriented. Besides, a close coordination of CLARIN and DARIAH exists there, with a single leadership for the two corresponding networks.

Challenges: There seem to be a strong support for DARIAH at all levels in Austria and the main challenge will be to provide the adequate visibility within the ERIC, in comparison to the "bigger" partners, probably through the identification of focused services and competences.

5.1.7. United Kingdom

Background: The UK has somehow been a pioneer in the domain of infrastrcutrues in the humanities by establishing the Art and Humanities Data Service (AHDS) in 2000, as a joint initiatives from the AHRC and and JISC. Despite its unexpected closure in 2007, the AHDS has served as a model or reference for many similar initiatives in Europe.

Analysis: The UK has not yet managed to have a political decision be taken as to its participation in the DARIAH ERIC as full member. Still, the strong presence of King's College in DARIAH activities, as well as the existence of many national initiatives in the Digital Humanities domain, will make UK a strong component of the future development of DARIAH

Challenges: Because of the instrinsic structure of the academic field in UK, there remains to create a strong network of partners that will contribute to the DARIAH activities.

5.2. ESFRI

The initiative taken by the European commission to launch the European Strategic Forum on Research Infrastructures (ESFRI) has been a major initiative, in particular because it has put the human and social science on the infrastructure agenda.

At this stage, and without going deeply into the history and arcanes of the ESFRI process, it is important to keep the following observations in mind:

- The initial ESFRI process has relied on a quite random bottom up process, which has not always ensured an adequate representativity of the scientific stakeholders and has often been the sresult of opportunistic situations;

- Countries have hardly been in the loop of the design of the initial ESFRI roadmap which has resulted in difficulties in having them design coherent national roadmaps;

- The quite rigid funding scheme for preparatory phases has not allowed each project on the roadmap to actually develop its implementations plans in accordance to the actual priorities or expectations of the corresponding communities.

As a whole, it is not so clear whether the ESFRI process will manage to keep its current structure and missions in a context where more and more initiatives from the roadmap will actually become ERICs. As a whole, the driving forces for the future of ESFRI will relate to its capacity to deal with the following challenges:

- Contributing to the evolution of the infrastructure landscape in the humanities by providing a coherent plan for shaping new and old initiatives in the coming 5 or 10 years;
- Taking into account the increasing involvment of member states in defining their priorities (and roadmaps);
- Articulating this with a sound European funding schemes for infrastructure that will continuously feed innovation in the domain of services.

5.3. European funding organisations

5.3.1. ESF

The European Science Foundation is a consortium of European research funders, who have a greed to launch joint call for projects, which, when accepted, are further financed at national level. ESF does not intend to support large scale initiatives, but mainly focuses on financing workshops, grants or transnational networks. The NeDiMAH project is actually funded in this framework.

ESF has expressed recently some interest in adding the research infrastructure issue[17] on its agenda although the way this will actually be implemented is not currently clear.

It is important for DARIAH to resume contact with the ESF to identify how far a collaboration could be put together. The vision here would be for ESF to finance activities which may become direct imputs (in terms of new research questions or new methodological frameworks) to the services that DARIAH will offer in the future. It should be noted here that a more ambitious scheme had been put forward with the preparatory phase of DARIAH, which representatives of the commisions had encourage us not to pursue.

The future merge between ESF and Eurohorcs as Science Europe, should probably provide us with a more adequate interlocutor.

5.3.2. HERA

HERA (Humanities in the European Research Area) is a partnership between 21 Humanities Research Councils across Europe and the European Science Foundation (ESF), with the objective of "establishing the humanities in the European Research Area and in the European Commission Framework Programmes".

In 2009, HERA launched its first Joint Research Programme (HERA JRP) with a total budget of €16.4 M. It is currently funding 19 transnational projects under two themes "Cultural

[17] with the publication of its report: ESF Science Policy Briefing 42: Research Infrastructures in the Digital Humanities: http://www.esf.org/publications/humanities.html

Dynamics: Inheritance and Identity" and "Humanities as a Source of Creativity and Innovation".

Because of the natural thematic proximity, and the presence of some DARIAH participants in the HERA decision process, DARIAH should soon or later officially relate its activities to the funding priorities of HERA. Given the prospect of having a stronger presence of the humanities in the next EU framework program and the forthcoming reorganisation of the ESF, we may want to wait until mid-2012 to actually take a more official stance.

5.3.3. ERC

The European Research Council is funding innovative research projects centered on a leading scientific figure. It has been very succesful in the recent years in providing a funding scheme that is complementary to the more result oriented framework offered by EU programs. Although no contact has been established so far, we would consider useful to provide more awareness about DARIAH to them.

5.4. European Union

As an ERIC, DARIAH will become a consortium of national members and as such would not have theoretically to establish direct contact nor reporting with the commission. Still, the commission has been so seminal in concretizing the ESFRI roadmap, that we should keep very close relations with them. This is all the more important that the EU is likely to pursue its policy of funding infrastructural or research projects that work closely with established research infrastructures. We should thus make that, on the one hand, we provide regular managerial reports eliciting the services we have deployed and identifying domains where specific investment should be made, and on the other hand, we spend the necessary time to help the EU define a coherent vision for e-Infrastructures in the humanities.

A future version of this report will specifically address this issue in detail.

5.5. What about scholars?

Individual scholar will have very limited capacity to impact on DARIAH if we do not proactively try to interact closely with communities, with further joint projects (cf. affiliated projects) deployment of concrete services or targeted dissemination activities. The two complementary approaches (short-term user oriented and long-term data oriented) that we advocated in earlier sections in this report should indeed be the way to relate communities and political bodies at national and European levels.

6. Environmental analysis

6.1. A reading guide to existing initiatives

The variety of initiatives and projects that are currently taking place at European level makes it very difficult to provide a coherent collaboration strategy. In order to make the global picture more legible we have considered each initiative according to their proximity to DARIAH in terms of thematic interest, institutional anchoring and intended user communities. From this analysis, we have identified the following main categories as basis for our collaborative framework:

- *Affiliated projects*: they are initiatives whose activity and even existence are closely dependant from the technical and political background established by DARIAH. We suggest to have a strong and homogeneous collaborative framework with them;

- *Sibling initiatives* correspond to projects or consortia that serve similar topics or communities as DARIAH, and with which we need to concertate on a wide range of issues;

- *Cultural heritage initiatives* will always be at the interface of our remit as they will provide access to data sources that researchers in the humanities will want to study further;

- *Technological initiatives* are projects or consortium that are working towards the implementation of specific IT services. Experts involved in VCC1 will typically have to follow the work carried out there;

- Finally, the *larger circle* are initiatives we need to be aware of, but with whom we have identified no priority to collaborate closely.

6.2. Affiliated projects

6.2.1. EHRI

The main objective of the Euroepan Holocaust Research Infrastructure (EHRI) is "to support the European Holocaust research community and help initiate new levels of collaborative research through the development of innovative methodologies, research guides and user-driven transnational access to research infrastructures and services. To this end, EHRI proposes to design and implement a Virtual Research Environment offering online access to a wide variety of disparate and dispersed key Holocaust archival materials and to a number of online tools to work with them. Building on integrating activities undertaken over the past decades by the 20 partners in the consortium and a large network of associate partners, EHRI sets out to transform the data available for Holocaust research around Europe and elsewhere into a cohesive corpus of resources."[18]

EHRI is a very important showcase for DARIAH as it is the opportunity for us to experiment how we can articulate our workplan with the specific requirements of a user community. We also need to make sure that DARIAH is visible in relation to EHRI and bring continuous feedback on their progress within our technical (VCC) meetings.

6.2.2. CENDARI

The relation between the EU CENDARI project and DARIAH is clearly stated in its description of work: "The Collaborative EuropeaN Digital Archive Infrastructure (CENDARI) will provide and facilitate access to existing archives and resources in Europe for the study of medieval and modern European history through the development of an 'enquiry environment'. This environment will increase access to records of historic importance across the European Research Area, creating a powerful new platform for accessing and

[18] Excerpt from EHRI's description of work

investigating historical data in a transnational fashion overcoming the national and institutional data silos that now exist. It will leverage the power of the European infrastructure for Digital Humanities (DARIAH) bringing these technical experts together with leading historians and existing research infrastructures (archives, libraries and individual digital projects) within a programme of technical research informed by cutting edge reflection on the impact of the digital age on scholarly practice."

CENDARI has even clearer statements concerning the reuse and contribution to the DARIAH technologies, which makes it theoretically an optimal affiliated project, with which we should even not have to sign a Memorandum of Understanding (MoU). Its kinship to EHRI (similar king of organisation between researchers and infrastructure specialists; interfacing archives) will also facilitate the creation of synergies with DARIAH.

6.2.3. DASISH

The DASISH project results from an initiative of the EU commission to make the various infrastructures in the domain of social sciences and humanities work together in identifying common methodologies and technical components that could be jointly deployed by all of them. DARIAH is participating in this project through three of its technical contributors, namely King's College, University of Göttingen and DANS.

The focus of the project covers the general domains of data quality, data archiving, data access and legal and ethics, but a closer look at the technical annex show how much there are specific activities intended uniquely for social science surveys (e.g. data quality) and which may not benefit directly to DARIAH. There is also a strong influence of the technical work carried out at MPIPL in Nijmegen, which, even if perhaps fit for our CLARIN colleagues, may not be of direct interest to DARIAH.

All in all, the recommendation for the DARIAH related partners would be to take the opportunity of the DASISH project to deploy some concrete services that may directly impact on our capacity to serve our communities (in particular in the two core work packages "Data Archiving" and "Shared Data Access & Enrichment").

6.2.4. NeDiMAH

Based on the model initiated by the *AHRC ICT Methods Network*[19] in UK, NeDiMAH[20] is an ESF project that aims at being an exchange forum on the use of computer-based methods in the humanities. The partnership integrates some major European academic sites with strong experience in digital humanities and covers the essential topics in the domain (geo-temporal information, visualisation, ontologies and "linked data", edition, development and usage of digital collections).

As such, NeDiMAH is an essential background activity for DARIAH. It is likely to bring essential expertise to VCC1 (existing technical environments for the management of digital collections; prospects for the development of new environments), VCC 2 (stable methods that

[19] Funded April 2005-March 2008; see (the project homepage is now blank!) http://www.kcl.ac.uk/artshums/depts/ddh/about/affliate/ahrcict.aspx

[20] Network for Digital Methods in the Arts and Humanities - http://www.esf.org/index.php?id=8752

should be further disseminated in the humanities) and VCC 3 (recording good practices which DARIAH can take up as a reference). Conversely, DARIAH can act as a sound box for the various achievements of NeDiMAH, since we have the capacity to outreach the humanities community widely.

NeDiMAH would thus contribute to providing the methodological layer for DARIAH and would be at the forefront of the projects with which we should have a clear collaboration scheme. Such a scheme (MoU) could be articulated along the following action lines:

- endorsement of NeDiMAH as a DARIAH affiliated project;
- dissemination of NeDiMAH results under the auspices of DARIAH;
- integration of NeDiMAH results within the DARIAH work-program.

NeDiMAH is also of strategic importance for DARIAH as it may bring in countries that would not necessarily be in the position of being DARIAH members, but still would want to participate in our activities.

6.2.5. Code name - DiXit

As a complementary layer to the NeDiMAH network, "DiXit"[21] is a Marie-Curie proposal dedicated to foster training activities in the domain of digital editions. This application has been explicitly initiated by the DARIAH management and taken up by the colleagues in Cologne (P. Sahle). If successful, it should be the first in a series of Marie-Curie network that should be focused on specific methodological fields in relation to DARIAH.

6.2.6. CULTURA

CULTURA[22] is an EU funded project exploring the technical challenges in dealing with large and complex cultural corpuses of digital data. The project is based on the two following collections:

- the 1641 Depositions, held in Trinity College Dublin,
- the IPSA Digital Herbal Archive, held in the University of Padua,

on which natural language processing techniques[23] and experimentation on possible interfaces should enhance usage to scholars.

Even if no reference to DARIAH is made in the project descriptions that are currently available, CULTURA has been put forward as part of the contribution of Ireland to DARIAH.

It is not clear at this stage if we should make it an affiliated project, but discussion with our Irish colleagues should take place in this respect.

[21] Digital Scholarly Editions Initial Training Network, which is to be submitted within the framework of the 7th Framework Programme FP7-PEOPLE-2012-ITN of the European Community
[22] http://www.cultura-strep.eu
[23] using IBM's UIMA platform

6.2.7. General strategy with regards affiliated projects

From the very definition of what affiliated projects are, DARIAH should contribute to make them part of a seamless network of actions contributing to build-up our European infrastructural capacities in the humanities.

We need to systematically enter into a sustainable relation with them that reflect the idea of a win-win strategy resulting from, on the one hand, an affiliated project carrying out targeted research and development, and, on the other hand DARIAH, taking up the corresponding results[24], promoting them and making them further accessible within the European research area.

The actual requirements that we may want to impose on affiliated projects should remain quite minimal, and comprise probably the visibility of DARIAH in project related activities, the compliance to DARIAH's policy in the domain of information dissemination and finally, as times goes on, the taking up of DARIAH's recommended good practices and technological background.

At some point, we may want to define an actual branding for affiliated project ("a DARIAH initiative") and make our support more and more professional from helping project preparation to provided basic technical infrastructure (communication, archival, pooling together publications, etc.). More elaborate collaboration should involve progressively inviting the coordinators or main technical partners of affiliated projects to the appropriate DARIAH meetings and in particular make them central partners in the preparation of the annual DARIAH days[25].

We should note here that the policy towards affiliated project should not be limited, *stricto sensu*, to EU initiative at large, and it can be the case that national or multilateral projects may become an essential partner in establishing the European footprint of DARIAH. For each initiative that we will consider in the future, we should just manage the trade-off between potential contribution and the potential overhead with having too many affiliated initiatives.

6.3. Sibling initiatives

6.3.1. TEI[26]

The Text Encoding initiative (TEI) could be seen as the oldest and probably most successful research infrastructure in the humanities. Initiated in 1987 by an international group of text archives aiming at unifying their practices in the domain of digital text representation, it produced in 1992 a set of SGML-based guidelines[27] for representing all types of textual genres (prose, drama, poetry, dictionaries, transcription of speech, etc.). Over the years, these guidelines evolved to become the reference XML-based platform for any project in the

[24] It should be part of the work plan of all VCCs to gather and pool intellectual, editorial and technical developments across affiliated projects

[25] It has occurred to me on several occasions that SDH may just prevent us from pushing the idea of such DARIAH days

[26] With the help of James Cummings – comments and proposals are my own responsibility

[27] The TEI was indeed an early adopter of the SGML standard published by ISO in 1987. The experience thus gained has lead to the TEI experts to be seminal in the definition of what was to become XML

humanities having to deal with textual documents[28]. It has put in place an editorial infrastructure allowing it to be extremely reactive to community need and now publishes two releases per year.

There are several factors that are making the TEI an essential component of the digital humanities environment:

- the TEI guidelines represent one of the most mature and technically sound XML application based on a complete specification environment;
- the TEI has a network of users outreaching in nearly all fields of the humanities: classics, philology, linguistics, history, musicology, etc. "Hence it is not only a community-driven standard but over time a reflection of the evolving concerns of the DH community"[29];
- the TEI has now reached a high level of institutional recognition, making it a default standard for grant application to ANR, DFG and NEH, among others;
- because of the advanced technological background it has developed and the width of its community, the TEI has probably become the place where the best experts in XML related technologies and text encoding for the arts and humanities are to be found.

As a consequence, DARIAH should define a long-term framework to foster what we could call a "symbiotic development" with the TEI. DARIAH should rely on TEI technical contributions and participate to its governance (encourage membership at EU level; encourage early adoption of TEI in new digital projects; integrate TEI contributions in VCC activities. VCC1: tool developments, VCC2: TEI education, VCC3, TEI as reference standard for text encoding). Among the possible action points, DARIAH should encourage participation of its members to TEI related work, either through a participation to the TEI technical council activities, or through the contribution to the development of TEI related tools.

From a strategic point of view, we can anticipate that the TEI may become an organisation providing representation guidelines to a much larger community than just scholars interested in digital texts. The recent taking up of the EpiDoc[30] proposals or the ongoing collaboration with MEI[31] have shown that the TEI framework is well adapted to the integration of additional descriptive modules. It is thus to be expected that the maintenance of formats in digital humanities will in the long run be the remit of the TEI, acting in particular as a clearinghouse of developments initiated within specific communities.

6.3.2. CLARIN

The CLARIN initiative was launched in early 2006 during a meeting organized at CNRS headquarters in Paris which put together, among other participants, representatives of the two

[28] See http://www.tei-c.org/Activities/Projects/ for a list of project which have been registered as using the TEI; thus only representing a small number of the actual user population
[29] J. Cummings, personal communication, with permission
[30] http://epidoc.sourceforge.net/
[31] Music Encoding Initiative http://music-encoding.org/

major networks that had existed in the preceding years, namely Parole[32] and Telri[33]. This context was to bear a strong influence on the CLARIN research infrastructure proposal, as it stood in September 2006 on the ESFRI roadmap, with an emphasis on language resources and associated language technologies. CLARIN is now, with DARIAH, one of the only two ESFRI initiatives related to the humanities and has followed over the year a very similar timeline (preparatory phase and ERIC submission in 2011).

The CLARIN consortium may be characterized as being a large network of small entities—many of which being actual research teams in NLP—whose driving force has for a long time been the data centre of the Max Planck Institute for psycholinguistic in Nijmegen. As a result, it has had a quite technology-driven agenda, focusing on core building blocks that are potentially generic for other types of infrastructures (PIDs, AAI, pipeline architectures, etc.). In a way, more then having to address new needs or new communities like it is and will be the case for DARIAH, CLARIN is more a coordinating structure for a well-defined group of interested parties.

We should also be aware that in the domain of language technologies, there are several other initiatives that make the picture difficult to decipher at times. Without making an in-depth analysis for each of them, we can mention:

- ELRA/ELDA: which had initially been conceived as a public service for distributing language resources within the European research community and has basically become a private structure that has, had as a whole, a negative impact on the free dissemination of language resources and in general on the development of an actual language resource dissemination policy in Europe;

- Meta-Net, an initiative to put together an applied research consortium in language technologies, comprising also partners from the private sector.

The community is also structured around a major event, LREC (the Language Resource and Evaluation conference), organized by ELRA/ELDA (…) and which somehow reflects well the heterogeneity of the groups interested in language resources at large.

On the positive side, the language technology community is by far more technologically mature than many other groups in the humanities, as it has since many years developed numerous tools and has managed to get quite organized in defining a portfolio of standards within a dedicated ISO committee. It has also established strong political connections with the EU commission, which may influence future funding schemes in this respect.

CLARIN focuses on specific technologies and processes for managing linguistic content, but this should not prevent us from putting textual documents at the centre of our interests. This strong overlap is one of the reasons for encouraging designing a more coherent organization of HS infrastructures in the future.

CLARIN and DARIAH are doomed to be working even more closely in the future, an orientation that may be supported by the following arguments:

[32] http://www.ilc.cnr.it/viewpage.php/sez=ricerca/id=63/vers=ita
[33] http://telri.nytud.hu/

- In many cases, the communities that may benefit from CLARIN or DARIAH services are overlapping. Working with language resources is not specific to a field (e.g. linguistics) but belongs to the methodological building blocks of most research activities in the humanities;

- Having two ERIC infrastructures related to the humanities in Europe with such a potential overlap may be a real overhead for both the involved communities but also for the governments. We may have to think of a more coherent administrative scheme in the long term;

- The preceding point has already been reflected in the efforts carried out on a national basis to gain support for both CLARIN and DARIAH, where we could have been better as a whole in coordinating our efforts:

- With a more forward-looking view on existing affairs, we could imagine within ten years' time a single infrastructural framework in the humanities which would allow specific sub-networks (for instance associated to disciplines) to develop as autonomous endeavours while benefiting from one single coordination.

We thus suggest adopting the following action points within DARIAH to pursue a closer partnership with CLARIN:

- Identifying possible joint working groups such as those already experimented in DARIAH-DE;

- Taking actions to reduce the administrative overhead (joint conferences like SDH, collocation of general assemblies, systemizing national budget records)

- Explore with national member bodies of both CLARIN and DARIAH ways of optimizing our activities, and thus identify the possible of joint national roadmaps;

- Work with the commission

- Start thinking at what a single ERIC in the humanities could look like in the future.

6.3.3. Bamboo

The Bamboo[34] initiative started in 2008 as a (mainly) northern American initiative to create a research infrastructure in the Humanities. Financed by the Mellon foundation, it has developed along the following phases:

- An initial "Bamboo planning project", which has gathered input from the research and technological communities through the organisation of workshops all around the world;

[34] http://www.projectbamboo.org

- The current "bamboo technology project", which focuses on the development of a technological framework ("Bamboo research platform") coupled with the identification of possible demonstrators;

- A more sustainable phase whose blueprint is currently under definition.

In its initial phase, Bamboo has had to face several difficulties mainly related to the fact that it wanted both to be too wide in its scope and too specialized for the various communities. Its bottom-up development, through the organisation of open workshops worldwide, has made it difficult for the community to identify in which clear direction it wanted to go.

In the current technological phase, "the actual work involves securing textual content for scholarly use, e.g. from the Hathi Trust Research Centre, and making it available with tools for textual exploration and curation, which will mostly be developed within the project"[35]. The University of Oxford being also a partner in this endeavour, there would probably be concrete ways in, at least, making the corresponding content available to our users and further to see how we could share some technologies and good practices in the domain of textual scholarship.

6.3.4. Other digital humanities initiative

This section is a placeholder to record initiatives that are potentially part of this siblings category, but not yet studied in depth:

- http://www.dhcommons.org/

6.3.5. Other HSS initiatives on the ESFRI roadmap

Beyond CLARIN and DARIAH, the ESFRI roadmap contains three initiatives in the social sciences that, just like the ones in language technology and the humanities have been supported by the EU in their preparatory phase and have been or will be implemented as ERIC. These are:

- CESSDA[36] (Council of European Social Science Data Archives) is a consortium of national data centres, which have a long tradition of joint collaboration at EU level. They have submitted their ERIC application, which should be hosted by Norwegian Social Science Data Services (NSD);

- SHARE[37] (Survey of Health, Ageing and Retirement in Europe) and ESS[38] (European Social Survey), two quite specific European surveys, which have had their network of participants far before the ESFRI roadmap existed.

Despite the fact that the three initiatives appear in the same group within the ESFRI roadmap and that, indeed, we have a partnership within the DASISH project, we should not over emphasize our need to establish close collaborations. As analysed in this report we have

[35] Martin Wynne, private communication
[36] http://www.cessda.org/
[37] http://www.share-project.org/
[38] http://www.europeansocialsurvey.org/

potentially many other more important priorities which should make us busy to establish a stable research infrastructure in the humanities as such.

6.4. Cultural heritage initiatives

By cultural heritage initiatives, we consider all European endeavours covering the library, museum and archival sectors, which aim at providing a better identification and access to the corresponding assets. As analysed in the section on a possible data-oriented strategy for DARIAH, we need to anticipate on ways to provide such content to the research communities in the humanities, while taking into account from a technical and usage point of view how much we need to influence the corresponding decision process.

6.4.1. Europeana - The European Library

Europeana is an initiative resulting from the wish of several European national libraries to pool together their catalogues and assets and offer a central portal providing the corresponding access facilities to the wide public. It is by essence cross-domain (museums, libraries, archives, and audio-visual collections) and currently groups together over 90 cultural organisations. It thus boasts at present access to around twenty million cultural artefacts—books, films, sounds and pictures.

The work carried out in Europeana is complementary to the TEL (The European Library), which aims specifically at networking all library catalogues, including research and local libraries, in Europe. The library-oriented view has allowed TEL, right from the onset, to develop an integrated search on existing catalogues and thus develop a reference expertise on meta-data interoperability.

It is essential for DARIAH to have an early strategy as to the possible collaboration with Europeana-TEL[39], including work on meta-data interoperability, document transcription and enrichment, as well as legal issues.

A typical case where collaboration would be particularly easy to establish is the *Europeana Regia* project[40], which is working on methods and tools to make manuscript descriptions visible through the Europeana infrastructure. In particular, they are working on mappings to make contents expressed in EAD or TEI (using the msDesc component) exchangeable in EDM (Europeana Data Model).

Such initiatives are very important for the scholarly community, since it anticipates a scenario according to which TEI encoded scholarly editions will be directly accessible as part of the materials that can be queried through Europeana.

6.4.2. DC-NET

The Digital Cultural Heritage Network (DC-NET) is a European consortium of national cultural institutions (involving Ministries of culture representatives), whose objective is to coordinate national policies in the domain of the creation and management of digitised cultural assets.

[39] The two should form soon a single entity
[40] http://www.europeanaregia.eu/

Although the DC-NET EU project proper will end in December 2011, the corresponding network will remain an essential political actor with whom we need to concertate to determine better access possibilities for our scholars to digital cultural heritage assets.

As soon as DARIAH is established as an ERIC, we should see, through the corresponding national contact points in DC-NET, how we can have a joint vision at the service of scholars. This should also be put in perspective in a context where research and culture are not always closely coordinated in all European states.

6.4.3. ARROW[41]

ARROW is developing a 'rights information infrastructure' for Europe. What this means is that the 'book metadata' from libraries, publishers and rights holders has been made interoperable, currently in four countries (France, Germany, Spain and the UK). In the follow-up project, ARROWplus, the system will be made to work in additional 12 countries. The rational behind ARROW was to develop a system to speed up the copyright clearance process so that more 'in-copyright' material could be included in portals such as Europeana.

The issue of in-copyright material comes regularly on the agenda of scholars working with contemporary literature for instance and in the long run, DARIAH will not escape having to consider these issues. However, we should recommend to just make a passive watch on the corresponding results, at most to be able to give advice as to where the best experts are.

6.5. Technological initiatives

6.5.1. EGI[42]

The European Grid Initiative (EGI) has been established as a foundation under the Dutch law to pool together computing resources and the associate expertise in the domain of grid computing. EGI federates national grid initiatives in Europe and is articulated around virtual research communities (VRC), corresponding to scientific communities having similar needs in the domain of data intensive computing.

In the last year, we have had continued contact to EGI with the idea to work on possible links with them, although we are moving bottom-up. Both DARIAH and CLARIN are cooperating closely with the Julich Supercomputing Centre, the KIT, the GWDG and RZ Garching to establish data infrastructure for the humanities. This includes PID, AAI, and bit preservation—several issues which has been mentioned the letter of intent signed jointly by CLARIN, DARIAH and EGI. This is also being implemented in the EUDAT project.

Still, it is not clear if we should support them putting together a Virtual Research Community for the Humanities? The risk of interference with a proper development of our own work program is not to be neglected.

[41] http://www.arrow-net.eu/
[42] With the help of Andreas Aschenbrenner—comments and proposals are my own responsibility.

6.5.2. EUDAT

EUDAT is a 3-year (2011-2014) EU project[43] that groups together computing centres wanting to share basic technological components across the services they deliver to the scientific community. This project partially echoes the already mentioned report "Riding the wave" in that it aims at dealing with the management of large amount of data[44] in science. Its main action lines are articulated around the notions of compatibility, interoperability, and cross-disciplinary research in a context of data intensive science.

Looking more closely at their work program, we can see that the main focus is on basic services that may facilitate long term data preservation and access, namely: persistent identifier service, workflows and web access, federated AAI, etc. It also aims at working for specific scientific communities by deploying meta-data and data mining services.

Seen from a distance EUDAT shares several objective, tasks and partnership with DASISH, as well with EGI. Whereas we should maintain a light technology watch on their achievements, we do not need to plan any specific collaboration scheme with them.

6.6. Larger circle of infrastructural initiative

6.6.1. OpenAIRE

OpenAIRE[45] is a direct follow-up to the DRIVER[46] project and aims at accompanying the EU open access policy for FP7 and FP8 programs. More specifically the objective of the EU in launching the call that led to OpenAIRE was to establish a publication repository infrastructure that would be the background for a publication deposit mandate that is likely to be issued for FP8 projects. The following characteristics of OpenAIRE can be put forward:

- The organisational model of OpenAIRE mainly relies on a network of university libraries;
- as a consequence, OpenAIRE is organized as a network of "institutional repositories" (see Romary & Armbruster, 2010) corresponding to the very partners of the project;
- It offers a limited support outside this framework, by means of an "orphan repository" hosted by CERN;
- The focus being on FP7 and FP8 projects, the search portal specifically searches through publications in relation to an EU funded project.

As a result, the current configuration of the OpenAIRE project does not bring services that would be directly useful to DARIAH. Still, the importance of a) having a publication repository strategy in DARIAH and b) relate this to the strong EU willingness in this respect, should make us consider being part of any future follow-up that would address specific

[43] FP7 e-Infrastructure Call 9 (WP11): INFRA-2011-1.2.2: Data infrastructure for e-Science (November 2010)
[44] As mentioned earlier the situation in the humanities may not be so much one of quantity but of complexity of the information one has to deal with.
[45] http://www.openaire.eu/
[46] http://www.driver-repository.eu/

scholarly domains. This should be related to a DARIAH open access policy as outlined further down in this report.

6.6.2. APA

The Alliance for the Permanent Access to the records of science (APA) started in 2004 as an initiative by major national libraries (BL, KB) to explore their possible involvement in the archiving of scientific information (publications and data). It was not until 2007 that the initiative took shape as a consortium on subscribing institutions, following the opportunity of an EU support through some targeted call for projects. Current members in the APA include CERN, ESA, ESF, STFC (UK), MPG (DE), CNES (FR), BL (UK), DNB (DE), KB (NL), and the STM[47] association.

The analysis of the current structure and activities of APA may lead to the following observations:

- The APA is conceived as a political entity aiming at acting as a direct interlocutor to the European commission in the domain of scientific information;
- Still, the heterogeneity of its membership base, and in particular a) the central role of national libraries and b) the presence of the STM association[48] prevents it to be a real partner for the academic world;
- Its extremely large scope combined with its reduced staff capacities[49] has prevented the alliance to yield any significant results so far.

The appointment of David Giaretta as director probably ensures that the APA will still be listened to at the commission level and will thus receive regular funding. However, the APA serving no real community, nor having clearly identified objectives, there is little perspective in pursuing active collaboration with it.

6.6.3. GRDI2020

GRDI2020 (Global Research Data Infrastructures) is an EU project funded within FP7 under the Capacities Programme. The partners behind are basically the remaining participants of the DELOS network, which, coming from the theoretical realms of the digital library business are now surfing the wave of eScience. The recent report[50] they produced is a good basis to get acquainted to their focus of interest, namely covering all aspects of "Data Security, Data Preservation, Data Interoperability, Open Access, Data Quality and Curation, Virtual Research Environments".

[47] International Association of Scientific, Technical and Medical Publishers
[48] The STM association has for instance always prevented the APA to put open access as an agenda item in its activities
[49] David Giaretta is close to being the only staff of the APA
[50] Thanos, Constantino 2011, "A vision for Global Research Data Infrastructures", GRDI2020 report. I feel responsible here for interacting with the author of the report to mention the existence of some project in HSS on the ESFRI roadmap.

With no real background in eScience, no real application domain and no real target user, we may even ask ourselves the actual role that such an initiative could play. My recommendation is to be politically polite but avoid loosing too much time in related concertation activities.

7. Next steps

This report expresses many possible priorities and orientations for DARIAH, which at this stage should be exclusively seen as expressing the opinion of its author. The next step is now to articulate these proposals within the DARIAH DCO and Coordination Board to assess how much consens they can gather and when applicable derive a future strategic plan for DARIAH.