Sustainable data for sustainable infrastructures

Laurent Romary

Sustainable data for sustainable infrastructures

Laurent Romary

Inria, Directeur de Recherche

DARIAH, Director

Abstract

DARIAH, the Digital Research Infrastructure for the Arts and Humanities, is committed to advancing the digital revolution that has captured the arts and humanities. As more legacy primary and secondary sources become digital, more digital content is being produced and more digital tools are being deployed, we see a next generation of digitally aware scholars in the humanities emerge. DARIAH aims to connect these resources, tools and scholars, ensuring that the state-of-the-art in research is sustained and integrated across European countries.

To do so, it is important to understand the actual role that proper data modelling and standards could play to make digital content sustainable. Even if it does not seem obvious at first sight that the arts and humanities would be fit for taking up the technological prerequisites of standardisation, we want to show in this paper that we can and should integrate standardisation issues at the core of our DARIAH infrastructural work. This analysis may lead us to a wider understanding of the role of scholars within a digital infrastructure and consequently on how DARIAH could better integrate a variety of research communities in the arts and humanities.

DARIAH – a digital infrastructure for the arts and humanities

In recent years there have been two major trends that have directly impacted on the establishment of DARIAH as an e-Research infrastructure in the humanities:

- A remarkably growing interest for digital methods in nearly all research domains in the humanities at large[1];
- The development of generic eScience initiatives, usually anchored on strong political activities at national and European levels[2].

In this context, DARIAH faces a double challenge of a) possible difficulties with focussing 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 on liaising (or concerting) with other ongoing (maybe ephemeral) projects and/or political bodies which see their own activities as related to ours.

[1] See the growing success of the Digital Humanities conferences (http://adho.org/conference)
[2] The latest development of which is the overarching Research Data Alliance (https://rd-alliance.org)

To circumvent these difficulties we can identify some core strategic orientations based, on the one hand, on the essential steps in the digital scholarship workflow and, on the other hand, on 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 acknowledge that the technological context is also an important factor to consider, and will indeed appear at several points in our presentation of the institutional landscape, we do not provide any specific background analysis here.

The history of DARIAH began in January 2006[3] when representatives from four European institutions[4] met to identify how they could join efforts in providing services to the research communities they served, with a strong focus on the humanities. The idea behind this initiative was to move towards a consortium of institutions which would ensure the long-term sustainability of the underlying infrastructure and a strong political voice towards the EU. Each institution played a role in coordinating or developing digital services in the humanities at national level, and could thus speak from a national perspective.

Within a hazardous context in which the idea of going digital is not necessarily mainstream in the humanities, DARIAH has managed to move forward to a stage where it is about to become one of the most stable components in the eHumanities landscape. 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:

- A research infrastructure in the humanities should be able to provide concrete short-term services that may lend it scholarly recognition;
- At the same time, it should have a clear vision of 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 their

[3] Just a few weeks earlier, the first meeting of Clarin took place at the same location in the Headquarters of the CNRS in Paris, grouping together the previously existing Parole and Telri networks

[4] 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.

highly fragmented scholarly structure, and, on the other hand, of their low technical literacy. Besides, the humanities are usually subject to comparatively low budgets, which leaves even less leeway for dedicating funding to infrastructural activities. 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.

A user-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 mean the now quite large community of scholars who have to deal with digital content[5], regardless of whether they master the technical background related to the creation or management of these resources.

The sufficiency with which services fulfil expectations will rely a great deal on 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 responding to 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, 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[6].

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 that they can explore, study, and finally reference in their own research. To help achieve this, DARIAH works on deploying services along the following lines:

- Discovery portals that acts as single entry points to existing online resources[7];

[5] Usually because they benefit from a research grant where they engaged themselves in delivering digital content or applying digital methods.

[6] see also: "Reinventing research? Information practices in the humanities", Research Information Network report, April 2011. http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/information-use-case-studies-humanities

[7] See the exemplary service provided by Isidore at CNRS (http://www.rechercheisidore.fr)

- Recommendations on optimal web searchability (e.g. *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 agencies 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[8] or CENDARI[9]) 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.

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 prioritizes the provision of services that help scholars to quickly learn how to work autonomously 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[10];

- Provision of editors in a box that point to a reduced set of environments that can be directly installed or used online to create relevant scholarlydigital content;

- Advertise and/or organize training workshops all over Europe so that scholars or newly hired students can be trained and gain quick autonomy.

These services should be strongly articulated with the standardisation strategy we will delineate later in this paper.

Preserving and disseminating content

Once digital assets have been created, it is essential that researchers are not left wondering how to make them widely accessible while ensuring that the resources will be trustfully used and cited. To this end, the DARIAH short-term agenda includes the following priorities:

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

[8] http://www.ehri-project.eu
[9] http://www.cendari.eu
[10] 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

but also encourage the deployment of national initiatives for researcher identification[11];

- Provide an online service for research asset PIDs. In this context, we should strengthen our relationship with EPIC[12] and DataCite;

- Provide recommendations on a core set of meta-data they have to apply in their resources to make them 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 applied in 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 by offering real hosting services (e.g. XML database farms)

Additional service related to publications

Although scholars may not request it from the outset, 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 papers in a publication repository;

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

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

Overview of short-term priorities

We understand that DARIAH can benefit scholars by offering modest but targeted services. DARIAH should also be able to boast this modesty to external actors (members, EU) and show how it is part of a long-term strategy to develop an infrastructure for the humanities.

---

[11] See for instance the IdRef service at ABES in France (http://www.idref.fr)
[12] EPIC – the European Persistent Identifier Consortium; http://www.pidconsortium.eu

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 in the creation of more elaborate virtual research spaces[13] based on a more data-oriented perspective, as outlined in the next section.

A data-oriented view for DARIAH

Towards a stable perspective for DARIAH

Contrary to the short-term strategy, the long-term vision of DARIAH should somehow go beyond a purely user-centric view. Indeed, given the speed at which technological awareness is presently evolving, it is nearly impossible to anticipate what scholars will actually request from a digital infrastructure in the humanities over the next five years alone. In this context, our duty is to create a sound and solid background that is likely to ensure the stability of digital assets in the long run, but also the development of a wide range of as yet unanticipated services to carry out new forms of research on these assets.

This data-centred strategy echoes various reports and statements that have been issued recently, in particular "Riding the wave"[14], which has placed 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 aspects (curation, standards) and sociological aspects (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 andin the way we 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 research. This wide notion typically covers objects such as manuscripts, artefacts, sculptures, recordings, statistical data, observations, questionnaires, etc.

Surrogate – definition

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

---

[13] cf. Laurent Romary "Scholarly Communication", in Mehler, A. and Romary, L. *Handbook of Technical Communication*, de Gruyter (2012)", http://hal.inria.fr/inria-00593677

[14] http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204

Surrogates can take a wide variety of forms ranging from metadata records, scanned images of a document, digital photographs, transcriptions of a textual source, or any kind of extract from or transformation[15] of existing data.

The notion of a surrogate is at the core of digitally based scholarship since it is intended to act as a stable reference for further scholarly work, as a replacement for – or complement to – the original physical source it represents or describes. By definition, it should always contain some minimal information to refer to the source(s) upon which it is based.

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 of this on our delivery of services.

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 standardisation landscape itself by supporting participation in 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 endeavour to define interoperability conditions between heterogeneous worlds (e.g. EAD – TEI relationship).

Finally, we need to assess the consequences of an extremely widely 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.

[15] E.g. the spectral analysis of a recorded speech signal

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 invest 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[16];

- Spend meaningful effort on 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 with 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[17].

Licensing issues – open access strategy

The evolution of the digital humanities towards a complex and interrelated data landscape will require a strong policy concerning 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:

- Theideological factors in the debate provide that each scholarly production financed by means of public funding is in essence a public good[18]. 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.

---

[16] 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.

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

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

All in all, the core reasons why we have no choice but to work towards an open data space are well identified and boil down to the issues of[19]: more efficient scientific discovery and learning, access for other researchers—and the wide public—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 provide 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.

To take a further example fromthe 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 prevent 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 the lightest possible licensing schemes are applied to the data made available to scholars.[20]

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 within DARIAH alone. 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 to what extent potential data 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:

[19] 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.

[20] 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[21] 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 access to one or the other data provider more transparent;

- Identification of possible scenarios in which the archival location of versions of records is clearly identified and, by the same token, enrichment mechanisms are contemplated[22].

Role of standards

The main issue in defining a policy about standards is to understand what they actually are. Standards are documents informing about practices, protocols, artefact characteristics or data formats that can be used as reference for two parties working in the same field of activity to be able to produce comparable (or interoperable) results. Standards are usually published by standardisation organisations (such as ISO, W3C or the TEI consortium), which ensure that the following three requirements for standards are actually fulfilled:

- Expression of a consensus: the standard should reflect the expertise of a wide (possibly international) group of experts in the field
- Publication: the standard should be accessible to anyone who wants to know its content
- Maintenance: the standard is updated, replaced or deprecated depending on the evolution of the corresponding technical field

Standards are not regulations. There is no obligation to follow them except when one actually wants to produce results that can be compared with those of a wider community. This is why a standardisation policy for DARIAH should include recommendation as to which attitude the scholarly communities could or should adopt with regards to specific standards.

The preceding characteristics outlined for standards put a strong emphasis on the role of communities of practice and the corresponding bodies that represent them. Ideally, a good standard reflects the work of a relevant community and is maintained by the appropriate body. This is exactly with the case for the Text Encoding Initiative for text representation standards and, to a lesser extent, for EAD, whose maintenance is taken up by the Library of Congress with support of the Society of American Archivists.

Because there is no obligation to use a given standard, it is essential to provide potential users with a) awareness about the appropriate standards and the interest to adopt them, and b) the cognitive tools to help them identify the optimal use of

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

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

standards through the selection and possibly customisation of a reference portfolio. In our experience with working with numerous projects (including those cited in this document) that were in the need of adopting existing standards, there was always an initial phase in which scholars should be made aware of some core standards that are systematically related to the definition of interoperable digital objects. We call these core standards a *standardisation survival kit (SSK)* and outline in Table 1 a first group of such standards. As we will see later in this document, the SSK should be part of several concrete actions for DARIAH in the domain of education and interaction with funding agencies.

An important aspect in this dissemination strategy is that projects should be told to refrain from defining their own local formats and instead first demonstrate that their needs are not covered by the wide varieties of already existing initiatives in the digital humanities landscape. This is also why DARIAH should avoid taking any specific lead in the definition of new standards[23], but should have a pro-active role in helping communities to participate in standardisation activities where they exist. Such a strategy will also contribute to the actual stabilisation of existing conceptual and technical knowledge within ongoing projects, as well as providing a channel for the wider dissemination of the corresponding results.

ISO 639 series	Codes for the representation of languages and language families
ISO 15924	Codes for the representation of scripts
ISO 3166	Codes for the representation of country names
IETF BCP 47	Standard for encoding linguistic content, combining ISO 639, ISO 15924 and ISO 3166
ISO 10646, Unicode	Universal encoding of characters
ISO 8601	Representation of dates and times
XML recommendation	Provides the basic technical concept related to XML documents

Table 1: Outline of a standardisation survival kit

Recommendations

The preceding sections could potentially lead to many possible action points for DARIAH. At this stage, we can boil these down to the following concrete recommendations:

- Define a basic curriculum on data modelling comprising awareness about digital surrogates, meta-data, versioning, multiple publishing, annotation and re-use
- Re-design the schema registry activity to focus on designing data models and formats toolkits for research projects
- Define and maintain a Standardisation Survival Kit that corresponds to the baseline of an awareness and recommendation activity on standards

[23] In this respect we should strongly depart from the strategy adopted in Clarin with infrastructure-internal format developments such as TCF or CMDI.

- Support and coordinate (VCC2 and VCC4) standard awareness workshops targeted at specific scholarly communities
- Encourage DARIAH members to allocate means for their participating institutions to contribute to standardisation activities

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 setting 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.