Technologies, services and user expectations–Prospects for DARIAH

Laurent Romary

DARIAH-EU

Technologies, services and user expectations — Prospects for DARIAH

R EU 4.3.2

Version - 15 February 2013
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

Document status: Final Draft
Dissemination level: Public (CC-BY)

Autors: Laurent Romary (SUB)

Revisions:

Datum	Author	Comments
15 February 2013	L. Romary	delivery

Prospects for DARIAH..... Erreur ! Signet non défini.
1. Introduction **5**
2. Background projects **5**
2.1. Berliner Intellektuelle 6
2.2. Berliner Repertorium 6
2.3. Laudatio 7
2.4. Cendari 7
3. Technologies for digital sources – do these projects have something in common? **8**
3.1. Identifying the core components of a digital editorial workflow 8
3.2. Edition — setting up scholarly-relevant data 8
3.2.1. Ad hoc editing environments 8
3.2.2. Native XML editing — “hiding the code”? 9
3.2.3. Dedicated environments 10
3.2.4. Editing vs. importing 11
3.3. Manage 11
3.3.1. The general issue of a repository solution 12
3.3.2. Long term archiving 12
3.3.3. Persistent identifiers 12
3.3.4. An essential mechanism – versioning 13
3.4. Publish 14
3.4.1. Generic publishing environments 14
3.4.2. Dedicated environments 15
3.4.3. What about books? 16
3.5. Wrapping-up — recommendations 17
4. Conceptual Tools **18**
4.1. Getting a clear idea of what a source is 18
4.2. The FRBR model (<pron>ˈfɜrbər</pron>) 18
4.3. Data modelling 21
4.3.1. When both the model and the format exist – the Cendari case 21
4.3.2. Linking model and format — the Laudatio case 22
4.3.3. A memory of modelling choices 23
4.4. Role of standards 24
4.5. Recommendations 25
5. Political tools **25**
5.1. Why help is needed 25
5.2. Digital primary sources — Copyright — plagiat 25
5.3. Relation to the libraries, archives and museums 26
5.3.1. The TEL – Europeana scenario 26

5.3.2. When the library is going digital… designing digital services 27
5.3.3. The visionary scenario – co-development of digital services 27
5.3.4. Recommendations 28
5.4. Publication as part of the infrastructure for research 28
5.4.1. A missed opportunity? 28
5.4.2. Considerations for DARIAH 29
5.4.3. Conclusions and recommendations in the domain of publications in the humanities 29
5.5. A wider political plan 29
6. Educational aspects 30
7. Going further 31
8. Annex — Cendari's "Data Sharing Agreement" 31

1. Introduction

This DARIAH-DE report gives a complementary view to the first of its kind, entitled "Partnerships, relationships and associated initiatives ", where we analysed the tenets of the DARIAH infrastructure and provided an environmental analysis in order to outline a strategic plan for DARIAH. The present report focuses on "Technologies services and user expectations" with a view to better understand what kinds of concrete services we want to offer to our users, given the developments made within the various DARIAH partners in general, but with a special view on contributions of DARIAH-DE.
As such, the title is already misleading because its various components are probably in the wrong order. Are we supposed to know about services and even more technologies without understanding who our stakeholders are and what they can expect or gain from the establishment of a Europe-wide eInfrastructure in the humanities?
An essential aspect of putting together a research infrastructure is indeed to fulfil expectation from its potential users. But how can we speak about expectations in general without actually keeping track of the ongoing cultural changes that occur so rapidly in the humanities at present? We thus decided to solve this contradiction by presenting an intentionally biased view on the difficulties that newcomers to the digital world actually experience, on the basis of our observations within a reduced group of ongoing projects.
Still, such an approach relates to the vision we have that the core stakeholder group for DARIAH activities are indeed research projects within the humanities that have received a national or European grant and whose work program comprises an important move to digital methods. Such projects are essential for several reasons:

- They are anchored on a clear scholarly domain, thus providing a precise insight on the underlying research issues
- They are likely to have clear needs in terms of digital data management and tools
- They have actual funding for their own grassroots developments, which are likely to bring new tools and services for the corresponding scholarly community at large
- They have a clear view that they have to take sustainability measures for their results

We thus think that this group of core users are and will be for quite a long time our target stakeholder group. By addressing them, with the effect of such projects taking the lead in their corresponding communities, we will magnify our impact and actually, although at times indirectly, reach out a wide community of scholars in the humanities.
We cannot guess what the users expect; we can only provide them with means to express their expectations.

2. Background projects

As background for our analysis, we will take four specific projects which, at various degrees, can be seen as digital humanities projects, and which are all based on a similar workflow of compiling digital sources at the service of a scholarly interrogation. In all these projects, I have been in the situation of observing how scholars themselves were striving to get support for defining their digitization workflow and I have tried to identify from these interactions the kind of short-term and mid-term needs that such scholars would need and how these could actually be addressed by DARIAH both as a social and a technical infrastructure.

2.1. Berliner Intellektuelle

The project *Berliner Intellektuelle*[1] (henceforth *BerlInt*) is a 5-year project financed by the DFG under the *Nachwuchsgruppe* (new research group) program. It aims at studying the interactions between the main literary and scholarly figures of the early 19th century in Berlin, through the analysis of their written productions, mainly made of letters and authors' manuscripts. The sources play a multiple role since they inform the scholars on the genesis of a work, its possible influences and its impact at the time it was published. This is why correspondence, with all the cross-references they contain, but also the study of the author's library when it has subsisted are important sources in complement to original manuscripts.
Most of the documents are spread across a small number of archives and libraries and in particular in the State Library of Berlin, which has accepted at an early stage to provide extensive digital surrogates (scans). The research group uses the scans to create full-text transcriptions, with additional inline annotations concerning persons, places and works. All scholars actually carry out their transcriptions themselves and only a few were done by students since it requires specific skills: experience with multiple scripts (roman and gothic characters) and languages (among which French, German, Latin, and Greek) as well as knowledge of the actual literary surroundings (abbreviations of the names of mentioned entities).
The group started with hardly any knowledge about the digital process and put together an editorial workflow where they could at each stage find the optimal trade-off between efficiency and their own learning curve. In order to move on quickly and attract additional skilled students, they put together a seminar at Masters level on the creation of digital transcriptions in compliance to the TEI guidelines.
They actually succeeded, with the contribution of some DH and LIS students to have a complete workflow in operation with both an online publication[2] and an accompanying blog[3] recording the issues they encountered along their ways.

2.2. Berliner Repertorium

Our second DFG supported project, Berliner Repertorium, deals with the analysis of German translations of religious texts as they appear within a wide corpus of manuscripts from the late medieval period. Their "primary sources" is thus a complex combination of catalogues, where the corresponding manuscripts are being recorded, manuscripts whose description will help tracing the history of the corresponding documents and thus the possible geo-temporal influences, and the psalms themselves. It is interesting to observe that their actual corpus is just a subset of what could be a complete digital library of the corresponding manuscripts, as well as paradoxically an extension, since they record information about documents which do not have any contemporary existence any more.
This has led the project to depart from a one level document structure for their documents (as we have in most text based projects such as the Berliner Intellektuelle for instance) and, as depicted in Figure 1, organise their corpus as two-layers with first the standard representation of the manuscripts (meta-data and general document structure) and an interpretation level, which groups together all which groups together all the relevant information for each translation, from transcription to precise annotations.

[1] http://www.literatur.hu-berlin.de/berliner-intellektuelle-1800-1830/
[2] http://tei.ibi.hu-berlin.de/berliner-intellektuelle/
[3] http://digitalintellectuals.hypotheses.org

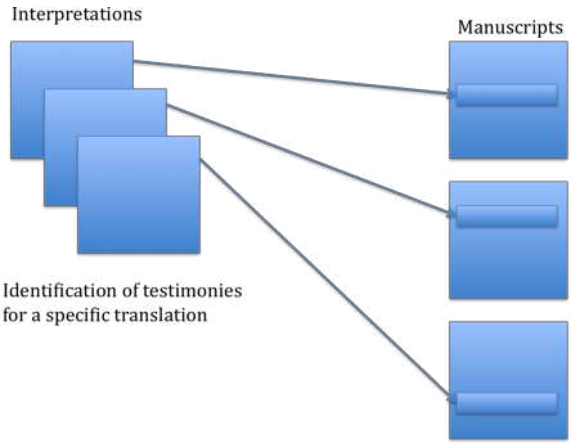


Figure 1: Simplified Document architecture of the Berliner Repertorium project

2.3. Laudatio

The Laudatio project, also financed by the DFG, has a more infrastructural orientation as it brings together researchers in corpus linguistics with colleagues from the central computing services of the Humboldt University in Berlin. The core objective is to design generic corpus management mechanisms that would serve as a basis for a generic infrastructure in corpus linguistics.

In this case, the scholars have a strong technical background since not only do they come from a scholarly field, corpus linguistics, which has since decades developed computer based methods, but also because the group has designed in the recent years one of the major corpus management and query environment in the field. Still, moving from the traditional view in corpus linguistics of a scholar scrutinising his/her corpus by applying a variety of tools within his local workbench to the design of a generic corpus delivery and study platforms clearly required a profound move to other domains of digital object management.

One of the main tasks in the first phase of the project has thus been to design and implement a corpus description model based on a meta-data structure allowing to document all aspects of a complex corpus such as its origination, its actual linguistic content, and the various annotations it actually comprises.

2.4. Cendari

The European Cendari project has been designed from the onset as a DARIAH affiliated project. Based on the idea of defining a virtual research environment allowing historians to identify, combine and use materials available within a network of archives across Europe, it is organised so as to facilitate the emergence of stable editorial and technical solutions. Even if the project itself has a focus on medieval and First World War materials, it is expected that most of the worked-out technical components would easily adapt to any kind of archival-based historical research.

The project was conceived as a combination of historians, technical partners (from the DARIAH network) and archives. The historians are involved in two main activities: first they contribute to the data acquisition by describing archives or liaising with them; second, they form a network of competence by exchanging experience and students within a transnational access program.

From the point of view of their prior digital competences, scholars presented a strong contrast between medievalists who had already participated in the creation or use of large-scale

digitization projects[4], and WWI historians with very little technical background, even if a strong understanding of the archival domain.
The first period of the project was thus dedicated to two main tasks: first making technical partners and historians to get a common understanding of the kind of digital process that could be put together, but also, in a very pragmatic way, putting together the basic building blocks of an editorial workflow to acquire archival descriptions, collection descriptions or, mainly for the medieval domain, document level information.

3. Technologies for digital sources – do these projects have something in common?

3.1. Identifying the core components of a digital editorial workflow

The projects we presented in the previous section are just a small sample of a very large group of research projects for which the acquisition, edition, management and possibly publication of digital information directly impact on the capacity of the scholars to carry out their research. Through the observation of their evolution and in particular how they devised their own digital editorial workflow, we can come to conclusions and recommendations that actually echo the wider observation of "young" digital humanities projects.
In the following sections, we will organise the description of such workflows along the following categories:

- *Edition*: where the data is acquired, either from an external digital provider or from an analog source;
- *Data management*: with all back-office functionalities ensuring the update and further re-use of data
- *Publishing*: which covers all the functionalities related to the access and dissemination of data

3.2. Edition — setting up scholarly-relevant data

It all starts with a corpus. We make the assumption here that the scholars come with a clear research question in mind and, closely related to this, a precise idea of what the optimal corpus could be. Such a corpus is basically a combination of primary and secondary sources ranging from written or printed documents, audio-visual data or actual artefacts (e.g. military uniforms). Sometimes, the sources come with existing documentation in the form of printed catalogues or online web pages. All in all, we are still in a phase of digital scholarship where such sources already exists in the form which will allow the scholar to perform his research with no acquisition or transformation overhead. At the end of the day, his/her aim is to have corpus of digital surrogates, comprising meta-data, one or several representations (scans, 3D, transcriptions) and maybe preliminary annotations he/she may have gathered during this editing phase[5].

3.2.1. Ad hoc editing environments

"OK. We need to enter some data. Shall we do this in Excel?"
"No. XML is too complex for the students. They can type in transcriptions in Excel."
Not sure that the quotations are precise to the word, but this is exactly what we can hear as a first reflex when data acquisition is at hand. It can indeed be tempting for scholars with no

[4] In particular *Manuscriptorium* (http://www.manuscriptorium.com), with extensive TEI compliant transcriptions.
[5] One cannot understand the role of the scholars during this editing phase without the understanding that this is usually the stage at which he/she becomes deeply acquainted with his source material.

background in digital methods or when the complexity of training students that will do the actual data entry is too high to resort to the basic idea that bringing data in digital form can simply rely on general desktop tools that everyone can work with.
Tools such as word processors, spreadsheets and even desktop database environments may give the illusion of providing adequate structure around some digital representations of the data. They also allow a very quick deployment with the capacity within a few seconds, and the identification of a couple of useful fields or layout features, to start entering data.
However, this trend creates real sustainability problems, with hardly any further possibility to re-use the data after the edition stage, either internally for searching or annotating or externally by other research groups. The main difficulties usually occur after a while when the evolution of the corresponding software platforms prevent the corresponding proprietary representations to be legible any more.
This also creates huge difficulties to come back to a solution that would be more appropriate for the management of their scholarly sources. All in all, when the projects are not prevented from going in this direction, they are stuck in a data silo that may last for years.
This is why, it is important that DARIAH has a proactive role on informing scholars, at a very early stage, about the danger of a wrong data management strategy. While not part of the technical support proper, we will see how this could be reflected in a more political recommendation.

3.2.2. Native XML editing — “hiding the code”?

The quite opposite alternative to the extreme situation presented before is probably to consider to have the data entry directly be made in the target format intended for further handling by presentational or analytical tools.
The pros and cons of native XML editing are regularly part of any digital humanities debate dedicated to source material[6]. In order to better understand the impact on the services that an infrastructure such as DARIAH should be offering to the scholarly communities, we can boil them down as follows.
Arguments in favour of offering a native XML access to scholars for them to edit, correct or enrich data are basically twofold:

- The resulting data, when expressed in the appropriate format (and/or standard, see further in the section on standards), is immediately usable for stages of more scholarly related work;
- The scholars thus acquire a deep understanding of their own data so that they know precisely which concepts are actually embedded within it.

Conversely, as a kind of mirror to the two preceding arguments, the view that scholars should not be put in contact with the XML source of his own data is essentially supported by the two following ideas:

- There is a steep learning curve that slows down the data acquisition phase and may even deter the scholar from pursuing along this path;
- The actual result may just be inappropriate with regards to the underlying format because the scholar does not actually master the actual implications of his/her encoding.

[6] Ranging from basic impact on student evaluation as in Edward Vanhoutte keynote "So You Think You Can Edit? The Masterchef Edition" at the TEI conference in Würzburg to general aspects of data modelling in the corresponding workshop organised at Brown University in March 2012 (http://datasymposium.wordpress.com/2011/09/14/data-modeling-in-the-humanities/)

In this context, expressing an opinion will always be spotted as bearing an ideological stance. Still, there are a few guidelines that can be set as background before an actual decision is being taken:

- Various factors related to the project itself should be taken into account before taking any decision: how much is the data likely to be reused? Do we have to do with a short-term project where the data acquisition phase can be "quick and dirty"? How much resources (time and man power) does the project have to justify the learning implication of native XML editing?
- The existence, or not, of a possible alternative to native XML encoding. We will address in the following section the issue of more elaborate environments, but the risk exists that the alternative may be ad hoc editing environment;
- Even if the scholar does "encode" himself the data, it is essential that he/she understands what the actual concepts embedded within the codes are. As we will also see where we will address the issue of data modelling, he/she should be able to ascertain how much the encoding reflects his/her own scholarship in relation to the primary source;
- The decision also depends on the actual availability of close support for the setting up of a native environment. Whether from a local competency centre, some specific introductory material provided by a DH infrastructure or isolated initiatives[7], the scholar has to be closely accompanied in his technical endeavour.

At this phase of the history of digital humanities, where we do not have fully transparent virtual research environments which somehow could hide the actual data representation and just show the underlying intelligence, we make a strong recommendation that encoded data should never be hidden to the scholar. Just as the dust of the manuscript is essential for who visits the archive, the precise XML encoding is currently the only way to acquire the flavour of the data.

3.2.3. Dedicated environments

A good example of a transitional editing environment may be XET, developed at Inria by Jean-Daniel Fekete. XET is an open-source system based on codemirror, implemented in Javascript on the client side, but also a mix of Python (for Django) and Javasacript (for Node.js) on the server side. XET allows one to edit an XML document, but with three main additional features:

- It is controlled by an underlying schema that validates data entry on the fly
- It can be linked to authority files to facilitate contextual completion (e.g. for names, institutions, etc.)
- It provides a real-time XSLT formatting in the browser (XHTML) allowing one to visualise data on the fly

It is currently used to edit fairly long TEI documents, and is planned to be an online editor for other formats within the Cendari project.
We encountered a more specific, yet interesting, example of a dedicated data entry environment in the context of the Cendari project for the purpose of providing archive and collection descriptions. We came across using ICA Atom[8], which was developed for several years as an archivist's workbench for entering data. In short, ICA Atom allows one to provide data that reflect the International Council on Archives ('ICA') standards and possibly export this data in a subset of EAD.

[7] TEI by example
[8] https://www.ica-atom.org/

Whereas the tool looked particularly attractive for us in the context where we wanted to have usable data entered by colleagues with hardly experience in XML technologies, we soon had to face several hurdles that were really problematic for our project:

- It did not have a versioning management solution, which created stress on the editors when more than one person could actually modify the same data
- It could also not relate data entries to their creators, with the consequence that no attribution, and thus trust, could be attached to them
- The underlying EAD schema could not be customized to follow the evolutions requested by the Cendari scholars

Paradoxically, ICA-Atom being an open source project, it proved to be a good platform when one had to develop additional feature (e.g. new export facility for EAG), which raises a real strategic issue from the wider perspective of DARIAH. Our recommendation would be to work in two phases a) identify reusable and open source platforms such that are likely to impact on a wide community of users within DARIAH and b) take a strategic stance as to whether we dedicate capacity to the evolution and maintenance of such platforms to make them reference services within DARIAH.

3.2.4. Editing vs. importing

In many cases, what we called the editing phase of a workflow for digital sources comprises the integration of data acquired from third party providers.
In the Cendari project for instance, the manual description of archives comes together with a variety of possible other sources. These can be archive or collection descriptions provided by archives such as the German Federal archives, bundles of item descriptions exported from Europeana, or even full collections of transcribed manuscript from the Manuscriptorium repository at the Czech national library. In the same way, the state library in Berlin can produce EAD exports of their collections that can be used as a basis for the creation of TEI based representations.
Such data come in a variety of formats that may not always correspond to the actual "target" delivery format of a given project. Still, our experience within the particularly complex case of Cendari allows us to identify some possible guidelines as to how to deal with such situations:

- Always keep track of primary digital objects to ensure the best possible traceability of information
- Identify modification and reuse of information as part of a versioning concept (see below)
- Preserve all forms of data and their accompanying versions within a single, yet structured, data space[9] that informs all the data complexity of the corresponding project as well as the various scholars' contributions
- Integrate in this global data concept, all dissemination formats (LOD, EDM, pdf, etc.)

3.3. Manage

The issue of data management usually comes very close together with that of editing, unless of course one keeps his/her own data on a personal disk space. Maintaining and sharing data requires that some specific functionalities be implemented, which are central to the definition of a repository solution for the corresponding type of data at hand.

[9] Which can be inspired from the concept of Blackboard, see B Hayes-Roth (1985) "A Blackboard architecture for control", Artificial intelligence, Volume 26, Issue 3, July 1985, Pages 251–321

3.3.1. The general issue of a repository solution

Independently from the specific features we will address in the following section, it is probably necessary to provide a quick overview of the current trends in repository solutions in the digital humanities domain. Since this topic could be the basis of much longer development, we will limit ourselves to a surface analysis that will inform our further recommendations.
There are various levels of repository solutions that may come to the mind of newcomers in the digital humanities world:

- At the lowest level, one may be tempted to use general-purpose shared directory services, as available in most academic institutions. Even if this can be a start to actually share some data within a local research group, it should be recommended not to pursue such solution since it does not allow the integration of any further data management feature, and in any case will not be fit for online publication;
- The second possibility is to use a digital management system either based on a CMS (Drupal is among the most favoured environment in the DH domain) or some generic platforms such as DSpace or Fedora. Such solutions usually require to have a strong technical support at hand and in particular additional development means to adapt the environment to the specificities of the project;
- Finally, more data type oriented environment such as Islandora (based on Drupal) or of course Textgrid, offer specific services for digital objects relying upon a well-defined structure (in both cases cited here, TEI documents).

Given this picture, the general feeling of projects that are making a move towards the digital world is that there is a huge overhead in deploying most of the above-mentioned solution.

3.3.2. Long term archiving

Once data starts to be created in a systematic way, long term archiving (LTA) is usually the first requirement that comes into place. This is basically due to the fear of losing information, but also being able to guaranty the project funder with some kind of sustainability for project results. We refer here to bit preservation, taking as given that the data fulfils the basic requirements about formats and standards (see further) necessary for long-term legibility.
Most of the time the expectations are locally fulfilled with the existence of a local facility from the University or the hosting institution, but there is general request on having such an LTA facility in close relation to the hosting of a repository solution. As we shall see, this has some impact on the kind of recommendations we should make for the “archive-in-a-box” activity.

3.3.3. Persistent identifiers

The requirement of having a persistent identifier service comes usually second in place after the LTA one, typically when a project starts putting information online and is obliged to move place for some reason. The prospect to actually use the capacity for themselves and for others to refer to the digital objects in a coherent way makes them think of requiring a service in this direction. However, our experience is that they are not necessarily aware of the concept of persistent identifier itself.
There are various possibilities to implement a PID system. The most common ones are based on DOIs and Handles, for which there are good operational solutions for several years. Such solution can also be explicit, in the sense that users have to make a specific request to get a PID from a given service, or implicit when integrated within a repository of a referencing

service. A good example of the last type is the Isodore portal[10], where each object referenced within the portal is being given a PID that can then be further used by the initial data provider. The current work within VCC1 has shown the necessity to have an open and decentralised solution for PID within the DARIAH-EU communities. Such solutions should be anchored in the general identification environment under deployment within DARIAH and combine step by step the capacities that national DARIAH members may offer to the wider community. The most operational one is currently offered by DARIAH-DE at GWDG, but it is essential that more services shall be officially integrated within a global portfolio.
One central issue to be solved in the immediate future is to offer a service not so much to individuals but to a project as a whole. This means that the project has the responsibility to select the appropriate contact person (whether actual or virtual[11]), but also ensure the maintenance of the underlying redirection mechanisms.

3.3.4. An essential mechanism – versioning

The standard scenario that leads to the request for versioning mechanisms within a digital humanities project is that of multiple authoring from people with heterogeneous authority (like students and scholars) on the same data[12]. Still, such needs may be exacerbated in situations such as the Cendari project where one has to face a variety of data sources that may then be corrected, enriched, etc.
If we take some distance with basic error management issues, it is important to inform DH projects at an early stage about the need of a proper versioning concept and solution in order to deal with the following three roles that versioning may take in relation with digital data management:

- Traceability: in the editing process, identifying who has contributed to which information within a digital object
- Release management: in the publishing process, provide a stable ground for indentifying coherent groups of digital objects
- Referentiability: for any further use of the content of a digital repository, to maintain the coherence between any further reference (e.g. annotations or assessments) and the underlying digital objects that are referred to.

Figure 2 shows a typical scenario that could be deployed in the context of an editing workflow where both a single PID naming is required for the digital object as a whole and releases are delivered with their own identification, and specifically associated with a persistent archival stage. In this scenario, a first PID is requested to the DNPI upon creation of the object, for instance a new EAG entry in Cendari. This PID is immediately (manually) inserted in the header of the object and will act as the reference entry point to the object for the external world. We should note here that it is up to the service request to provide the appropriate URL within Subversion[13] that will always point to the most recent version of the object.
When the project want to make a release, for instance for providing a comprehensive public version of a series of consolidated archival descriptions in Cendari, a request is sent to the DNPI in order to:

[10] www.rechercheisidore.fr
[11] In concrete terms allowing an alias for the project to be authorised to request PIDs from one of the DARIAH services
[12] See also Razum Matthias , Frank Schwichtenberg and Rozita Fridman, "Versioning of Digital Objects in a Fedora-based Repository" , GES 2007.
[13] http://en.wikipedia.org/wiki/Apache_Subversion

1. get a specific PID for this release, which will point specifically to the corresponding version, but will also be inserted in the header of the object[14];
2. make sure that the corresponding object, is appropriately archived for long term preservation.

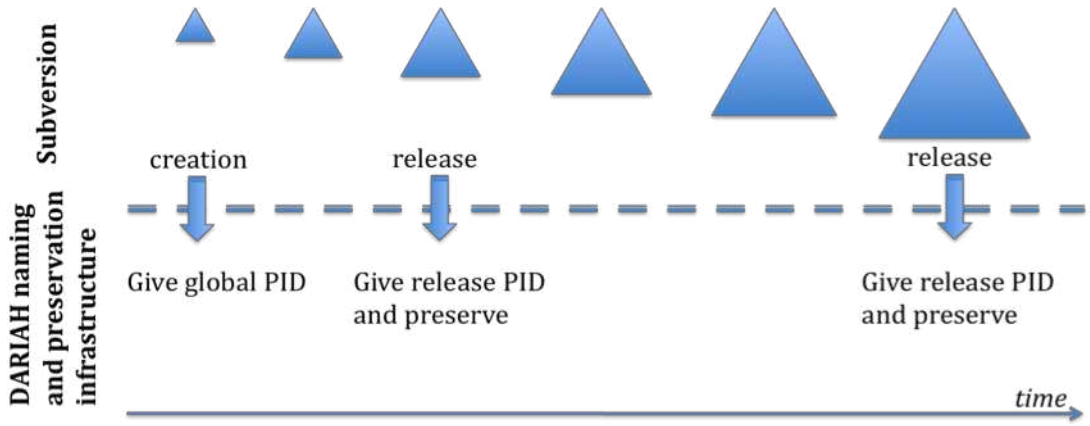


Figure 2: A naming and archiving scenario

All in all, it seems hardly possible to avoid deploying a versioning solution as soon as more than one person may edit a given digital source. The corresponding conceptual and technical overhead though is still high enough to make some projects just disregard this essential step in a standard digital humanities workflow.
Our experience within the Cendari project has shown that it is not so much an issue of finding a solution: there is indeed a wide range of possibilities starting with very simple environments such as an SVN server. The focus for DARIAH should first and foremost be educational: there is a need to spread the information about the role of a versioning environment and the provision of ready-made solutions in this domain.

3.4. Publish

The publication phase of primary sources in a digital humanities project[15] encompasses a variety of objectives that usually makes the issue difficult to tackle in the first place.
The initial aim is usually to "show" the results by providing an online legible (HTML) version that other scholars may simply browse through. The second level is to distribute the actual documents for other scholars to reuse the results. Finally, when a big enough community is concerned with a certain type of sources, comes the idea of actually pooling documents together to form a more comprehensive corpus of primary sources.

3.4.1. Generic publishing environments

There exist several generic solutions for publishing digital content online ranging from simple CMS (e.g. Drupal), general digital repositories (Fedora) or specialized XML databases (eXist). Whereas it is not our purpose here to give an exhaustive overview of such environments, we can try to delineate their possible roles in digital humanities projects and see to which point such solutions could be recommended.
As already mentioned in the section on data management, such generic environments require a great deal of technical expertise to install, but even more to fine tune to the actual specificities and needs of a given project. Since such customisations are likely to take place within specific communities and for specific types of data, we would thus recommend that

[14] in a <revisionDesc> for a TEI document, or <mainhist> in the EAG header
[15] or indeed a humanities project which is going digital, as reflected in our main use case.

DARIAH does not take any specific action in trying to be innovative in this respect but instead remains open to offer the necessary technical support when a specific customisation becomes mature enough to require sustainability.
In complement to this, we should provide some generic information about these environments within the data portal and I would recommend to get inspiration from the Plume project[16] to organise such work. The idea would be to rely on the network of expertise that DARIAH provides to solicit descriptive contributions that would be updated on a regular basis.

3.4.2. Dedicated environments

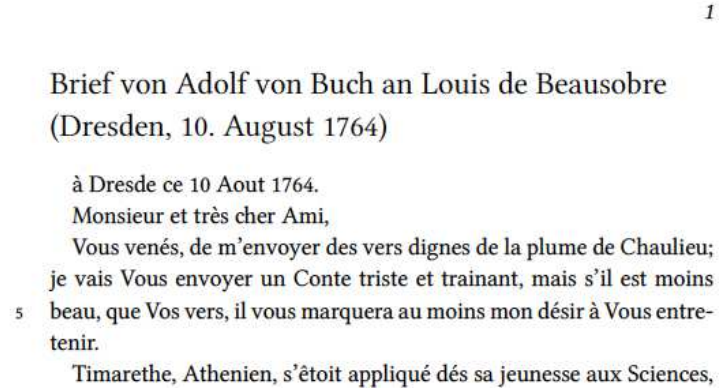
We quickly addressed the issue of dedicated environments when discussing repositories for data management. There is in some way a not negligible overlap between the two, but we want here to put some specific light on environments for the dissemination of legible and re-usable data online.
TEI boilerplate [17] represents in my view the simplest publishing environment for TEI documents ever, and could indeed be easily generalised to any kind of XML encoded source. Without any underlying repository infrastructure, it relies on the client-embedded XSLT processing capacity to apply a presentational XSLT stylesheet on a source document downloaded by a web browser.
A more operational environment for the online publishing of XML encoded primary sources is clearly XTF, which provides a good compromise between deployment facility and expressiveness. XTF is basically an indexing, query, display and browsing layer that relies on XSLT to implement all its functionalities. Being delivered with predefined stylesheets for the TEI and EAD formats, it requires no additional development effort for any one who wants to quickly install a ready made environment, but is also very simple to expand with new presentational features and indeed to take new formats into account.
In the Cendari project, where we actually wanted to have a quickly deployed editorial workflow to get our scholars integrate and use archival data at an early stage of the project, XTF proved to be a very flexible solution to adopt. Installed within a day on the GWDG virtual machine in Göttingen, EAD, TEI and then EAG data were indexed and delivered for scholars to see the result of their data entry activity, but also for the project participants as a whole to see the current status of the Cendari "data soup".
Whereas XTF provides an "archive-in-a-box" solution that has to be installed specifically (and customised) by each project, there are also options for highly centralised publishing platforms when one deals with specific data type. The best example we could mention here is the national image archive built for the scientific community at large in France[18]. Based on the same architecture as the French national publication repository HAL, MediHal offers the basic functionalities required for the publishing of scientific images: author identification with precise affiliations, generic meta-data set for images and of course long-term archiving. We would indeed recommend that such an environment be taken up and made available to the wide DARIAH community.
Finally, one could not address such dedicated environment without mentioning that sources do not come always as one single digital object but can actually lead to various targeted publication or distribution. An important scenario in this respect corresponds to the management of authority descriptions for persons, places or events as encountered, and usually annotated, in sources. Such entities are in general difficult to maintain within the sole scope of a given project, because their description relies on other sources, official (e.g. in

[16] https://www.projet-plume.org/
[17] http://dcl.slis.indiana.edu/teibp/
[18] http://medihal.archives-ouvertes.fr

national libraries[19]) or unofficial (in many other projects), but also because publishing them in isolation would usually provide little added value.
This is why we consider that an initiative such as the Personendatenrepositorium[20] (PDR) is so important, as it provides an integrated environment where several projects can query information about persons as well as adding complementary entries of their own. Such services are essential for providing a stable background for managing entities in digital humanities projects and should be part (cf. below) of the basic tools provided by DARIAH.

3.4.3. What about books?

One of the recurring questions asked by scholars is how much a digital edition of their sources may still allow them to produce a traditional paper edition of part or total of their corpus. We do not consider here the case where the publication of a book is just motivated by a possible gain of prestige[21], but when a stable printable, or viewable on a pad, edition will facilitate the in-depth consultation of the source by other scholars or even in some cases by a wider public.
The first thing to notice here is that the creation of a readable version of a source, although there exist in theory quite a few technical solutions to this purpose, is in general a specific task in a project, which is likely to absorb some man-power. Depending on the complexity of the transcription and the attached annotations, as well as the specificity of the layout that the scholars expect, some precise fine-tuning of the processing chain is to be expected. Still, when the corpus of digital sources is encoded according to reasonably well-adopted standards and the project takes up tools, such as Latex, that are well suited for the purpose of producing a well presented output, it may be worth it to offer an additional dissemination channel for the work.
Figure 3 shows a typical output from the BerlInt project inspired from some prior work done by colleagues at Atilf[22] (Nancy, FR)



1

Brief von Adolf von Buch an Louis de Beausobre
(Dresden, 10. August 1764)

à Dresde ce 10 Aout 1764.
Monsieur et très cher Ami,
Vous venés, de m'envoyer des vers dignes de la plume de Chaulieu;
je vais Vous envoyer un Conte triste et trainant, mais s'il est moins
5 beau, que Vos vers, il vous marquera au moins mon désir à Vous entre-
tenir.
Timarethe, Athenien, s'êtoit appliqué dés sa jeunesse aux Sciences,

Figure 3: an automatically generated "paper" edition of a letter from the Berliner Intellektuelle project

Our recommendation here is for DARIAH to have a space where exemplary[23] projects are publicised as inspiration but also as contact points for newcomers that would have similar needs.

[19] Cf. the http://www.dnb.de/EN/Standardisierung/Normdaten/PND/pnd_node.html in Germany
[20] http://pdr.bbaw.de/
[21] Cf. our section on publishing and the fact that the book has lost some of its usefulness as a scholarly medium.
[22] Bertrand Gaiffe, Béatrice Stumpf, "A large scale critical edition: first translation of St Augustine's City of God by Raoul de Presle", TEI conference, Würzburg, 2011
[23] With the meaning of not necessarily cutting edge project, but projects that can boast a comprehensive application of the state of the art

3.5. Wrapping-up — recommendations

Observing what was actually used, or often requested, within the various projects we observed, but more widely within the community of "scholars with a budget"[24], we identified the emergency to deploy within DARIAH a Basic Service Kit (BSK), comprising some essential technical facilities helping projects to move ahead in their move to digital methods and be able to focus on the issues that are actually relevant for their research. Inspired from the very good support already provided by DARIAH-DE in this respect, we would incorporate the following components in such a BSK:

- Have a European single sign-in environment for humanities scholars on DARIAH services. National DARIAH affiliation would automatically give the corresponding right at EU level and conversely;
- Portfolio of PID services offering full EU coverage;
- Pool of Long Term Archiving facilities with a European plan for deployment and maintenance (with accompanying budget plans);
- Portfolio of interconnected virtual machines for installing and deploying basic tools at the service of humanities projects;
- Basic portfolio of tools for project management and communication: Confluence, Wiki, Etherpad, Jira in combination to blogs from hypotheses.org
- Basic portfolio of tools for technical developments: in particular Subversion for systematic versioning and Jenkins for continuous integration.

DARIAH, through its partner networks should urgently offer this kit for all scholars in Europe by pooling in capacities available in at least a core group of major contributors (France, Germany and the Netherlands at least, but we know several colleagues in Austria and Denmark could quickly join). Such a portfolio of services would be maintained (and updated) in the long term under the auspices of VCC1.

In addition to this BSK, an EU wide policy, supported by several core European DARIAH members should design a roadmap on the maintenance and deployment of basic editorial, data management and publishing tools. We insist here, and thus referring to the previous discussion in this document[25], that simplicity should be the lead so that the best compromise is found in terms of ease of deployment and expressive power. Note that this would mean shifting the priority of the "Archiv-in-a-box" task in DARIAH-DE from national humanities centres to less technically aware projects, like suggested more generally in this document. Maybe a renaming as "Archiv-in-the-cloud" could be anticipated.

To conclude this section, we can identify the topics where we would not recommend that DARIAH be taking actions. In particular, to quote Manfred Thaller, we have disregarded "the analysis of that information by tools reflecting the methodological requirements of the specific discipline and research problem"[26]. In coherence with our strategic stance that DARIAH should work in complement to the ongoing digital humanities projects, we strongly advocate that infrastructural work should not focus on too innovative solutions, unless there have been prior experimentations and a strong urge from the scholarly communities.

[24] i.e. they have a funded project

[25] And the huge amount of requests in this respect. See for instance, http://editingmodernism.ca/2011/01/in-search-of-a-digital-humanities-repository/

[26] Manfred Thaller, "Controversies around the Digital Humanities. An agenda" Introduction to the special issue of Historical Social Research Vol. 37 (2012), No. 3.

4. Conceptual Tools

The preceding section insisted on the technical components of a typical editorial workflow for the creation and maintenance of digital sources. Still, this rather task-oriented presentation puts the most essential aspect out of sight, namely the actual definition of the conceptual objects that a scholar may want to identify and qualify within the sources.
Indeed, it makes no sense to put scholars in contact with digital data if they are not in the position to judge or even shape the adequacy of the actual data that the technical environment presents to them or asks them to produce. The scholars should not become disappropriate of the essence of their work, namely the understanding of the conceptual organisation of their field.
As we can observe in all four projects that we used as reference for this report, scholars are indeed knowledgeable at two complementary levels:

- Their knowledge of the sources, their origination and possible history, and above all how a source can be part of a wider corpus of information with possible lacunae;
- Their own relation towards the sources as scholar, which depends on the underlying research question they have and how they see that the sources can provide evidence.

The aim of this section is to identify some possible directions to follow if we want to provide the scholars with the appropriate conceptual tools that would allow them to work smoothly with their sources in a digital context. Without any attempt to be exhaustive, we will give snapshots about possible core issues, existing models and actual priorities to set, in a context where we keep in mind that DARIAH should play an important role in disseminating these.

4.1. Getting a clear idea of what a source is

There are various ways of understanding data modelling issues in digital humanities. In the present report we centre our analysis on the source proper, whether from the point of view of its documentation (meta-data) or its description (transcription and/or interpretation).
As a matter of fact, data modelling should contribute in letting the scholar have a better understand of what the source, as a digital object, may be. In particular, it aims at qualifying when not distinguishing what pertains to the objectivity of the source and what results from the subjective interpretation of the scholar in the context of his research. This impacts on many notions that will have to be further implemented on the digital representation such as attribution, inline or stand-off representation, definition of the version of record for the digital library, etc.
In the following sections, we will give some hints and examples about the way a data modelling could be carried out to finish up with a specific focus on the notion of standard.

4.2. The FRBR model (<pron>ˈfɜrbər</pron>)

It would be very difficult to speak about conceptual tools without mentioning the FRBR[27] model, introduced in the library world to offer a better conceptualisation of the abstraction levels associated with a documentary item. There are two main reasons why this model is particularly important for a better understanding of digital process in the humanities:

- It is abstract enough to be easily understood and offers a clear articulation between the levels of representations of a human production. It is easily transposable to the digital world and helps understanding the digital ubiquity, both from the point of view of the life-cycle of digital objects and the appropriate meta-data descriptions that have to be maintained accordingly;

[27] Functional Requirements for Bibliographic Records, see http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

- Providing a better understanding of this model in the humanities is likely to create a close conceptual link between scholars on the one hand, and librarians and archivists on the other hand. This is essential to ensure the long-term collaboration of the two groups.

To illustrate this, we will now expose the main principles of the FRBR model and apply it to the two cases of meta-data description and digital document workflow.
The FRBR model is based on a four-level description of human cultural productions, depicted in Figure 1, and which can be roughly explained as follows[28]:

- The *Work* level is the highest level of abstraction (FRBR: "distinct intellectual or artistic creation") to the identification of a human cultural production, usually associated with a differentiating name, and possibly a specific author. *Homer's Iliad*, for instance, can be abstracted away to refer to the story it tells or the mythological elements it depicts, in comparison to Homer's other works or other works by different author;
- The *Expression* level reflects that a work can be realized in various ("intellectual or artistic") forms that bear their own specificities. This may indeed relate to various versions of a work, various translations, or any substantial difference that allows one to actually compare or confront the various expressions of a work. In the classical tradition, the various linguistic versions of the *Iliad*, form a typical group of expressions;
- The *Manifestation* level relates to the "physical embodiment" of an expression. Manifestations can be differentiated by the analyses of the physical, presentational or perceptive properties they have, but also by their specific *origination* in time and space[29]. The French 1975 edition of *Homer's Iliad* in the folio *Classique* series (with a Preface of Pierre Vidal-Naquet[30]) differs from other editions of the same translation (= expression) in layout, illustrative corpus, or even spelling;
- At the bottom of the FRBR model, the concrete (singular) existence of a manifestation is reflected in the *Item* level, seen as a "single exemplar of a manifestation". An exemplar of a book, with its characteristics of where it stands on the shelf of a library, and possible bundled with annotations on its pages will be a typical item in this respect.

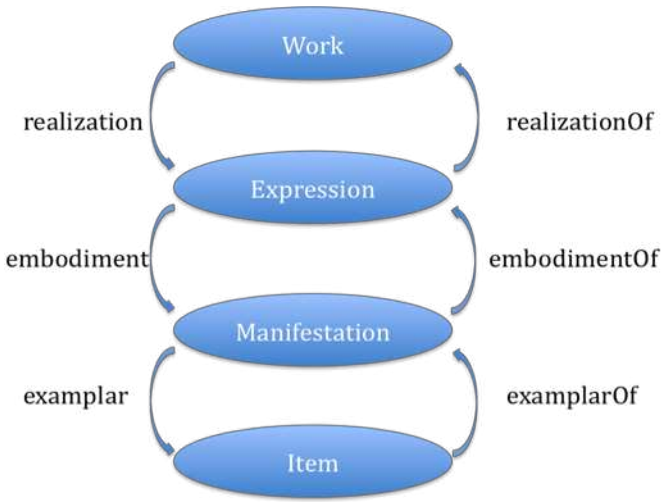


Figure 4: The main components of the FRBR model

[28] When appropriate, we recall the FRBR wordings; examples are ours.
[29] In the FRBR model, this is related to the fact that a manifestation may have a "producer".
[30] Note that Manifestations (but this is also true for the other sublevels of the FRBR model) may be compounds objects for differentiated entities at a higher level.

It is important that beyond the quite "simple" example of a written work, and its dipping through the FRBR model until one has a physical object in his hand, the actual boundaries between the levels of the FRBR model may not necessarily be clear when moving to artefacts with vague authorships (e.g. uniforms or weapons during a war period) or when the possible uniqueness of an object (e.g. a manuscript) makes it more difficult to actually implement all the levels of the FRBR model. Still, such a challenge is exactly part of the kind of intellectual activity (foregoing a more elaborate data modelling activity) that a scholar must take up when going digital to actually identify which information he wants to be represented, depending on his own view of what, for him, a work, expression, manifestation or item are. Indeed, the FRBR model may help to characterise also the absence of one or the other level in the model and thus characterise a lacunal information landscape.

As an illustration, let us see how the FRBR model could help us understand the representational level, and thus the meta-data, that a project such as Berliner Repertorium should consider in relation to the specific research issues that the scholars have.

In the Berliner Repertorium project the granular object at the core of the research is the psalm, seen as an abstraction across all its possible intellectual and physical occurrences, and thus clearly corresponding to the *work* level in FRBR. A psalm can be characterised by various meta-data descriptors, which depending on the tradition, may correspond to some reference classification schemes or simply be named by its first verse from Latin reference version.

Further down, the *expression* level corresponds to the various translations of the psalm, where the precision of dialectal variation is relevant here, since it is the fundament of the research project. Identifying clearly this level requires for the scholar to have a clear strategy as to how to distinguish translations from pure scribal variation, and mark his sources accordingly.

The *manifestation* level is more subtle to delineate because, contrary to the usual book example, there are no clear-cut groups of items that form a stable and coherent entity within a given expression. As a consequence, it becomes part of the scholarly work to qualify sequences of manuscript copies (*traditions*) for which similar qualifiers in content, time and place apply, and above all can be seen as research unit about which specific commentary work can be carried out. Indeed, a significant part of the scholarly work could be characterised as providing a comprehensive organisation for this manifestation level.

Finally, the *item* level is clearly associated with a specific manuscript, whether the artefact still exists within an archive or because its existence can be traced back or inferences from other pieces of evidence.

As we can see, such a FRBR based analysis has an impact on the management of the corresponding digital assets for the project at two complementary levels:

- Meta-data: the model may help the scholar to identify the adequate descriptors to be attached to a digital asset according to their relevance to one or the other level. This is indeed essential for corpus associated to several reference (printed) catalogues from which a coherent digital representation must be construed;
- Management of the digital content: depending on where the scholar wants to put the emphasis, or how far he wants his corpus to be reusable in different settings, he may want to focus on one specific level as the reference for his representation: e.g. manuscript per manuscript representation (item level), grouping editorial variants as one document per manifestation, or unifying the representation of all manifestation with one document per expression.

To conclude, even such a simple model as FRBR can lead to a better conceptualisation of source materials and their representation as digital content. We think FRBR could be (or should be) part of any introductory course to digitally-based humanities and probably part of any decent guidelines of good practices that an infrastructure such as DARIAH should publicize.

4.3. Data modelling

The importance for the scholar to participate to, if not directly lead, a data-modelling task in any digital humanities project should not be underestimated. This is usually the only way to make sure that a) his/her research concepts will be reflected in the data and b) he will get a clear idea about what kind of exploitation possibilities the data may offer.
Indeed, data modelling can be seen at identifying *affordances* on the data, that is reified (= encoded) concepts that offer ways for search, extraction, transformation or linking processes to be actuated. To take a simple example, in the TextGrid base encoding guidelines for texts, tokens are units that should systematically be tagged, and thus offer affordances for further search tools, or morpho-syntactic taggers, to actually anchor their work on this representation level.
As a consequence, modelling data is a strictly focussed activity that should not necessarily force the scholar to express his own language as an ontology. It will be all the more efficient for him/her, for the success of his project and for the reusability of his/her data to understand that the end product is subordinated to the actual operations that he/she expects to carry out on the data.
From a practical point of view, there is also a need to find the appropriate setting to optimise the scholar's contribution. Providing a too general modelling environment that is too open and far from his/her specific practices may frighten him/her away. In this respect, the CIMDI approach for instance, which basically provides a universal meta-data modelling environment, is not likely to be appropriate for most cases in the humanities. In the same vein, it would not make sense to offer the TEI guidelines in general as a basis for any type of data modelling purpose in relation to textual data. Identifying targeted subsets that already have a domain flavour is an essential starting point.

4.3.1. When both the model and the format exist – the Cendari case

The Cendari project offers an ideal scenario for the involvement of scholars in the data modelling process, in this case for the description of archives and their content. There is in Cendari several favourable factors that led to such a positive situation:

- The existence of both a conceptual standard for the description of collections in archives (ISAD(G)) and an XML format (EAD), implementing most of the components of the ISAD(G) standard;
- A core group of scholars (historians) with a very good knowledge of archival issues (by definition), but also of the reference ISAD(G) standard;
- Technical partners with data modelling skills with a strong experience in interdisciplinary work within digital humanities.

After an initial phase (mentioned above) where the project departed from an *ad hoc* data entry mechanism, a quick move was made to the definition of the optimal customisation of EAD for the (scholarly) purpose of the project. Within a few weeks, the historians managed to master the EAD format so well, that they could directly express their needs in terms of mandatory/optional elements in the schema, provision of specific fixed values or the addition of missing constructs. The corresponding results in turn are planned to be fed back to the EAD maintenance office to improve the future versions.
The lessons to be learnt here are clear and correspond to a generic scheme that we should keep in mind when working with similar projects:

- Importance of multi-disciplinary work: non of the parties would have been able to provide a coherent data model for archival description in Cendari. Scholars were to be acquainted with the conceptual tools and computer scientists had to transfer the lead to the scholars as they did so;

- Central role of standards as background for the expression of further needs: standards have been essential in providing a stable basis from which it was then possible to depart. The work would have never reach this quality of adequacy with the research expectation if we had been in the situation to design a format from scratch;
- Strong motivation of scholars to tackle technical aspects within a funded project: the goal oriented context provided by the project itself gave all parties the actual energy to reach a goal.

4.3.2. Linking model and format — the Laudatio case

Aiming at being a project with a strong infrastructural orientation, Laudatio has contemplated, right from the onset, to both design a generic model for the representation of meta-data associated with deeply annotated historical corpora and comply as much as possible to the representational background offered by the TEI guidelines.
As can be seen in the UML diagram in Figure 5, the model provides a very precise description of both content and annotations where in particular, the responsibility behind each information source is precisely provided. Taking distance and trying to understand the role of the model for the scholarly process, we can indeed identify a twofold objective, mainly related to the capacity of scholars with little technical background to actually "understand" the data:

- Offering ways for scholars to foresee what kind of queries will be applicable to the data
- Help scholars who are about to contribute and compile some specific historical corpora to adopt a clear methodology to design the corresponding meta-data they want to attach to them

It should be noted that the actual model is by far more abstract than the final TEI implementation[31] so that it is easy to disseminate even to communities that are not (yet) au fait with the TEI technicalities.

[31] Implemented as an ODD specification

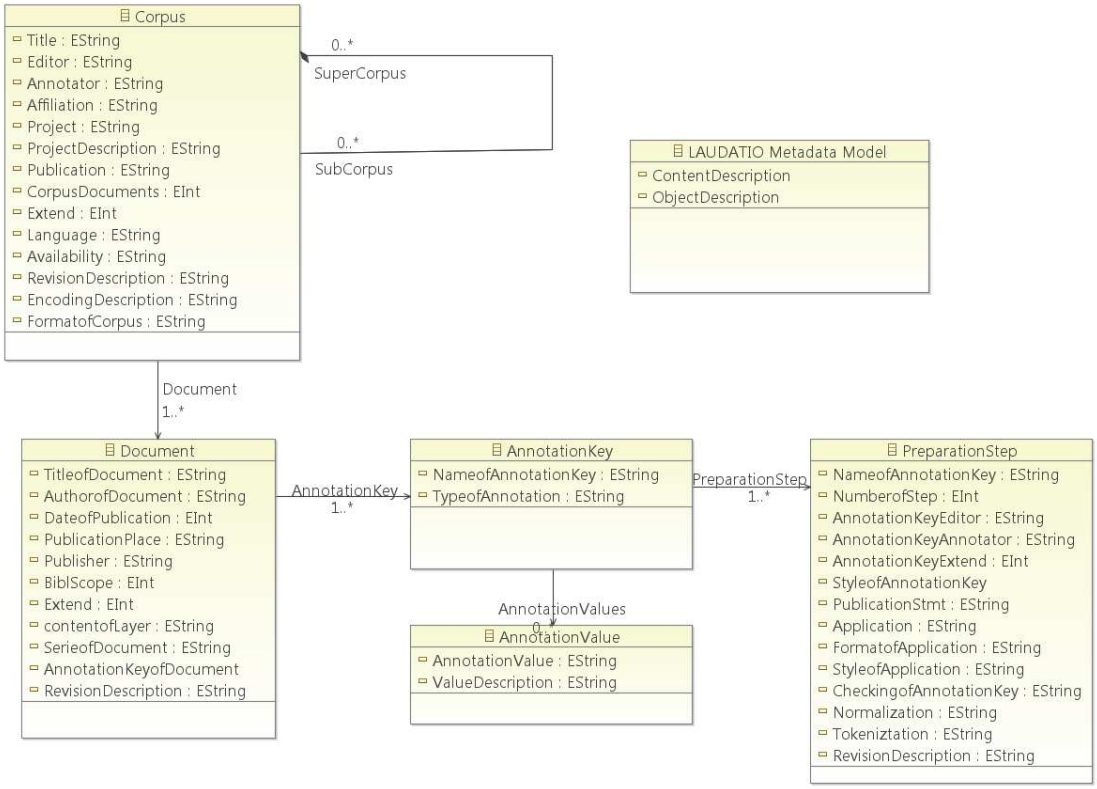


Figure 5: The meta-data model in Laudatio

4.3.3. A memory of modelling choices

Not only is it important to have a modelling tool at hand that is adapted to both scholarly needs and current technological context, but one should also have the capacity to record modelling choices and further share them within and outside the scholarly project.
DARIAH should thus have a clear policy to accompany scholars in stabilising their conceptual choices, usually reflected in the formats and standards they decided to use. Such a strategy should clearly reflects the various stages at which such an activity can take place within a project and provide the corresponding editorial and technical infrastructure.

- At an initial stage, projects should be provided with the means to specify and document their own practices. Possible actions range from basic awareness in data modelling to technical education on modelling platforms such as UML or ODD;
- When the data model, and corresponding formats, are to be maintained by a wider group (usually with a pre-normative objective), DARIAH should offer the corresponding subset of the Basic Service Kit: Subversion for maintaining versions of the model, Jira to record evidence about evolutions in the model, a Wiki for gathering the appropriate documentation;
- At the standardisation stage, as we will see in the next chapter, DARIAH should stop having a core role but support scholars to join the appropriate standardisation bodies.

We see that such a framework does not leave room for the deployment of a centralised schema registry for which we have not observed a real demand from ongoing digital humanities projects. We advocate here some simpler solutions based on simple tools that are likely to be quickly deployable among a quickly evolving digital humanities landscape.
Besides, we can see how central standards are in such a strategy for DARIAH, as we shall see in the next section.

4.4. Role of standards

The main issue about 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 three following requirements on 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

A standard is not a regulation. There is no obligation to follow 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 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 what happens with the Text Encoding Initiative for text representation standards and to a lower extend 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) the awareness about the appropriate standards and the interest to adopt them, and b) the cognitive tools to help them identify the optimal use of standards, through the selection and possibly customisation of a reference portfolio. Our experience acquired while working with numerous projects (among which those cited in this document) which were in the need of adopting existing standards is that there was always an initial phase through 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 this 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 first actually 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[32], 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

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

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

4.5. Recommendations

The preceding sections could potentially lead to many possible action points in DARIAH. At this stage, we could boil this 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
- 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

5. Political tools

5.1. Why help is needed

The proceedings sections were closely related to the needs of humanities projects in relation to their core scholarly and thus technical endeavours. Still, we should not disregard here the need for some additional support as to how the scholars may position themselves with their political surroundings when they either want to acquire or disseminate information. In the first report, we outlined for instance some of a possible open access policy for DARIAH, which should be further instantiated as concrete services (i.e. support) to scholarly communities to go in this direction.
In this section we will thus address three complementary topics for which we think DARIAH should provide guidance, namely the dissemination of primary sources, the relation to libraries and finally scholarly publishing.

5.2. Digital primary sources — Copyright — plagiat

After having moved ahead to a point where most of their research output, in particular their primary sources, are available in digital format, humanities projects are confronted with the need of getting some concrete guidance as to how they can actually disseminate their results widely online. Concretely, the following questions are recurrently asked by scholars going digital:

- Will I be spoiled from my own results if I put all my content online?
- Under which licensing scheme should this be disseminated? Should I ask a lawyer?
- If some other works have been done on the same source, especially if some commercial publishers published these as books, may fall under some accusations of plagiarism?
- How much can I point or reuse to some existing published sources, such as catalogues, bibliographical information, existing editions or even secondary literature?

As a matter of fact, the global access and re-use situation in the humanities is still hampered by the traditional culture of transferring rights to commercial third parties rather than

favouring barrier free access to the scholarly community. This is not just the fault of publishers, who are just bringing their business forwards, but of the academic environment that has not explicitly developed an wide spread understanding of this issue and accordingly has not been in the position to take action.
We think DARIAH has a very important role to play in to provide more awareness, but also to give guidance as to which practices should be adopted to go towards an open scholarly space of digital sources and publications. We should strive at giving clear and simple messages that should in no case reflect a fully comprehensive coverage of the issues, but contribute to the necessary cultural evolution we are looking for. The core messages that we should contribute to disseminate should be based on the following principles:

- Never transfer exclusive rights to a commercial third party
- Favour licensing schemes that will increase the fluidity of digital scholarly knowledge, while guarantying adequate quotation of the source
- Put as much of the scholarly production online, comprising the actual source formats

In this perspective, the Cendari project has issued an exemplary "data sharing agreement" which articulates the issues of data re-use, production and dissemination, mainly based on the application of the Creative Commons CC-BY licence. The statement in available as an annex to his report for other similar projects to inspire from it.

5.3. Relation to the libraries, archives and museums

Many domains in the humanities have relied for years upon the availability of documents, cultural objects or artefacts in libraries, archives and museums as the basis for their research. It is thus important for DARIAH, which aims at coordinating digital services and competences in the humanities to identify how this may relate to a similar shift to the digital world that is taking place in cultural heritage institutions (CHIs).
Still, the topic is potentially so wide and the actual policies are just being shaped in CHIs that we can only present here some general trends that we illustrate by means of specific use cases.

5.3.1. The TEL – Europeana scenario

At the highest level of (possibly disembodied) abstraction, we can present the possible role of aggregator such as Europeana, or rather, possible interface between the scholar and the aggregator.
In a context where TEL and Europeana potentially represent an extraordinary wealth of data for scholars, it is important to identify the main stumbling blocks that could hamper a closer collaboration scheme between the two communities:

- First, as a kind of paradox, the actual holdings available from Europeana are often felt as too large for scholars who develop a trustful relation with specific collections where they know they will find their primary sources[33]. Conversely, specific libraries or archives may just be missing within the Europeana network whereas they are considered as essential for some specific scholarship activities;
- Second, the precision at which data is described is often too low for a proper scholarly work, even if useful at the discovery stage. As we have encountered in the Cendari project for archival content or BerlInt for the manuscripts, scholars may have very specific requirements on the precise description of collections or documents;
- Finally, there is a political issue with regards the licensing scheme, where the one adopted by Europeana, based on waiving rights associated to data (CC0) may be contrary to the basic scholarly principle of attributing any information to its source.

[33] In literature, they will globally know where manuscripts are; in war studies, the core archives are well identified; etc.

Still, we are doomed to establish a long-term dialogue between scholarly infrastructures such as DARIAH and cultural heritage institutions at the service of a better access and dissemination of research results. This should lead to precise schemes by which scholars will be able to search and retrieve as much information as possible from the CHI network in Europe and conversely, have simplified means to make their results visible through Europeana and TEL portals.
Such a relation does not indeed prevent, on the contrary, the establishment of closer relations between scholarly communities and specific libraries. We will see two such cases in the next sections.

5.3.2. When the library is going digital… designing digital services

To understand better the developments that are needed to create a synergetic scenario between scholars and the library, let us observe the recent development at the university library of the Humboldt University. In the recent years, most library activities in the university have focused on planning, building up and launching the quite extraordinary Grimm Zentrum, an integrated library environment offering very high standards of student oriented services (large on-hand access to books, extended opening hours, multimedia facilities).
When appointed in 2011, the new director of the University library identified the need to provide digitization services within the library and to do so, made a general call for application within the university departments to identify actual needs. The call was indeed exemplary as it combined two constraints related to a) the expression of a real research need and b) the knowledge of a sizable and coherent corpus that could back up this research.
This targeted digitization campaign is indeed successful and both the BerlInt and Berliner Repertorium projects benefited from these services. Moreover, it was the opportunity for the library to fine tune its strategy by a) offering persistent reference and long term archiving for the digital assets and b) experimenting further state-of-the-art OCR technologies to bring in searchable full text that could in some cases also be used for bootstrapping a transcription.

5.3.3. The visionary scenario – co-development of digital services

In the case where libraries are more advanced in their digital agenda, we can foresee more elaborate services where library and research symbiotically develop data management and re-use scenarios. We use here the example of the collaboration that has taken place for two years between the BerlInt project and the State Library in Berlin (StaBi).
The StaBi has for several years pursued its digitization process in a classical twofold approach:

- it developed an integrated digital catalogue of its collections coupled recently with the design of a completely standardised export interface in EAD[34];
- it has initiated a digitization strategy whereby they offer to scan specific collections needed for scholarly activities and offer the corresponding basic infrastructure (long-term archiving, persistent identifiers).

During its collaboration with BerlInt, the StaBi clearly understood that going further with the digital content, that is providing informed transcriptions of the available documents, strongly relies on the expertise of the scholarly group that has a close interest in the collection. Following various discussions and actual experiments, the following scheme was thus designed:

[34] There is indeed a strong similarity between historical and manuscript collections in libraries and archival collections like described in the General International Standard Archival Description (ISAD(G)), the background conceptual standard for the EAD format.

- The library provides a deep access to its meta-data in EAD to bootstrap the meta-data of the TEI transcription[35];
- On a regular basis, the research group identifies "stable" versions of their transcriptions (aka *releases*), which are then transmitted to the library to be integrated in their digital repository;
- These releases are curated and made public with a persistent identifier and long-term archiving, so that other groups can re-use the corresponding information;
- The transcriptions are also made available through the project's own repository environment with links back to the StaBi scans.

This collaborative scheme provides several advantages which to our view make it a seminal construction for future library-scholar collaborations:

- It relieves the scholarly group from thinking of core data management services such as persistent identifiers and long-term archiving;
- It creates a virtuous circle whereby both groups contribute to the elaboration of trustful data with the best of their respective knowledge;
- From a more general perspective it provides a concrete way of fluidifying the exchanges between scholarship and the library world.

5.3.4. Recommendations

While we agree with Manfred Thaller's observing that "It is extremely welcome, that librarians nowadays take an active interest in providing access to digital information.", we consider that it is the role of DARIAH to develop a generic partnership framework to explore such interactions for major library environments (and maybe archives).
To do so, we need to abstract away from various experiences that are currently being carried out to shape a general charter of good practices for libraries and research projects. Such a charter, which would be annually updated would state some core principles related to the respective roles of both sides and aim at facilitating the simple dissemination and re-use of knowledge.

5.4. Publication as part of the infrastructure for research

5.4.1. A missed opportunity?

In the early years of DARIAH the topic of scholarly publications has been intentionally left out from our work plan for three main reasons: a) it was not clear in which direction the A&H community wanted to go in the digital world[36] , b) there was not a clear offer of new services in Europe beyond what we had with the private publishing sector and c) there was an explicit assumption that all the corresponding issues would just be dealt with by Driver/OpenAire.
In between, the situation evolved dramatically with various communities in the humanities discovering social networks and the immediacy of online communication. It has thus become essential for DARIAH to incorporate scholarly publishing as part of its work plan in VCC3[37].

[35] I.e. feeding in the TEI <msDesc> element.
[36] With institutions mainly focussing on ranking existing journals, cf. http://www.esf.org/research-areas/humanities/erih-european-reference-index-for-the-humanities.html
[37] This change of policy has somehow been validated with the ERC asking us to contribute to the definition of their open access policy in the humanities

5.4.2. Considerations for DARIAH

Above all, it is important to perceive the complexity of scholarly publishing in the humanities considering the variety of domains involved, which somehow corresponds to even more variable publishing cultures. This said, there is a wide-spreading feeling that the current publishing environment is inadequate because it offers a limited dissemination of the research results, but also because the prominence of books is not always adapted to the reality of modern scholarly practices.
For DARIAH, the emergence of new communication modes such as blogs[38], which offer an immediate commentary channel during a research project and the stabilisation of major partners such as OpenEdition within the European community has offered some means to answer these expectations. Still, it appears to us impossible to contemplate offering "new" services without actually reflecting the existing debate about open access to publications and the corresponding core information to be provided to scholars (importance of depositing in repositories, current economy of scholarly publishing, new models, etc.). Finally, such a task should not be carried out in isolation but integrated within a comprehensive scholarly publication policy (primary and secondary).

5.4.3. Conclusions and recommendations in the domain of publications in the humanities

Even if the topic of scholarly publishing was not addressed during the preparatory phase of DARIAH, it is now incorporated on our working plan as a new task "Publishing and dissemination" in VCC3. Concretely, we need to contribute to moving journals and books to public open access platforms as part of the normal infrastructure provided to scholar.
To achieve this we make the following strong recommendations for the future involvement of DARIAH in the domain of open access:

- Contribute to the dissemination of open access awareness in the humanities
- Provide an orphan archive for scholars in the humanities that do not have access to a publication archive. Such an archive could also host/mirror DARIAH related publications
- Liaise with OpenAire to guaranty that our policy is compatible with EU requirements for Horizon 2020 and ERC projects
- Widely propose services from DARIAH-FR (OpenEdition) to the general DARIAH-EU community
- Contribute to setting up a coherent picture of the publishing landscape that could help inform decision makers to adapt their current policies, in particular in terms of scholarly recognition

5.5. A wider political plan

Rec.: contribute to establishing a systematic liaison between national DARIAH members and the corresponding funding agencies. Such a liaison should foster various levels of possible collaboration:

- Establishing a forum of projects, where DARIAH can identify both new needs and results that should be further spread within a wider community;
- Work out a basic set of recommendations that new projects should fulfil;
- Provide contact partners at European level where projects could confront their methods with colleagues having similar scholarly objectives;

[38] 489 French scholarly blogs on hypotheses.org as of 31th Dec. 2013, and already 39 on de.hypotheses.org/

- Encourage multilateral funding schemes between funding agencies to facilitate such exchanges when they lead to joint projects

6. Educational aspects

The preceding sections have shown the importance of *collective* intelligence in DARIAH and the need for scholars to identify spaces where they can benefit from the experience and contributions of others. This has some impact on the educational policy that we should foster in DARIAH.
Indeed, we should consider educational and training aspects as an endeavour (carried out in VCC2) that involves various levels of disciplinary and geographical competences, namely:

- *Local networks of multidisciplinary scholars* involved in teaching digitally based methods in a variety of humanities areas. Such networks are likely to favour cross-fertilisation across disciplines and help less advanced domains benefit from the experience of others. When there is a critical mass of scholars, DARIAH should encourage the identification of a portfolio of digital humanities seminars allowing students to benefit quickly from a variety of training possibilities;
- *Structured communities* that are likely to organise, usually at a national level, awareness and training events to show return on experience from existing projects and demonstrations of the most recent tools. Such communities should be the focus of attention of a training policy in DARIAH by offering them the necessary organisational support, but also by encouraging DARIAH members to provide the adequate funding for such events;
- *Wider European network* such as the recently selected DiXit within the Marie Curie framework that provides a space where students within a semi-focused community of practice can move around in Europe in the context of their Master or PhD project. DARIAH should help identify mature domains in the digital humanities where networks similar to DiXiT could be put together.

Within this landscape, general teaching events such as DH or DARIAH summer schools should be clearly positioned in complement to other types of training possibilities according to a subsidiary principle. Their focus should be on a) offering some core training in digitally based methods (for instance through the presentation of the Standardisation Survival Kit and the Basic Service Kit) and b) offering a picture of the variety of communities in the humanities that have moved to digital method. This should also be a place where recent advances from the computer science discipline in searching, visualisation or information extraction could be shown.
To conclude, the spreading of digital humanities methods across a variety of communities should be accompanied, not necessarily through the creation of over-specialized DH departments, but more as a momentum towards all possible disciplines. A particular support could be provided by DARIAH by encouraging the creation of two complementary resources:

- A digital curriculum evaluation grid, which would continuously identifies the skills and competences acquired by students in curricula that would boast a DH (or DARIAH) flavour. This could lead to the creation of a label “DARIAH seminar”;
- A review journal that would provide in depth state of the art articles in all fields of humanities computing in order to provide reference material (existing projects, overview on possible methods, tools and data models). In the format of the Living Review series, such a journal would be branded by DARIAH and would accompany its development.

7. Going further

The present report has obviously taken a very personal approach in identifying the priorities for DARIAH, seen from the point of view of its construction phase, in the domains of services and technologies for its users.
The main lesson I have learnt in working with so many projects in the digital humanities domain in the last 20 years (!) is to identify the major tension that exists between them striving to be original in their endeavour and the actual simplicity of the technical solutions that could be used to fulfil most of their infrastructural needs for dealing with their digital objects. There is here an important issue of scholarly recognition where it is difficult for a project to boast being "just" in the state of the art, because they have used widely spread standards, because they have used simple technological components, or because their projects are just easy to apprehend from a digital perspective.
And this is indeed good so. Scholarly achievements can often be elsewhere when the added value is more on the quality of the resulting data or all in all in the actual contributions made by scholars in their own field. It may not be necessary to have a cutting-edge paper at the DH conference to be a very good digitally-based scholar. Still, taking up Manfred Thaller's leitmotiv, we should not disregard the important role of analytic tools. We have not addressed these so much in this report, but the issue is clearly related to the final point we will make about our core stakeholders.
As alluded to several times in this report, we need a global strategy with regards funded projects in the humanities to accompany their move to digital methods. Such a strategy should be based on the recognition of some basic services about which DARIAH members and DARIAH-EU make a strong commitment. As an outline, and indeed recommendation for such systematic services, we should consider:

- Disseminating information about the project, probably by offering long term hosting of their web site
- Provision of ground services needed for securing the projects' technical work
- Taking up any re-usable development or expertise to feed DARIAH VCC activities
- Offering a joint approach to making the project results sustainable
- Feeding back the projects' results to the corresponding wider communities as well as the funding agency (as best practices)

Implementing such a strategy will require the combination of three complementary factors:

- Awareness of the DARIAH actors that funded projects are our main stakeholders for service provision
- Maintenance of a strong dialogue with national and European funding agencies
- Favouring the emergence of methods and training networks such as NeDimaH and DiXit, but also local or thematic competence networks, to facilitate the exchanges between projects and DARIAH actors.

As a final word, we would like to stress that DARIAH should play a complementary role to the actual nationally or EU funded projects whereby such projects are the place where *innovation* is carried out, whereas *consolidation* aspects are conferred to DARIAH.

8. Annex — Cendari's "Data Sharing Agreement"

Data Sharing Agreement

As a research infrastructure project funded by the European Commission and a formational contributor to the ESFRI roadmap through its parent infrastructure, DARIAH, CENDARI's primary concerns are to facilitate the access to and interrogation of the primary data

required for humanistic research. CENDARI is not intended as an access portal for the general public (though members of the public may find and use it), nor as a primary delivery mechanism for a broad range of cultural content (though a lot of content will be available through it). Instead, CENDARI's primary role is to become a powerful workbench for the deep investigation of humanistic questions, capturing the tacit knowledge of the scholar-user and the archivist and enhancing those perspectives through its data curating, mining, query and visualization tools.

As such, the project places a very high value on *attribution* of the data it delivers, creating traceability for all elements in the infrastructure back to a single source. We strongly believe this is the only way to create a scholarly digital ecosystem with equivalent conditions to those of traditional research environments for the results to be transparent and reproduceable, the most meaningful quality standard for any scholarly work.

A single project cannot mandate usage and attribution patterns for all data and all institutions. What follows here, however, are the principles by which CENDARI will determine appropriate standards for sharing and exposing content clearly within its own technical control, and what the project's baseline recommendation to its scholarly and archival collaborators and contributors in this respect will be. As such, CENDARI hopes to set a precedent for sustainable good practice within digital humanities research and research infrastructure.

This good practice will be based on a commitment to the widest possible usage of the CC-BY standard. CC-BY (Creative Commons Attribution 2.0 Generic) allows the user/researcher is free to share (to copy, distribute and transmit the work), to remix (to adapt the work) and to make commercial use of the work under the condition that the work is attributed in the manner specified by the author or licensor. See: http://creativecommons.org/licenses/by/3.0/

1. Data or documentation created by CENDARI for CENDARI: this data will be available automatically for use under a Creative Commons CC-BY license, with reference back to CENDARI itself or the contributing partner.
2. Publicly visible data created by scholar-users within CENDARI and disseminated through its platform: this data will also be available under a Creative Commons CC-BY license, with reference back to the author. User-data within CENDARI will only be made public at the user/creator's request and this standard need not be applied in the case of data created using CENDARI tools or content, but hosted and disseminated independently from it, or to data created and hosted within CENDARI but only held for private visibility.
3. To maintain this standard, archival metadata and digital content available through CENDARI will also need to be available as CC-BY, attributable back to the partner archive. We realize that this may be a complicated standard for some archives to apply to some data, and will work with our partners to ensure maximal inclusion of their data that does not create conflicts with any other preexisting copyright agreements. As CENDARI will not normally be hosting full data-sets from participating archives, but only a registry, lower-quality facsimiles (eg image thumbnails), metadata, links etc., CENDARI's services to partner archives will be limited to accurately reflecting holdings, and providing access to them in seamless combination with other, cognate collections. Every effort will be made to respond quickly to any problems in the CENDARI register of partner data, to protect the rights of the content owners, including but not limited to maintaining data security, and to ensure that data provenance is clearly reflected within CENDARI. Any required transformations to archive data will be specifically agreed with the archive in question, and the results of any enrichments will be shared at no cost with the partner archives.