COLLABORATIVE EUROPEAN DIGITAL ARCHIVE INFRASTRUCTURE

Project Acronym: CENDARI
Project Grant No.:   284432
Theme: FP7-INFRASTRUCTURES-2011-1
Project Start Date: 01 February 2012
Project End Date: 31 January 2016

Deliverable No. :	2.4
Title of Deliverable:	Sustainability Plan
Date of Deliverable:	31 January 2016
Revision No.	1
WP No.:	2
Lead Beneficiary:	TCD
Authors and Contributors (Name and email address):	Jennifer Edmond edmondj@tcd.ie Francesca Morselli morsellf@tcd.ie
Dissemination Level:	PU
Nature of Deliverable:	R
Abstract:	The CENDARI project committed in its original description of work to the production of a Business Plan (Deliverable 2.4), outlining a method statement for ensuring that the outputs of the CENDARI project would be available to future users. The project team decided to rename this deliverable the CENDARI "Sustainability Plan" and this document sets out the plan which is based upon a multidimensional conceptualisation of what CENDARI is and the value of its assets, as well as on the fundamental understanding of a digital project as useless if it does not 'evolve and involve.' This document presents not only the actions that have been and will continue to be taken by the CENDARI team, but also interrogates the topic of sustainability of research infrastructures

	in the humanities and social sciences. This document describes and guides the actions that will guarantee the CENDARI project's continued impact and continuity after its funded period, and it also provides a model for other, similar initiatives, seeking to plan for and conceptualise their own sustainable futures.

WHAT DOES IT MEAN TO SUSTAIN A RESEARCH INFRASTRUCTURE?

Experiences and recommendations from the experience of the CENDARI Project

Authors: Jennifer Edmond, Francesca Morselli (Trinity College Dublin)
Contributors:
Natasa Bulatovic, Milica Knezevic (Matematicki Institut Sanu u Beogradu)
Emiliano Degl'Innocenti (Societa Internazionale per lo Studio del Medioevo Latino /Fondazione Ezio Franceschini Onlus)
Jörg Lehmann (Freie Universität Berlin)
Carsten Thiel (University of Göttingen)

Table of Contents

1. Introduction

The CENDARI project committed in its original description of work to the production of a Business Plan (Deliverable 2.4), outlining a method statement for ensuring that the outputs of the CENDARI project would be available to future users. In the course of the development of this document, we have come to understand that CENDARI is anything but a "business" (an organisation to support commercial transactions between suppliers and customers): indeed it is many things -- a set of communities, a set of tools, a knowledge base, a set of resources -- none of which meets that definition of a business. The attempt to develop this plan in a manner that neither under- nor over-estimated the nature of the project and the value of its outcomes has been a fascinating challenge, met by the text that follows, which we have decided to rename the CENDARI "Sustainability Plan."

It was felt at the outset of the project that sustainability for the project outputs would not be an issue, and that our close relationship with DARIAH would guarantee that our results would be maintained. But a complex technical infrastructure cannot simply be frozen in time and expected to continue to meet evolving needs. The communities that have built CENDARI will need a more active platform to ensure this, and the assets produced will need a broader base than even DARIAH can provide in order to maximally benefit its potential users. For this reason, the plan that follows is based upon a multidimensional conceptualisation of what CENDARI is and the value of its assets, as well as on the fundamental understanding of a digital project as useless if it does not 'evolve and involve.'

This document therefore presents not only the actions that have been and will continue to be taken by the CENDARI team, but also interrogates the topic of sustainability of research infrastructure in the humanities and social sciences. Based upon the current research and practice of other, similar projects, it proposes a comprehensive understanding of infrastructural assets, and a model for sustainability that considers not only the most obvious end product of the project (in CENDARI's case, a virtual research environment), nor even just the project's tangible, reusable and customisable components, such as data and tools, but also the project intangible achievements, such as the know-how and the expertise developed during its duration, and the community of users that has been built through the project's testing and outreach phase.

Although the primary function of the document is to describe and guide the actions that will guarantee the project's continued impact and continuity after its funded period, it should also provide a model for other, similar initiatives, seeking to plan for and conceptualise their own sustainable futures. Research infrastructures and digital projects in the field of the arts and humanities have struggled to establish successful models according to which they can

strategize and plan for the period after their funded development ends, when seeking reuse and ensuring maintenance become the greater challenges. This is not because the challenge of sustainability is new, but merely that best practice remains an unresolved issue in this area, an already challenging space due to the natural tendency to view it as a late action for a project, rather than an awareness to be woven throughout activities from the conception phase.

1.1. What does it mean to sustain a complex Research Infrastructure?

Recently, the term "sustainability" has achieved such prominence that one has the feeling that the problem of sustaining research data and technical elements of an infrastructure project had emerged only recently as a requirement. In reality, issues of sustainability have been discussed and debated from the time of the first digital libraries and other memory institutions that first explored digitisation and metadata encoding of their holdings. Already in 2003 the *Council on Library and Information Resources* published a booklet entitled "A survey of digital cultural initiatives and their sustainability concerns", where the author, Diane M. Zorich, introduced a number of the factors that need to be considered when discussing sustainability, elements she characterised as 'threats' to the development, such as continuity of funding, data preservation (including choice of standards and longevity of data) and flexibility to change business models.

As digital research infrastructures flourish and proliferate in number and form in the field of Arts and Humanities, the urge to somehow 'solve' the problem of sustainability has become ever more urgent. Like any enduring challenge, sustainability of complex digital resources has been subjected to both scholarly investigation and practical reflection. What a survey of this literature shows, however, is that the details of any individual interpretation of how to face this challenge are highly dependent on how the project or infrastructure views itself. For example, the digital library example cited above reveals a very traditional stance, in which the digital project is viewed not so much as a new kind of initiative, but an institution equivalent to its bricks and mortar precursor, which does not need to prove and define its value on its own term, but to manage threats to its fledgling existence. Another common indicator of this conceptualisation of the digital project or infrastructure development as a form of organisation or institution is the focus on creating a 'business model' for maintaining the services and the organisation (including staff) created[1]. APEx[2] for example established a Foundation at the end of its funded development, as this organisational continuity was

---

[1] See, for example, Nancy L. Maron, K. Kirby Smith and Matthew Loy. "Sustaining Digital Resources: An On-the-Ground View of Projects Today". http://www.sr.ithaka.org/publications/sustaining-the-digital-humanities/
[2] http://www.archivesportaleuropefoundation.eu/

considered to be critical in order to maintain the vital functions of the infrastructure and to further connect with other projects and similar initiatives.

Alternatively, when project outputs are viewed as a *"tool or technical platform"* a sustainability proposal will primarily take into consideration issues and practices such as migration and curation of elements such as the repositories where the data are stored and continued maintenance of work environments and specific tools. The TextGrid project[3], for example, has been hugely successful in rolling its activities forward over a long period, continuing to make its services available to users. At its best, this approach results in a broad focus on software durability, documentation of processes and the modularity of services.[4] A further possibility along this line is for a project to view itself as a provider of *data*. At its worst, this paradigm can lead to the kinds of fast-ageing, unknown data silos and digital editions that littered the early days of digital humanities project development.

All of these perspectives originate from that of the resource creator, however, and it is not surprising that sustainability statements formulated by people who build things would tend to focus on ensuring the availability of the things they have designed, built and can (to a certain extent) control. But sustainability has a user side as well, and it is in the nature of short-term, project-based funding that project trust, visibility and tuning to the needs of these users, out there, in the wild, may still be in its infancy as the project team adjourns. The importance of these assets cannot be underestimated, however, as these are the elements that will assist project results to be used and more importantly reused. The LAIRAH[5] project has contributed significantly to our understanding of what factors enable digital projects and tools to be found and adopted by users, and these findings don't always match to the high-level statements of sustainability plans. From the results of this project we can see another model for the sustainable project to emerge, in which the *"communication and branding"* of the project is a key element of its success, and *"handover to the user community"* is a part of the thinking behind the long term use model.

The CENDARI Sustainability Plan seeks to build upon the good work done according to these models and propose a new state of the art of the concept of sustainability for research infrastructures. This new paradigm will take into account all of these perspectives, following a hybrid approach that seeks to understand the many facets of what the project has created, understand their value for current and future user groups, and sustain those elements in one or more formats that will best allow them to connect with their users. By tracing previous experiences, successes and pitfalls, this document elaborates a "toolkit" for the sustainability

---

[3] https://textgrid.de/en

[4] Buddenbohm, Stefan, Harry Enke, Matthias Hofmann, Jochen Klar, Heike Neuroth, and Uwe Schwiegelshohn. "Success Criteria for the Development and Sustainable Operation of Virtual Research Environments." *D-Lib Magazine* 21, no. 9/10 (September 2015). doi:10.1045/september2015-buddenbohm.

[5] Warwick, C., Terras, M., Huntington, P. and Pappa, N. "If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data." *Literary and Linguistic Computing* 23, no. 1 (2008): 85–102.

of the CENDARI project and its infrastructure, in the hope not only of sustaining CENDARI and contributing to the DARIAH ERIC's capacity to advise on sustainable development, but also of supporting future project teams in thinking about their own projects and their legacies.

2. State of the Art of Sustainability Plans of EU Projects and Research Infrastructures

2.1. Recommendation from the Digital Libraries Experience

As Digital Libraries (DLs) have the longest tradition in the preservation of digital objects and their related metadata, it is advisable to investigate which route they have taken in relation to the topic of "sustainability". From a literature review it is notable that when it comes to sustainability, DLs have adopted a twofold approach. On the one hand they have approached the technical side of sustainability, by focussing on plans related to the actual preservation of digital objects, data and tools. On the other hand they also focussed on aspects related to project management and financial sustainability.

This second approach has been analysed in a research undertaken by JISC and Ithaca University from 2008, which was consequently published in the report "Sustaining Digital Resources: an on-the-ground- view of Projects Today"[6] [7]. The research project analysed digital libraries and digital projects from a number of countries in different continents; moreover, it is interesting to notice that the research continued until 2014 and resulted in a number of subsequent recommendation documents.

The analysis of the JISC/Ithaca report depicts a scenario where project management strategies are crucial in the sustainability of a project after the end of its funding period. The report identifies a set of activities and indicators that reflect the wellbeing of a digital project and its ability to establish a sustainable project:

1. Dynamic and engaging leadership
2. Identify a value proposition: the identification of the end-users is crucial in the development and maintenance of a digital project. End users are those who are going to find the resources valuable. Moreover the community of users can constitute a powerful community, trustworthy and interested in the choices and development of the digital library will undertake.

---

[6] http://www.sr.ithaka.org/wp-content/mig/reports/4.17.2.pdf

[7] The same Institute has later in 2014 developed a Sustainability Implementation Toolkit, available at http://www.sr.ithaka.org/publications/sustainability-implementation-toolkit/

3. Manage costs creatively: keep the costs down, for example by seeking partnerships and manage these partnerships to keep them healthy; or by being hosted by a stable institution/ organization which could help to reduce some costs.

4. Cultivate diverse source of revenues i.e. subscription model, individual donation models,

5. Establish realistic goals and a system of accountability

On a more pragmatic level, the document suggests a series of actions, which are recommended to every project that is about to end its funding period:

1. Desired post grant impact: the identification of the goals and impact that the DL wishes to accomplish after the end of the funding period is necessary to realize the costs that could be involved in the realization of that impact.

2. What the set of goals are?

3. Activities involved (content, technical requirements, staffing…)

4. Costs

5. Revenue plan

An older document, dated 2002, and written by researchers within the University of North Texas Library[8], focusses on the technical aspect of data preservation within Digital Library projects. This document indicates a series of steps to be taken for a "healthy" data sustainability plan:

1. Life Cycle Assessment of the Digital Resources: The review pays particular attention to the preservation and management of metadata sets, which are needed to support various preservation approaches including migration and emulation.

2. Draft of a metadata architecture: Metadata, to be most valuable, both for the users and owners, needs to be consistently maintained throughout the process. Creating documentation that governs and informs the metadata creation steps and procedures in a consistent and uniform manner is among the most important steps in metadata creation.

3. Metadata Creation Workflow: This problem is coupled with the fact that metadata creation is not sufficiently incorporated into the tools for the creation of elements' record to rely solely on the creation process. As standards groups and vendors move to incorporate XML and RDF architectures in their word processing and database products, the creation of metadata as part of the origination of the object will be easier.

4. Metadata Creation Tools: appropriate means for creating a metadata repository database, appropriate indexing and harvesting software and search engines to use

More recently, the IFLA Manifesto for Digital Libraries[9] focuses instead on the concept of sustainability from the point of view of data interoperability: "*Interoperability and*

---

[8] Daniel Gelaw Alemneh, Samantha Kelly Hastings, and Cathy Nelson Hartman, A Metadata Approach to Preservation of Digital Resources: The University of North Texas Libraries, 2002.
http://firstmonday.org/ojs/index.php/fm/article/view/981/902
[9] http://www.ifla.org/digital-libraries/manifesto

sustainability are key to the vision of DLs able to communicate with each other. Digital libraries that conform to commonly agreed open standards and protocols improve world-wide knowledge dissemination and access." Metadata are interoperable when they have the same structure, or are structured in such a way that they can be aligned and reused for external or third party projects; however metadata in different format can be interoperable as well, as long as they can be aligned to a common metadata standard.

The Europeana Data Model (EDM)[10] represents a concrete answer to the necessity of interoperability among DLs. This data model has been indeed developed from the collaboration of data experts from a number of cultural heritage institutions: the strength of this data model is represented by its capacity to be extended for different spheres of cultural heritage institutions, such as museums, libraries and archives, and to capture important information of each of these fields-related data model. In this sense the Europeana Data Model has represented a big step forward in the dialogue and the connection among different data models, by allowing the aggregation and the comparison of data, which was just unthinkable a few years ago. The benefit that such an interoperable data model can bring to the field of DLs is considerable in that it facilitates the aggregation of data, not only on the Europeana portal itself, but by third party projects as well, such as CENDARI.

The aforementioned documents and research projects focus on different but complementing aspects of the sustainability of DLs, namely: Aspects related to Project Management (such as the identification of the value proposition and its users); Documentation of the creation of Metadata for Digital Objects, Interoperability of data formats and standardized data exchange protocols (such as oai-ore, API). Although they were useful in the conceptualisation of sustainability for CENDARI, we felt that much was missing from them. For this reason, we also investigated the sustainability plans of other European projects, in the hope of finding something more holistic - a review of these findings follow.

2.2. Archives Portal Europe

The project APEnet (Archives Portal Europe Network) ran from 2009 to 2012 and continued until 2015 with the name APEx. Apart from building a strong network of European archives, Archives Portal Europe has also created a portal where researchers can freely access the electronic versions of the finding-aids held by the participating archives (https://www.archivesportaleurope.net/). Moreover, Archives Portal Europe has a strong relationship with the Europeana Foundation and publishes its content on the Europeana portal, for easy access to its users as well as the general public.

As the project was approaching the end of the first funding period, it was already clear that it was necessary to think about the long term sustainability of the project's outputs. In 2012

---

[10] http://pro.europeana.eu/page/edm-documentation

APEnet published a sustainability report[11], which focussed both on the large network of organisations that collaborated with APEnet (and therefore valid candidates to sustain its project's outputs and community) - both of researchers as well as of archivists. The identified organizations were EBNA (the European Board of National Archivists), EAG (the European Archives group), ICA (the International Council of Archives), EURBICA APEnet Liaison Group (EURBICA being the European Branch of the ICA).

Moreover, this document focused on the risk assessment and improvements that the project had to take into consideration in order to guarantee a sustainable future for the project. Among new developments to be planned, the following were investigated:

- *Metadata Standard Implementation* (importance of XML in the archival standardization)
- *Ingestion of content by new content providers* (possibility to ingest content not only from Nat Archives but from local archives as well)
- *Easier Ingestion Procedure*: develop more automated ingestion and knowledge transfer
- Need of a viable *support financial plan for the decades to come*, in order to sustain and further develop what has been done so far
- *Dissemination of Knowledge and know-how*: against the risk of the project becoming stagnant and monolithic over past achievements
- *Usability as a main factor for sustainability*: APE envisions to develop a platform for the exchange of collections and feedback between researchers and archivists.
- *Future chances via Europeana*: it is possible that the APE portal will act as a dark aggregator to Europeana, collecting archival material, the same that The European Library is currently doing for libraries.

As the second funding period for the APEx project was about to finish in 2015, its project leaders, as well as the community of archives that formed the network, realized how important it was to sustain the objectives that the network had achieved; so, the idea of a continuation of the project, under different clothes, started to take shape. The Archives Portal Europe Foundation kicked off in September 2015 and among its main objectives, the future tasks of the foundation will verge around the following points:

1. International cooperation and collaboration
2. Sustaining the digital archival infrastructure (methods and experiences of related projects with regard to governing models, funding and financing methods such as public-private partnerships and membership models etc. as well as the question of integrating the wishes and demands of content providers as well as users, into further developments)
3. Assessment of use cases of digital service: infrastructures and the re-use of data (by other cultural heritage portals or specific user groups)

---

[11]http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/APEnet/Deliverables/D5.4%20Sustainability_Report.pdf

4. Impact (provision and use of archival resources, data protection, legal issues and the potential of open data)
5. Possibilities and perspectives of digitising archival content for access and possible re-use

At the time of this report, the APE Foundation is still in the process of establishing itself and its future roadmap, though the authors have been informed that the development of an open API for the reuse of this data, a highly welcome development, is in the planning phase.

2.3. EHRI, ARIADNE, DASISH

Although most closely related to CENDARI, neither the European Holocaust Research Infrastructure (EHRI) nor the Advanced Research Infrastructure for Archaeological Dataset Networking in Europe (ARIADNE) network were able to provide much formal input for the CENDARI project, as they are both ongoing projects at the time of CENDARI's close. Hopefully they will be able to benefit from CENDARI instead.

The Data Service Infrastructure for the Social Sciences and Humanities (DASISH) did close well before CENDARI, but did not leave a strong footprint in terms of sustainability documentation, perhaps due to the networked nature of its goals.

2.4. DM2E

The project DM2E[12] ran from February 2012 to January 2015 and it was coordinated by the Humboldt University in Berlin. The goal of the project was to build the tools and the communities to enable humanities researchers to work with Manuscripts in the Linked Open Web. The project has worked on different levels of knowledge creation and transfer in order to enable scholars - as well as third party projects such Europeana - to reuse the content they have gathered and to enrich it with annotation tools such as Pundit.

DM2E has created a number of tangible and intangible assets, namely:
1. aggregation of content and data interoperability
2. adaptation of existing tools to the needs of the DM2E targeted users
3. creation of a community of researchers

In 2014 the project released a deliverable on the sustainability of the project, entitled "Result Transfer Plan"[13].

As the connection with Europeana was established from the very beginning of the project - both in relation to the dissemination of the content aggregated and to the data mapping and interoperability - this document is strongly focussed on the sustainability of the data and the technical infrastructure developed. Europeana committed to sustain these two assets,

---

[12] http://dm2e.eu/

[13] http://dm2e.eu/files/D5.10_1.0_Results_Transfer_Plan_140730.pdf

however a tension between the EDM-RDF developed by DM2E and the capacity to ingest and sustain this format by Europeana (Europeana's infrastructure is still bound to a XML version of EDM) emerged, and it would require the adoption of a new technical infrastructure by Europeana to overcome this problem.

In the meanwhile, DM2E data will need to be preserved and sustained, however. In this context, DM2E will provide a complete archive of the data on the server hosted by the Computer and Media Service Repository of Humboldt-Universität, where the data can be exported via the API. In addition, the project will send to every data provider an archive of the data in EDM-RDF, creating therefore a sort of distributed archive while enabling the single partners to reuse their original data in a linked open data form.

As for the annotations produced by the researchers within the platform Pundit, they will also be archived in the servers of the Humboldt University, which will make them accessible via export through the API.

The Ontologies developed within the project (a sub-set of the EDM data model) will be made public through the DM2E's own ontology publishing service, an installation of the publishing system *Neologism*. Moreover they will be accessible via external ontology libraries (as http://lov.okfn.org/, http://vocab.deri.ie/, http://metadataregistry.org/, http://datahub.io/, https://onki.fi/).

DM2E developed a set of tools in order to facilitate the ingestion of data from its content providers, in order to transform the original data into RDF-EDM and to enable researchers to annotate resources via the DM2E interface. All the tools will be available via GitHub and released under a double licence (A-GPL 3 + commercial licence as described in http://thepund.it/license). Even though a complex annotation tool like Pundit cannot be integrated into the Europeana infrastructure, nonetheless it could be adopted by some of the Europeana channels (which are mini-projects and portals, within the Europeana Sound project).

As for the documentation and official deliverables produced by DM2E during its lifespan, they will be hosted on the Europeana Pro website (http://pro.europeana.eu/), which ensures long-term and easy access to all the Europeana partners.

Overall, the documentation prepared for the sustainability of the DM2E project reflects a comprehensive approach, as well as the understanding that not only data and software need to be preserved, but documentation and know-how as well. DM2E also demonstrates a pragmatic approach, circumnavigating the challenges presented by the interoperability of data and infrastructure, by proposing valuable alternative host institutions for the archiving - and where possible, dissemination - of its data and software.

3. Actions for Sustainability

3.1 The CENDARI Approach to Sustainability

3.1.1 What Sustainability means for CENDARI

As mentioned above, the Digital Humanities landscape is littered with projects that were not sustained by or for their intended user community. Reuse of tools and processes is also not a part of the culture. The resulting inefficiency slows the cycle of development in digital humanities.

Final results of an investment on the scale of the CENDARI project should not be allowed to suffer this fate, however. Resisting the pull and push of the forces that prevent sustained reuse will not be easy, if for no other reason than the simple reality that the end of a funding stream indicates the end of staff dedicated to working on and for a resource. It may be possible for a digital project to thrive without investment, but not without commitment. In fact, a truly successful community tool or piece of software is often looked upon as a liability, a long term responsibility that must be maintained but no longer generates significant research advancement. The project partners that build something new, such as the university research groups that have driven CENDARI forward, are seldom able to shift their focus and mission enough to comfortably absorb this other mode of operation, and research funders are seldom interested in sustaining services, rather than projects.

There were two fundamental concepts that needed to be defined before this Sustainability Plan could be developed, and which form its most basic assumptions: **sustainability** and **assets**.

Sustainability, in this context, has no easy definition. It is certainly not as simple as keeping a server running somewhere, or even as clear-cut as the preservation of a curated data set. In fact, merely 'putting something somewhere' is probably the opposite of sustainability, for while it may ensure the potential of digital availability, it will certainly also ease the path toward invisibility and ossification. What it means to sustain a resource is unique not only to that resource, but to its user communities and to what they value about it. The transferal of ownership from the project team to these communities is a moment at the end, but also at the beginning, of a long process of engagement and co-creation. DARIAH's strong focus on communities of practice and working groups should serve this focus well, but the challenge is to know: first, what the valuable assets of a project like CENDARI are; second, who would benefit from them; and third how and when they can be effectively communicated to potential successors. For CENDARI's purposes, therefore, we defined sustainability as comprised of the visibility, activity and availability for reuse of the project assets.

CENDARI's 'Assets,' are also not simple to define. Although the culmination of the CENDARI project's work exists in the form of a digital platform, this platform will not necessarily make all of the project assets accessible to future users. In fact, the potentially useful assets of a project like CENDARI come in a full range of forms, from relatively tangible (data, archival research guides, publications) to highly intangible (processes, best practice, know how, communities). Deciding what is of value and to whom may be the first step, but deciding how to make it concrete may be a further layer of required activity.

As a part of its sustainability planning process, the CENDARI team developed a list of candidate asset classes for sustaining. These classes were as follows:

- Technical Infrastructure: Portal, Services and Tools
- Research Data: Unique and Aggregated
- Internal Publications and Knowledge: Archival Research Guides (ARGs), Toolkits, and Knowhow, Management Data and Assets.
- Communities: People, Networks and Relationships

Each of these categories and the specific mapping exercises and sustainability actions related to them will be discussed in the following sections of this document.

3.1.2 How this plan was developed

From its inception, CENDARI's medium- to long-term sustainability has been linked to the DARIAH ERIC: indeed, the project Description of Work states clearly that: "The research aspects of this bid will be led by strong DARIAH partner institutions so as to integrate seamlessly with DARIAH's own development. This aspect of the CENDARI project will address the need to create a robust facility for the long term that is integrated with the major infrastructural developments for the sector, in particular DARIAH, but also many others (see Section 3.1.2). By bringing together the concerns of the scholars forming these strong transnational networks with the work of DARIAH, CENDARI will contribute to the alignment of DARIAH's developments with scholarly needs."[14]

---

[14] See also, from the CENDARI DOW: "The work required to best model for the sustainability of the CENDARI network and electronic presence is strategic, designed to ensure that CENDARI continues to exist as an accessible e-infrastructure; to grow and develop over time and in line with user needs; to be able to ingest content as it becomes available in digital form; to support the development of capacity for digital integration among the smaller archives of Europe; and to be responsibly and efficiently managed and delivered. It will also look toward the extensibility of the platform to include research areas beyond the test case, becoming an extensible resource and example of best practice in the vertical and horizontal integration of e-resources. The solutions to these possible barriers to longevity will be addressed from an early point in the project via a number of avenues, including: re-usage of existing DARIAH services, and the re-integration of aspects of the CENDARI infrastructure into DARIAH; efficiencies with other cognate data infrastructures, such as the EHRI; negotiation with commercial publishers and other entities as to any possible mutually beneficial model for cost sharing that might be developed. These activities will be guided by a CENDARI business plan, which will be created and agreed after the midpoint of the project development."

But DARIAH was then, and to a certain extent still is, an entity in its foundational stages, and the pathways for ensuring the long-term availability of relevant results from projects like CENDARI are not yet clear. For this reason, CENDARI took on the task of designing modalities and processes by which an independent, albeit linked, project could link in with DARIAH over the course of its development to co-develop a model for embedding sustainability planning with and for DARIAH reuse into a project development.

The first step in this process came about naturally in the second year of the project (2013). It was driven by recurrent queries from data providers we were approaching regarding whether their data would continue to be available after January 2016, when the CENDARI project would end. In order to ensure provider comfort with CENDARI's commitments to protecting their rights and reputations, we drafted a brief Memorandum of Understanding between CENDARI and DARIAH, outlining their complimentary roles and DARIAH's commitment to maintaining what CENDARI would build.

This ad hoc approach to identifying issues relevant for sustainability and creating measures to address was followed upon with a formal, 18 month-long sustainability planning exercise, conducted from July 2014 through January 2016. This exercise, which built upon the previous work to date, began with the preparation and planning of a stakeholder meeting in January 2015. The participants in this workshop were drawn from CENDARI and DARIAH, but also from cognate projects, such as EHRI and DM2E. Resulting from this workshop, a set of principles and processes for mapping and sustaining user value from a project for the medium and long terms were discussed, and a number of key actions were identified as key enablers for the planning, development and conclusion of digital projects, in particular for projects affiliated to DARIAH. Although both the generic process (which will be released as a sustainability toolkit at the end of the project) and the specific actions implemented by the project match on some level the specifics of the CENDARI development, they also reflect the reality, identified by Joris Van Zundert, of the "fluidity" of research infrastructure, caught up in both the digital information lifecycle and the creation of knowledge by end users, as well as the software components.[15]

The discussion and conclusions of the CENDARI Sustainability Workshop resulted in a roadmap for CENDARI, as well as an agreed set of generic actions projects should build into their proposals to promote sustainability (see Section 4 below for these broader recommendations). As far as the CENDARI process was concerned, the following actions were instigated following the workshop, and became the core drivers of the Plan you are reading. The agreed actions were:

1.   CENDARI must engage with DARIAH-DE to lobby for a continuation of the digital presence of the CENDARI virtual work space beyond January 2016.

---

[15] Van Zundert, J., If you build it, will we come? Large scale digital Infrastructures as a dead end for digital humanities. Historical Social Research Journal ; 2012, Vol. 37 Issue 3.

2. CENDARI must map the landscape of its tacit knowledge, and create a roadmap for publishing on the experience gained by the project that might be of use to the community and otherwise lost. To achieve this action, each work package should submit an assessment of activities or experience that have not already or are not already planned to be a part of a formal, publicly available deliverable. In doing so, project participants are encouraged to think of the utility of their experiential knowledge for future researchers, DARIAH affiliate projects, institutions or others. This task was in the first instance assigned to WP leaders to consult on and respond to, and resulted in a number of specific outputs, such as the CENDARI White Book of Archives.

3. A more detailed plan for the future sustainability and reuse of CENDARI's data (collected and created) must be developed in association with DARIAH. To achieve this action, we must first return to the early work of WP6 on data types and flows in the researcher's system of work, scoping out the sources and streams of data populating the CENDARI system. This list should be revisited and revised if necessary, and the updated list of data types should then be assessed for its size, uniqueness, openness, fluidity, personal or sensitive nature and any other characteristics that might affect the necessity or quality of an ongoing commitment to its sustainability. As a third step, possible appropriate measures to ensure long term sustainability for this data should be devised for different data streams (and, where appropriate, for intersections or relevant subsets of these streams), bearing in mind that sustaining the full environment through DARIAH DE is a separate conversation but that there may be other possibilities available through DARIAH, such as the VCC3 Common. This may include long-term deposit of some sort outside of the full CENDARI environment with DARIAH, handover in some format over to institutions, individuals or aggregators, placement on a public repository, closure or removal or other measure. Additional consideration under this item should be given to the aspects of data maintenance that require human oversight, such as curation and oversight of new archival descriptions or other data that may be added to the system after project end, the recruitment and maintenance of data for new users, the maintenance of ARGs as potentially fluid scholarly objects, etc. A small task force, composed of representatives of WPs 2, 5 and 7 (and possibly 6 and 8) was convened to develop an initial position on this issue.

4. CENDARI should develop a plan and a toolkit for enabling future user communities, including but not limited to Horizon 2020 Starting Communities to adapt and develop the tools and services developed by CENDARI.

These actions converged with our fundamental understanding of how we wanted to define sustainability and how we wanted to conceptualise the CENDARI contribution to knowledge to define a set of conversations, agreements and actions delivered over the course of 2015. The results of that process are described in the following sections of this document. As such, this document evidences a concrete response to the hypothesis that a successful approach to sustainability for Research Infrastructures needs to be comprehensive; an

approach that doesn't just consider data or technology, community, communications or processes, but in fact all of them simultaneously.

3.2. Technical Infrastructure: Portal, Services and Tools

The current physical home of the CENDARI servers is at the University of Göttingen. In order to ensure the basis for continued use of the portal as a whole beyond January 2016, DARIAH-DE has agreed, as a part of their contribution to DARIAH-EU, to keep the services running for three years. In some ways, this may seem a very short time but in fact, it is a very useful timeframe: long enough to prove the utility of the resource, but not so long that is becomes a question of trying to migrate tools and services wholesale to new environments and underlying platform versions that could cause growing instability in ageing.

Continuing availability, even at this time step, is not the same as sustainability, however: unless both the user base and the content in the system continue to grow, the system might as well be discontinued. Within 3 years, the software will either have proven its value or will have done so to the point of requiring a (funded) technical overhaul by new or existing communities to continue to be of sustained use. Within this time, new user communities will be recruited to take on the continued development of the system, potentially via a modularized installation ("CENDARI in a box"). Responsibility for the growth and development of CENDARI in this time will be within the remit of the CENDARI SUSTAIN working group described below.

But the full portal or virtual machine may not be the most interesting format in which CENDARI software can be reused. Having adopted a service-oriented architecture, CENDARI's tools and components are mostly modular, and most of these modules have an existence outside of CENDARI (sometimes being available as a stand-alone web service). The tools therefore will usually have a sustainable pathway outside of the portal already in place.

That said, however, connecting tools with potential user bases is a constant challenge. The software community practice of using GitHub to share software is a good one which has been adopted by CENDARI and should be adopted widely, but this only will impact the developer, rather than the potential end user. Tool registries generally become outdated before they can establish themselves as a recognized resource, but DARIAH may be in a position as an ERIC to develop a better platform or modality for facilitating this kind of awareness, e.g. through their Marketplace and their close connection to the many Working Groups. The CENDARI SUSTAIN working group (described in section 3.6 below) will maintain a watching brief over these possibilities as well.

3.2.1 What are the CENDARI Services?

The CENDARI research infrastructure is a complex system of applications. The infrastructure comprises four main layers:

1. ***Files storage and servers***
2. ***The Data Access Layers***: includes search engines such as Elastic Search and SOLR; includes the RDF triplestore Virtuoso
3. ***The Application Layer***: includes the open source data portal platform CKAN and the CENDARI Data API, that constitute the communication layer between the data stores and the presentation layer
4. ***The Presentation layer***: includes the presentation of different functions of the CENDARI data space, such as: Ontology Viewer (Pineapple); Notes VRE; CKAN User Interface; ATOM User Interface. The presentation layer includes external applications that connect to the existing infrastructure via the Notes VRE: TRAME (the harvester and scarper for the medieval data) and the Named Entity Recognition (NERD).

The image below shows the latest infrastructure architecture, including their relations and dependencies.

CENDARI Infrastructure Layer Model v 1.0.1

February 2016
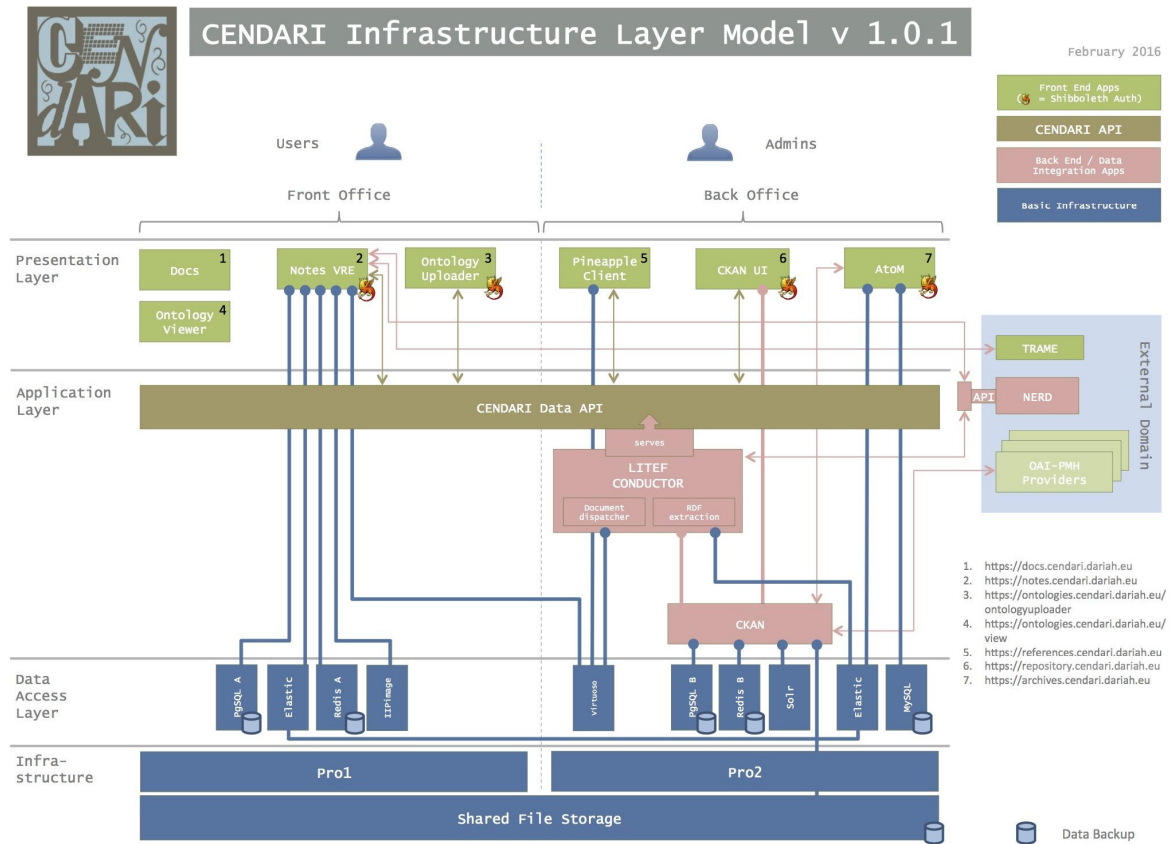
3.2.2 Sustainability of Tools and Services

An agreement for the sustainability of the tools has been signed between the Project Coordinator Trinity College Dublin and DARIAH-DE, represented by the CENDARI partner UGOE (University of Göttingen). This agreement will be in action until February 2019, three years after the end of CENDARI.

The following summary gives details of these components and lists the action points for sustainability of tools agreed between TCD and UGOE.

A.   Note Taking Environment (NTE)

The CENDARI Note-Taking-Environment (NTE) is the central research environment for historians developed by CENDARI. It allows historians to collaboratively collect notes and documents within projects with fine-grained access control. The NTE provides access to the archival directory and the repository.

TECHNOLOGY
The Main Application is written in Python using the Django Framework and based on the older version of the open source solution *Editor's Notes*. It uses postgres, the Data API, elasticsearch and Virtuoso at the backend. The relation to the DataAPI (thus CENDARI Repository) is very basic; it only checks the permissions of the repository data for displaying them in the faceted search. Only NTE project definitions are saved in the Repository as Dataspace. NTE Projects data (notes, images, documents) are not saved into the CENDARI repository. Further requirements are node.js, fabric, redis, flup, vips, iipserv.

STATE AS OF JANUARY 2016
Elementary documentation[16] on how to install a development environment exist; it is not possible to create installable releases. In particular, no instructions for setting up a production environment exist and the current production server is compiling from source using the development libraries on every update.

As there is no sufficient source code documentation at the moment, it is not clear, whether DARIAH will be able to cover minor security bugfixing for a long term provisioning of the NTE.

DARIAH-DE ROLE
Will be kept running as it is, as long as it works and no security implications arise. Should the component fail within 3 years, registered users will be advised to export their data in RDF format to provide continued data access outside of the NTE. This component would be one of the greatest concerns due to its complexity, but also because the partner that developed will not be continuing association with the CENDARI SUSTAIN Working Group (as will MISANU and UGOE). Through the CENDARI SUSTAIN Working Group, MISANU will evaluate the possibilities to export the data into the CENDARI Repository and provide redirect to its User interface in case the NTE will be discontinued.

B. Data API and Litef Indexing Framework

TECHNOLOGY
Developed by MISANU on top of CKAN in Scala, the DATA API itself is not publicly accessible.

STATE AS OF JANUARY 2016
Documentation on build and installation are provided and the API itself is fully documented[17]. Source code documentation exists according to official ScalaDoc style.

DARIAH-DE ROLE
The API will be running on the CENDARI servers for as long as possible. Maintenance, in particular if CKAN upgrades cause incompatibilities, cannot be provided.

---

[16] https://github.com/CENDARI/documentation

[17] https://github.com/CENDARI/documentation

Should MISANU have a prolonged interest, implementation against the DARIAH-DE repository or through the DARIAH ERIC could be investigated collaboratively.

C. CKAN

Open-Source Software Repository Solution developed by Open Knowledge Foundation UK.

TECHNOLOGY
Implemented in Python, requires SOLR instance to work for indexing of data for the web interface, which is not used by CENDARI.

STATE AS OF JANUARY 2016
Standard installation of an older CKAN release with a number of third-party plugins. Maintenance is performed mostly by MISANU who is also responsible for data ingest.

DARIAH-DE ROLE
The installation will be hosted as long as possible. Should security concerns arise, the interface may be taken offline, severely impacting the overall CENDARI user experience. Due to the lack of standard interfaces in CKAN, integration of the CENDARI repository within DARIAH's planned Data Federation Architecture will not be possible. This situation may change over time, however and will be monitored.

D. Pineapple

Browser for the Knowledge base, i.e. the contents of the Virtuoso triple store.

TECHNOLOGY
Implemented in PHP, directly accesses the Virtuoso triple store.

STATE AS OF JANUARY 2016
The tool is fully documented and deployed[18]. Pineapple is publicly accessible under https://resources.cendari.dariah.eu/ .

Pineapple communicates to CENDARI Virtuoso triple store (Resources, WW1 Ontologies) and in addition to the SPARQL end-point provided by Medievalists (TRAME Knowledge Base) for Medieval Ontologies.

DARIAH-DE ROLE
The tool can be hosted under the current conditions.

---

[18] https://github.com/CENDARI/documentation

E. Access To Memory (ATOM)
Open-Source software solution "Access to Memory" for collecting and describing archival institutions and holdings.

TECHNOLOGY
Implemented in PHP framework Symfony.

CENDARI MODIFICATIONS
- Modifications were made to the AtoM core to remove certain unwanted features.

CENDARI ADDITIONS
- Shibboleth authentication plugin
- CKAN sync script

STATE AS OF NOVEMBER 2015
Minor modifications have been made to the official release affecting only the presentation.

DARIAH-DE ROLE
Will be maintained and possibly extended to include wider subject area and/or new user communities. Software will be upgraded when new versions are released. The modifications will be ported if possible.
The Shibboleth plugin will be maintained as long it is supported by applications, the sync scripts will be kept as long as they work.

E. Ontology Upload Tool
The ontology uploader is developed by KCL, the prototype will be publicly available at the end of January 2016.

DARIAH-DE ROLE

As the ontology uploader is hosted by the Dariah-DE servers, Dariah commits to maintaining this service. [19]

HOSTED EXTERNALLY
In addition to the services described above, the following services are hosted directly by the respective project partner and provided to CENDARI as external services.

---

[19]

F. NERD

The Named Entity Recognition and Disambiguation service is hosted and maintained by INRIA and provided to CENDARI and other parties via a REST interface. DARIAH-DE expects that this will continue, and the partner is committed to the further development of the service.

NERD is used by the CENDARI Data Integration Toolkit (Litef) for automatic population of the Cendari Knowledge Base. Should this service stop working, this population will stop, but will not affect further functioning of the data API and Litef indexing Framework. A multilingual NERD has also been developed, but not made stable enough during the project development period to be rolled out as a full service (though this may happen in the future).

G. TRAME

The TRAME search interface is hosted and maintained by FEF/SISMEL and independent from the servers in Göttingen. The NTE provides access to the service through an API that has been developed within CENDARI and published on GitHub along with its documentation. TRAME is a sort of brand or umbrella that encompasses a number of different tools, services and datasets for the study of medieval culture:

- Resources discoverability (TRAME)
- Advanced Data Search & Browse (MeSo Search)
- Ontology Navigator (MeSo Browser)
- Semantic data aggregation (MeSo Knowledge Base)
- SPARQL endpoint

The development of the tools will be undertaken by the underlying scholarly community in the context of the Medievalist's Sources DARIAH-ERIC working group (already approved). A subset of the tools under the TRAME umbrella is provided as in-kind contribution to the DARIAH-IT national branch of DARIAH-ERIC.

H. CENDARI Website (Drupal)

The CENDARI website is hosted and administrated by an external Drupal hosting provider, with whom the project has contracted for three further years hosting at time of project close. This part of the infrastructure will be overseen and maintained by Trinity College Dublin as a part of its commitment to the CENDARI SUSTAIN Working Group.

In practical terms, this action translates for the CENDARI SUSTAIN Working Group in liaising with the Drupal hosting provider and in making sure that the costs related to the administration of the server and the services provided are covered.
The Drupal hosting provider will take care of:
- maintenance of the website, including software and plugins updates
- check spam comments
- check the stats, in order to monitor the traffic to the CENDARI website

Key Considerations for ensuring Sustainability of a Technical Infrastructure:

- **Identify a partner (or a group of partners) responsible for the maintenance of basic services after the end of the project**
- **Have a realistic expectation for how long software will remain useful if not under active development**
- **Design the infrastructure in a way that the single technologies can be reused and implemented by other research infrastructures**
- **Technical documentation of the tools and their integration should be openly available**

3.2.3 Tools assessment by KNAW-DANS

As a further element of the project's commitment to creating a best practice model for sustainable development, CENDARI's tools and software will be subjected to a pilot sustainability test within the 'Humanities at Scale' (HaS) project, led by the DARIAH ERIC.

As part of the HaS Project[20] coordinated by Dariah-EU, the partner DANS-KNAW[21] will perform a series of assessments based on capability maturity modelling and reuse readiness levels, for the Deliverable 5.3 due to month 24 "Service Integration Demonstrator. Showcase of Dariah assessed services".

The aim of this appraisal is to assess the appropriateness and the reuse readiness levels of the Dariah-EU in-kind contributions, including tools as well as services (i.e. summer schools or workshops). This exercise can be performed both at the beginning or at the end of a project: in the first case it provides suggestions and recommendations, while in the latter case, it assesses whether the service provided is mature enough - according to the user requirements developed both in the project planning and by the capability maturity level assigned by the infrastructure itself.

CENDARI has been selected as a case study for the maturity assessment of the Dariah infrastructure for a number of reasons: first of all, some of the tools developed within CENDARI will be maintained as contributions to DARIAH-EU; secondly, the set of applications developed within the project presents a variety of tools with different functions and technologies, yet they all contribute to build a unique and coherent research infrastructure for the humanities.

---

[20] https://dariah.eu/activities/humanities-at-scale.html

[21] http://www.dans.knaw.nl/en

This testing process will not only add to our understanding of what it means to maintain software based infrastructures in general, and CENDARI in specific, it will also help to concretise processes for future projects to use in developing their own sustainability programmes.

3.3. Research Data: Unique and Aggregated

The CENDARI project comprises a wealth of data from different sources and with different requirements and expectations for sustainability. These sources include both unique (CENDARI-created collection and institution descriptions) and deposited data (e.g. from contributing institutions), some of which may have a clear web presence (e.g. from Europeana) and some not, most of which will be public, but some not (such as users' working annotations).

The CENDARI data agreement and license were developed with DARIAH as a co-signatory, via a Memorandum of Understanding between the projects, so in many ways DARIAH has already agreed to sustain this data. But again, we had to ask ourselves what this commitment should mean. As a baseline, the data will be sustained by DARIAH DE as a part of the overall portal, but, from a sustainability point of view, it seemed reasonable that some data exist in more than one location and format, and that unique data might have different requirements to aggregated. Therefore both independent and shared approaches to the preservation and curation of data are taken and described in this section of the sustainability plan.

3.3.1 "Data Soup"

CENDARI has federated content on the model of a 'data soup,' including a wide variety of formats and levels of description, hierarchical interrelation and granularity. Where we have had control over the descriptions, we have tried to use recognised, interoperable standard to describe data, while also preserving the greatest level of data richness possible.

Collections descriptions encoded manually by the collaborators within CENDARI correspond with EAD (Encoded Archival Description), the archival standard used in almost every European and also non-European country, which is approved by the International Council of Archives. However, the descriptions available from data providers may be based on other standards not yet enclosed within the EAD. For different media types, different standards are applied (e.g. MODS – Metadata Object Description Schema, METS – Metadata Encoding and Transmission Standard, etc.). In addition, export functionality (through the "ATOM2CKAN" synchronization) has been provided to export the Archival Institution descriptions in accordance with the EAG standard.

Instead of applying a classical integration approach and defining a common description format, the CENDARI repository allows for coexistence of such heterogeneous content. The

classical approach almost always suffers from loss of information which occurs during translation from the original into a common format acceptable by the repository. By contrast, the method adopted in CENDARI is based on ensuring a set of common functions over diverse formats and allowing for an evolutionary approach in providing more specific and semantically rich services. The need to perform transformations over collections descriptions encoded in various formats in order to achieve a certain level of semantic integration is not avoided. However, the upfront efforts are lower and enable for incremental integration during time. The ability of the system to manage a variety of data types and standards led to the coinage of the term 'data soup', a hearty mixture of objects, descriptions, and sources and XML formats, database exports, PDF files and RDF-formatted data respectively.

As of December 2015, the CENDARI repository contains 829,087 descriptions represented in a number of data formats, including EAD, MODS, METS, DC, EAG, LOD and EDM. These data are stored in a repository called CKAN[22], an open source data portal platform, which currently hosts both manually encoded (collection and institution descriptions) and data harvested from Cultural Heritage Institutions. The table below shows the figures related to the data acquired by CENDARI and that will be accessible via the faceted search of the CENDARI Note Taking Environment (or Elastic search API, as an underlying component).

Format	Total per Format
EAD	48,503
MODS	44,347
METS	2,627
DC	44,347
EAG	1,315
LOD	198,883
EDM	486,065
PDF	3,000
Total	**829,087**

---

[22] http://ckan.org/

3.3.2. Sustaining Data Harvested from Cultural Heritage Institutions

The aggregation of data from pan-European aggregation projects and small, local archives represented both a challenge and an opportunity for a project like CENDARI. A challenge, because dealing with institutions that differ in their size, the scale of their digitized records and catalogues, and the technical solutions they have implemented has required great flexibility from our research team. On the other hand the variety of cultural heritage institutions that we have been in contact with represents one of the biggest assets of the project, which aims to guarantee a wide representation of institutions, according to their geographic coverage, type of collections preserved and data describing their analogue collections.

The Cultural Heritage Institutions which the CENDARI project has been in contact with range in size from local archives to pan-European aggregation projects. As different kinds of institutions have different kinds of internal organisation and financial support, their data formats, organization and accessibility vary deeply from case to case. Small, local archives usually lack staff and financial resources to invest in metadata standardisation and data storage. They often preserve and make accessible archival descriptions through spreadsheets and often don't have an online presence: for this reason, CENDARI has called these kinds of archives, "hidden archives".

In our experience, the situation is quite different for National and International Archives. Although sometimes operating under serious resource constraints, such institutions usually have a cataloguing and encoding department, as well as a dedicated IT person or office that takes care of the technical infrastructure in place to preserve their data. In spite of being quite advanced with their technical infrastructure, however, national and international archives often lack both the technical means and the policy framework to share their data with other institutions and projects. Having a larger scale and a national mandate does not necessarily mean that technical choices have been made with the same priorities as would support cooperation with an infrastructure project, and having more people involved in delivering the digital presence often means that no one person is informed or empowered enough to grant access.

Pan-European Aggregators, on the other hand, focus their activities on the aggregation and the re-use of data from cultural heritage institutions; therefore, they are very advanced in terms of data-sharing protocols and applications, such as APIs and OAI-PMH. They have been a great resource for CENDARI.

The table below shows the figures related to the data harvested from Cultural Heritage Institutions as well as aggregation projects.

Data Format	Institution/ Project	Total per Dataspace and Format
EAD	American Jewish Joint Distribution Committee (JDC Archives)	2
EAD	Archives Hub	1,189
EAD	Bundesarchiv	25,384
EDM	Europeana	287,182
DC	Kalliope	44,347
MODS	Kalliope	44,347
EAD	Kalliope	988
EDM	The European Library & Europeana - Newspapers Project	198,883
LOD	The European Library & Europeana - Newspapers Project	198,883
EAD	Ajalooarhiiv	5,253
EAD	Riigiarhiiv	4,895
EAD	Tallinna Linnarahiiv	1,207
PDF	Ottoman Archives	over 3,000 descriptions contained in 211 PDF files
EAD	Czech Department of Archives Administration	2,050
EAD	Landesarchiv Nordrhein-Westfalen	970
EAD	Landesarchiv Baden-Württemberg	36
EAD	Bayerisches Hauptstaatsarchiv	36
METS	University Library Heidelberg	2,627

Data Format	Institution/ Project	Total per Dataspace and Format
EAD	International Institute of Social History, Amsterdam	4,547

In all of these cases, CENDARI has been very clear that our remit is to facilitate the kind of federated search required to support transnational research, not to preserve unique data for the long term. As such, while we offer to share any transformations or results based on institutional data back with the archival 'owners' of that data (of interest particularly in the case of those 'hidden' archives) we did not feel that significant investment was required in providing alternate access routes or formats for it. In addition, we do not believe that the collections (selected to represent our two case study areas, medieval culture and the First World War) are likely to see significant new additions over time. Therefore we did not feel that we needed to make regular updates of the data into a feature of our Sustainability Plan. In particular for the modern case, it would have been impossible to create a fully comprehensive resource including all relevant content anyway, for which reason we felt that tasking the Working Group to focus on any new collections that might become available, rather than regular update what will largely remain static data, was the better compromise.

3.3.3. Sustaining Manually Created Data

As introduced in the previous section, the so-called "data-soup" approach represented the best approach to aggregate data from different institutions for the CENDARI project. CENDARI did not only aggregate data from local, national and international content providers, however: the CENDARI researchers also manually encoded information related to archival descriptions and archival institutions. This strategy was envisioned from the very beginning of the project in order to allow those small archives that lack a digital and/or online presence to be nonetheless visible to researchers looking for connections to support their research in the CENDARI infrastructure.

In order to encode the archival institutions and descriptions, the CENDARI project adapted the open source software "Access to Memory" or "ATOM" (https://archives.cendari.dariah.eu/). As of January 2016, the CENDARI instance of ATOM hosts 3,924 archival descriptions and 1,285 archival institutions encoded manually by the CENDARI researchers. Both collections and institutional descriptions are freely accessible and can be faceted according to the geographic locations, type of archive and subject. ATOM was chosen by the project after testing several options to act both as the tool that would support the creation of the archival institution and holding descriptions and as the repository for this information. As such the CENDARI instance of ATOM often represents the first point of access to the descriptions for modern and medieval historians.

These data comply with one of two standard formats, depending on what they describe: EAD or EAG. The collection descriptions have been encoded in the EAD format. EAD stands for Encoded Archival Description: AtoM uses the standardized format Encoded Archival Description (EAD) to present information about an archival unit, which can range from a huge collection to an individual item in an archival institution. The International Council on Archives (ICA) defines an International Standard Archival Description (General) (ISAD(G)) as "The creation of an accurate representation of a unit of description and its component parts, if any, by capturing, analysing, organizing and recording information that serves to identify, manage, locate and explain archival materials and the context and records systems which produced it."

EAD is used worldwide for data exchange between archival institutions and for Internet presentations of archival descriptive information; it is also the selected data format of international aggregation project such as Archives Portal Europe.

EAG has been created by the Spanish Ministry of Culture and it is used to encode information related to archive repositories (institutions). The standard is currently available in form of a Document Type Definition along with an EAG Tag Library (in Spanish). The correspondent ICA standard has been released in 2008 as International Standard for Describing Institutions with Archival Holdings (ISDIAH).

Every archival description in ATOM is linked to an archival institution entry. An archival institution has been defined as "an institution holding legal and physical custody of noncurrent documentary materials determined to have permanent or continuing value. Archives and manuscript repositories are archival institutions." This includes all institutions that manage, preserve and/or store historical source materials.

The figure below shows the number of Archival Descriptions (EAG) and Archival Collections (EAD) in CENDARI as of January 2016.

EAG	CENDARI Archival Descriptions	1,315
EAD	CENDARI Archival Descriptions	3,996

This data is accessible through the CENDARI ATOM application (https://archives.cendari.dariah.eu/) as well as through the CENDARI NTE search functionality. These data are made available as well in another place - in the CENDARI

Repository (CKAN) as EAD and EAG format as a separate Dataspace, which is synchronized daily. [23]

The idea of making this data available in another place and format was considered, but discarded as too difficult to do in a way that would be meaningful for any users. This alternative should arise over time, however, through CENDARI's involvement in DARIAH VCC3's data aggregation project, described in Section 3.3.5 below.

3.3.4 Sustaining User-created Data

Data created within the portal will be sustained as a part of the portal by DARIAH DE. Should major threats to data integrity ever appear in the system, users will be advised to export their data and hold a copy outside of the NTE also. The inherently small user base of this data (one person) means that any changes from the current portal configuration can be easily communicated with the interested parties.

3.3.5. Collaborating with DARIAH to Enhance Data Sustainability

Sustainability of data doesn't only imply its curation and existence in a repository, but it is deeply connected with its reuse as well. The research data that CENDARI has produced and aggregated in the last four years has a strong potential for reuse by external research communities, and in particular for modern and medieval historians. Users in fact will play a central role in promoting the resources after the end of the project: if CENDARI data will be queried and referenced within historical research, this will in turn attract interest by research institutions and funding bodies.

This will not occur through individual use alone, however. During its life-span, CENDARI has developed strong collaborative links with international networks in the field of the Digital Humanities as well as in the Arts and Humanities generally: these collaborations will serve to continue to engage a wide community of scholars and digital humanists to mitigate the risk that the data collected and enriched in the context of research projects, becomes obsolete and not usable anymore.

In particular, we hope that our close relationship with DARIAH will provide a second line of access and support for the CENDARI data. This cooperation has already been instigated through a workshop organized by the working group Dariah VCC3 and organized by DANS-

---

[23] In addition, through the Litef indexing Framework, and by use of CAO (Cendari Archival Ontology) these data are also available in an RDF formats. Via the Litef indexing framework all other formats which Litef supports are also available (RDF/XML, RDF/N3, JSON, Text – simple – extraction for NERDing)

KNAW in The Hague. The attendees were invited from projects like CENDARI, ARIADNE and EHRI as well from European research infrastructures and Digital Humanities Centres, such as CVCE and the Austrian DH Centre.

Among the "users' needs" collected among the participants to the Workshop, the following were identified:

- Content management system to collect good practices and documentation
- Shared Data model
- Services to provide content, such as web interfaces and API
- Policy (i.e. who can remove/ modify the data)
- Authorization/ authentication
- Multilingual User Interface

Among the data-registries that the group considered, **re3data**[24] was felt to be a strong possible candidate to fulfil this task. Re3data was launched in 2012 and is funded by the German Research Foundation. It is a registry that covers research data repositories from different academic disciplines. The main aim of re3data is to make accessible research data to researchers, funding bodies, publishers and scholarly institutions. Moreover re3data is recommended by the European Commission as part of the Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. Implementation of a DARIAH-led federation of archival data within the re3data platform is yet to go forward, however.

Although the development is still in the future at the time of the end of CENDARI, it is foreseen that data with shared utility across similar projects will be integrated into the work of DARIAH VCC 3, where it can be federated with similar records (such as the EAG descriptions of the EHRI project). Although this development will fall outside of the current time horizon for CENDARI, the facts that a) the data for CENDARI is 'guaranteed' durability through the DARIAH MOU, b) that the CENDARI data is being sustained alongside the technical components by DARIAH DE and c) that the CENDARI SUSTAIN Working Group will itself be founded under the DARIAH umbrella, all reinforce the potential for harnessing this possibility when it is ready to move forward.

Key Considerations for the Sustainability of the Research Data:

- **Use, where possible, accepted and widely shared standards**
- **Agree from an early point in the project who will sustain the data after project close**
- **Design a data ingestion cycle that is as efficient and simple as possible, capable of being rolled out as an easily managed service at project close**
- **Identify interested institutions or repositories that can provide support with**

---

[24] http://www.re3data.org/

the storage, dissemination and reuse of the data after the end of the funding period
- **Be clear about what you have collected your data for, and what it's value is (and for whom)**

3.3.6 Ontologies

A special category of research data created by the project has been its linked data and ontologies. For the project, two sets of ontologies have been created, for the WW1 and the medieval domain respectively. These domain ontologies are part of a larger Knowledge Organization Framework, designed to aid in the discovery process encompassing dispersed archives and the archival research guides and other documents in the CENDARI Note Taking Environment (NTE).

In terms of sustainability, it is necessary that the ontologies developed within the project are made accessible and reusable to other researchers or projects. This means that the ontologies needed to be published via a number of channels; the documentation and guidelines (on how to "read" and reuse them) is a crucial element as well for the aspect of reusability. To facilitate different kinds of users, these ontologies are currently deployed though the CENDARI triplestore *Virtuoso* in xml-rdf formats. In addition, they are fully documented in the project Deliverables 6.3 and 6.4, available at: http://www.cendari.eu/about-us/project-deliverables and available to download in xml-rdf form: https://repository.cendari.dariah.eu/dataset/ontology-element-sets.

The documentation can be seen as an essential tool for anyone who wants to reuse or extend the CENDARI set of ontologies: it explains indeed which set the CENDARI ontology builds upon - the contextual classes of the Europeana Data Model (agent, organization, place, concept, person, group, Timespan) and mapping examples. The documentation also shows examples of the existing WW1 related thesauri, which have been incorporated in the CENDARI ontologies, namely: WW1 Online Encyclopaedia, DBpedia Battles, Trenches to Triples. Moreover the documentation includes a section with guidelines for the creation of additional instances. It was important not to reinvent the wheel in terms of ontologies for these domains, as such a sloppy development would lead to only partial reusability, a proposition with very little value and much frustration in it for future users.

The instances of the Ontologies are available via the tool **Pineapple**[25], developed in the context of CENDARI: https://resources.cendari.dariah.eu/. At the moment the interface of Pineapple includes both the First World War and the Medieval resources, that have been

---

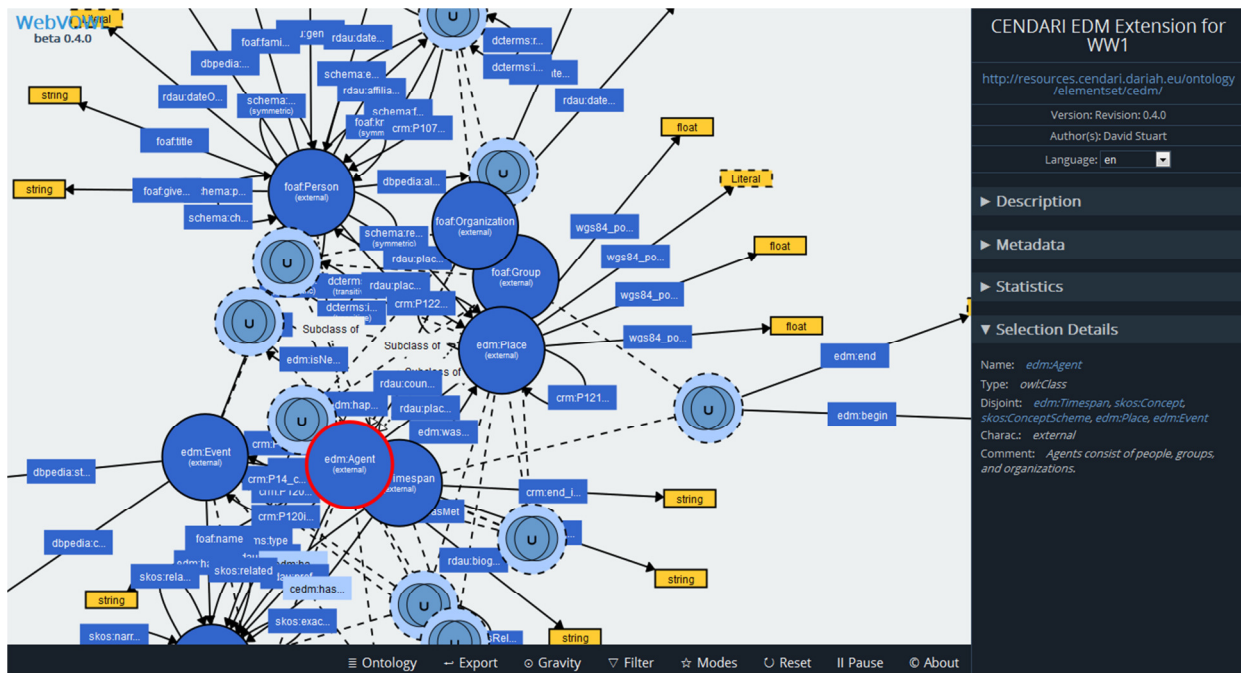[25] Pineapple is currently available only for users with a Dariah account.

extracted with the NER (Named Entity Recognition) among the wealth of archival resources available in the CENDARI data store.

For the **First World War case study,** both the Element Set and the Instances of the "modern" ontology are made available through Pineapple. The screenshots below show both the resources related to the modern period, and the visualisation of the Element Set in a WebOwl visualisation.

Instances of the First World War Ontology in Pineapple

CENDARI EDM Extension for WW1 viewed in WebVowl

For the **Medieval case study**, selected scholarly resources (i.e. medieval manuscripts, authors and texts) curated by SISMEL and FEF, have been exported from a traditional Relational Data Base Management System mapped to the DM2E medieval extension developed within CENDARI and imported into the MeSo Knowledge Base (based on Virtuoso TS).

As of January 2016, the following scholarly information is stored in the MeSo KB:

Medieval Authors	18,353
Medieval Manuscripts	115,105
Medieval Texts	95,727

In the Virtuoso Triplestore the following data is stored (according to the MDV DM2E model extension):

Number of triples	4,146,955
Number of entities	415,341
Number of classes	44
Number of predicates	257

Instances are available through Pineapple (see screenshot below) and/or via the services built on top of the MeSo KB:
- MeSo Search (advanced search over the data on the KB)
- MeSo Browser (Conceptual Navigation over the MDV model and instances)

By selecting the "visualize manuscript shelf mark" button, the user can access the LOD visualization page (MeSo Browser) with information related to the holding institute, manuscripts from the same author as well as other institutes that preserve a copy of the same manuscript.

By selecting the "visualize text" button, the user is directed to the LOD visualisation page with (among others) information related to other manuscripts by the same author and the linked holding institutions (from the MeSo KB).
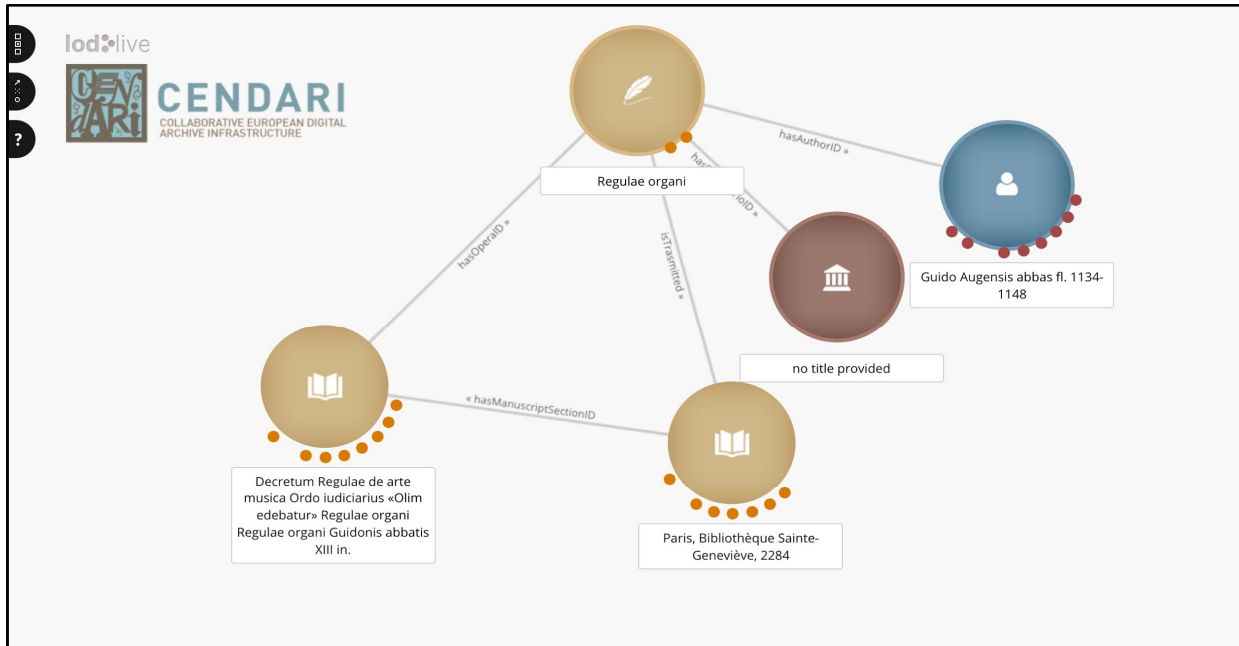
Instances of Medieval Ontologies visualized as Linked Open Data

The contents of the MeSo KB (the medieval LOD) are available through a dedicated search tool providing researchers with both a simplified and an advanced interface, allowing complex searches, faceted browsing and conceptual navigation (via the MeSo Browser)

The MeSO Browser tool for advanced search over the medieval LOD

Making the ontologies available in these numerous formats for direct application, browsing and reuse supports not only their sustainability, but also their baseline utility to underpin scholarly research. Were the ontologies to remain hidden within the portal functionality, then their impact upon searches and connections within the data set would also be hidden, yet it is precisely this level of understanding of the data they use that historians need most. By exposing the ontologies for interrogation, we also have opened them to transparent application and eventual revision for future scholarship.

Key Considerations for the Sustainability of Ontologies:

- **Before creating a new domain-specific ontology check whether similar work has already been done. When possible integrate existing open source ontologies into new work.**
- **Use open formats: Linked Open Data increases the possibility of interoperability with other ontologies**
- **Prepare and make the documentation accessible**
- **Ensure that any ontologies being used to structure data are also available to reuse, adapt, and browse in user friendly manners**

3.4. Project Publications and Knowledge: ARGs

The Archival Research Guides (ARGs) exist as unique CENDARI data, but their specific characteristics justify their discussion as an asset class in themselves.

The ARGs create richly described connections across archival collections with contextualised analysis and related information. They serve as entry points into the CENDARI resources, as well as to some of the transnational topics that will be of interest to CENDARI users, guiding them to different content and through the application of the tools and services available within the Virtual Research Environment (VRE). As such, they are not only finding aids or acts of scholarship, they also exemplify the enhancement of the traditional methods of historical research provided by the project.

The ARGs facilitate the retrieval of collections and complement the descriptions of the holdings within the CENDARI Archive Directory by providing historiographical context and methodological support to researchers. They cover broad thematic areas of historical research and serve also as examples for user-generated annotations, which may also be further developed in a guide format. Although a number of possible formats for the presentation of the ARGs were tested within the course of the project, in the end we decided that only the CENDARI NTE would enable implementation of all of the potential functionality we wanted to see in the ARGs, such as incorporation in the faceted search and browse, integration with the project ontologies, inclusion of multimedia elements *et al*. Finally, the

advances the ARGs have made in developing a new form of scholarly output that lays a structure over parts of the research process, formalising certain steps and creating a place for collaboration and sharing of preliminary results should also continue to be capitalised upon with the historical research community itself.

In the context of the Note Taking Environment, the ARGs can be accessed in a 'read-only' mode directly from the CENDARI website, at the web page: http://www.cendari.eu/thematic-research-guides/available-research-guides. Through this link, the guides are visualized in the Note Taking Environment, and a password is not required (a password is required, however, if the user would like to annotate them or use their content along with other functionalities of the NTE).

Although 25 exemplary guides have been completed by the CENDARI team, it is important to remember that these are not intended to be the only representatives of their kind. The NTE offers users the tools and facilities needed to expose their own work in this format as well, and it is hoped that the CENDARI workbench can therefore also become a showcase for the scholarship it facilitates. For this reason, however, we needed to ensure that future users would have the same access to some of the sustainable publishing modalities as we have built in to the original CENDARI guides.

Archival Research Guide "The Fall of the House Romanov" in the Notes Taking Environment

CENDARI considers the ARGs both as a resource and a tool for enhancing the research practices of modern and medieval historians. By connecting and contextualising archival collections, the ARGs allow historians to connect available archival resources and to

annotate them, to ease their retrieval and organisation. However the Guides can be considered as resources as well, as a digital publication through which the researcher disseminates her research outputs as well as accessing the ARGs in the same way they access academic articles. The license applied to the content in the ARGs is the CC-BY Creative Commons attribution – http://creativecommons.org/licenses/by/4.0/ following general rules of scholarly attribution practices and general practice within CENDARI.

The following sections investigate what is needed in terms of long term sustainability (defined in part, due to the nature of the scholarly publication and communications ecosystem, as visibility and distribution) of the ARGs when considered as scholarly publications.

3.4.1 Alternate formats and locations for the ARGs

As the ARGs are comprised of a number of elements, such as text, annotated entities, images and links, their sustainability is related to the capacity to preserve in the long term all of these elements. As such the sustainability of the ARGs in their fully realised format is related to the sustainability of the NTE itself: without the NTE, the ARGs are indeed not accessible in their interactive form. For this reason, much of their existence depends on that of the NTE, discussed above as a technical component.

We recognise, however, that not all historians are familiar with digital technologies and comfortable with interactive scholarly environments.   To facilitate the retrieval of the guides by such users, CENDARI is also facilitating access to the Guides in alternate forms available through alternate access points. An export functionality has been provided from the Note Taking Environment, through which the text and the entities contained in an ARG can be easily exported in RDFA-XML format; the file can be then opened by any text editor (i.e. Word; Notepad, Openoffice) and saved in a read-only format, as pdf. This functionality will be available to any users of the NTE for the management of their research work.

For the specific purpose of presenting and disseminating the CENDARI ARGs outside of the NTE, each guide has been exported through this RDFA - XML to PDF workflow. In addition, an In Design template has been created to give each publication the same visual identity and structure. These PDFs are being made available from the CENDARI website at http://www.cendari.eu/thematic-research-guides/available-research-guides.

CENDARI Archival Research Guide

ITALIAN BOOKS OF POETRY

Abstract

The aim of this Archival Research Guide is to provide the reader with a brief overview of the history of textual tradition of medieval Italian poets, which is connected to complex events that shaped the transcription of verses in manuscripts. The Italian lyric heritage, from the Sicilian and Siculo-Tuscan School to Stilnovo movement, survives in a small number of manuscripts, which preserve the entire corpus of 13th c. Italian poetry. Each of the main manuscript witnesses will be examined, stressing their characteristics and the organizational structures of the poetry collections, along with digital tools including databases and digital collections.

Introduction

Almost all early Italian vernacular poetry survives in small number of manuscripts. For a few specific authors there are many surviving manuscripts, but on the whole the manuscripts were designed to be anthological collections of texts. Many of these manuscripts have a clear organizational structure, in which the authors are arranged according their literary importance or their social authority and poems are classified by the poetic forms or in alphabetical order. Italian poetry of the 13th survives almostly entirely in three major lyric collections (canzonieri):

- Vatican, Bibliotheca Apostolica Vaticana, Vaticano latino 3793 (V)
- Florence, Biblioteca Medicea Laurenziana, Rediano 9 (L)
- Florence, Biblioteca Nazionale Centrale, Palatino 418 (P)

Florence, Biblioteca Nazionale Centrale, Banco Rari 217, f. 2r
Image by Biblioteca Nazionale Centrale

These manuscripts are books of poetry edited and copied by anonymous Tuscan scribes. They contain the works of 13th c. Sicilian and Tuscan authors who belong to the generation immediately prior to Dante Alighieri. The diverse linguistic origins of the three manuscripts (Pistoia, Pisa and Florence) influence the authors' choice and selection, with connections to specific territorial areas. While each manuscript witness is different, they provided standardized traditions that remained as the reference point for all future transmission of poetry. Guittone d'Arezzo is an exception, as in L he is the major author, while in P his work both begins and closes the most important section of the manuscript dedicated to poetry (canzoni). The V manuscript, while focused on the work of Florentines, includes Guittone d'Arezzo's poems.

All these manuscripts share the subdivision by poetic forms (songs, sonnets in L and V; songs, sonnets and ballads in P), however the final selection and organisation of the poems is different. P is possibly the oldest, and focuses on the courtly love theme, which

6

would be developed by the Stilnovo movement and find its full expression in the works of Petrarch. The P manuscript is also notable due to its large illuminated letters and an unusual independent section on ballads, which is placed before the sonnets. In contrast, L, which was copied mostly by two scribes (along with other figures) is a very large manuscript based on the work of the poet Guittone d'Arezzo. Apart from his poems, his letters are also quoted along with an anthology of predecessors and contemporaries. Codex V, containing about 1000 poems, was planned and copied by a merchant skilled in poetry. It is considered to be a supreme example of lay culture. V is organized chronologically, and offers a historiography of poetry, from the Sicilians to the Florentines Chiaro Davanzati and Monte Andrea, through to Guido Guinizzelli, Bonagiunta Orbicciani and Guittone.

The successive period, called Dolce Stilnovo according to Dante's definition, survives in collections of lyrical poems from early 14th century to mid-15th century. These manuscripts include both works in Stilnovo and the poems of contemporary authors, including comical-realistic sonnets and a few works from the earlier period. The oldest collection is from the Veneto region, dated around the first part of the 14th century, and compiled by the poet Nicolò de' Rossi's (Vatican, Biblioteca Apostolica Vaticana, Barb. lat. 3953 = B). Another manuscript of this type is the Escorial codex (El Escorial, e.III.23 = E). The most important manuscript of these collections is commonly named Chigiano (Vatican, Biblioteca Apostolica Vaticana, Chigi L.VIII.305 = Ch). It is from the Tuscan area, dates to the mid-14th century, and Dante is the principal author. Contemporary and independent from Ch is a lost Tuscan witness, whose text has been reconstructed from a sixteenth-century copy (1523) that Giulio Camillo had prepared for Pietro Bembo (Vatican, Biblioteca Apostolica Vaticana, Vat. lat. 3214 = V2). This last manuscript also contains the Novellino.

An important lost anthology is the so-called Raccolta Aragonese (1476-1477), which has been reconstructed through later manuscripts (Florence, Biblioteca Medicea Laurenziana, Pluteo 90 inf. 37; Paris, Bibliothèque Nationale de France, It. 554; Florence, Biblioteca Nazionale Centrale, Palatino 204). It was compiled as a gift for Frederick of Aragon (at his request) by Lorenzo de' Medici and Angelo Poliziano. It aimed to create a canon of Tuscan poetry from 13th c. Florentine lyrics to the work of Lorenzo himself.

The peak of the philological tradition was reached with the major printed anthologies. A fundamental example of this type is the Giuntina (Sonetti e canzoni di diversi antichi autori toscani in dieci libri raccolte, Firenze, Filippo di Giunta,1527), which was prepared by a group of humanists advised by Bardo Segni. The oldest period of Italian poetry can be reconstructed, based on the Raccolta Aragonese. The canon of the Giuntina was increased by other anthologies, such as the manuscript collected and prepared by Lorenzo Bartolini, called Raccolta Bartoliniana (Florence, Accademia della Crusca, ms. 53) or Lodovico Valeriani's anthology (Poeti del primo secolo della lingua italiana in due volumi raccolti, Florence, Gamba, 1816; vol. I, vol. II).

The Raccolta Bartoliniana is also important for reconstructing another lost manuscript which belonged to Ludovico Beccadelli in the early 16th century. This so-called 'Beccadelli's text' can be also partly reconstructured through excerpts that are found in the first section of Bologna, Biblioteca Universitaria, ms. 1289, which is an anthology of poetry collected by Antonio Giganti for Beccadelli.

7

Archival Research Guide "Italian Books of Poetry" in PDF

These 25 original ARGs will be further promoted through the production of two electronic publications (one for each case study) as collections of 'essays' edited by the leaders of the two case studies, Prof. Janz (Freie Universität Berlin) and Prof. Gumz (University of Birmingham) for the modern and Dr. Emiliano Degl'Innocenti (SISMEL/FEF) for the Medieval. The modern edition will be published as two PDF volumes and made available via the Freie Universität Berlin or other online repository for download. The medieval edition will be possibly published via the Mirabile platform and disseminated via the MeSo DARIAH working group. MIRABILE is the online publishing platform developed and maintained by SISMEL and FEF. It's a widely used scholarly platform, rooted in the international medieval studies landscape, with a strong user base both in EU and the USA, suitable as a platform for data and traditional (books, journals etc.) research products publishing.

Of primary importance with the ARGs, is therefore to ensure not only their accessibility, but also that they are read and used to the greatest possible extent. Their utility as educational modules and as resources for the collections holding institutions should not be underestimated, and finding ways to ensure they are known to the communities has been a part of the strategy around them.

3.4.2. Providing the CENDARI ARGs with Unique and Persistent Identifiers

In order to make the ARGs accessible as stable scholarly publications, the CENDARI team has also developed a workflow to ensure they are assigned a unique identifier - exactly as other scholarly articles available online. Persistent Identifiers, in short PID, represent a tool for making a certain publication suitable to be referenced and cited. This task may sound obvious from the perspective of traditional publications, but in the context of the ARGs the whole original concept of them as have the potential to be "evolving and collaborative works" requires a new conceptualisation for how to both allow them to be dynamic, while also "freezing" them in a stable format at certain points in their development process. Persistent identifiers for scholarly publications represent what an ISBN does for book; the main difference between the two reference codes is that the stable text referred to by the PID needs to be administered and stored somewhere. Exporting and assigning a PID to a guide when it reaches a stable 'version' plateau will allow it to be referenced and cited as a stable scholarly source. Moreover, by including a link in the pdf to the "interactive" version of the ARGs, we hope to encourage the audience accessing the Guides to continue their investigations with the online version in the Notes Taking Environment.

Deciding to seek PIDs for the ARGs was not enough to deliver them, however. The CLARIN "Persistent and Unique Identifiers"[26] report suggests that when disclosing research outputs, we should make sure of two aspects:

1. we should upload the resources in well-established repositories in order to make sure that the articles are disseminated as much as possible. ("which ensure of the existence, accessibility and authenticity of them to make them citable and referenceable")
2. the references pointing to the resources need to be stable, especially in an ever-developing environment where migrations are performed at different levels. [27]

Based on these considerations and the time-limit represented by the end of the project the PID strategy for an ARG has been defined as follows:

1. The 25 original ARGs have been exported as RDF and converted to PDF format as described above. This workflow will be described and recommended to future NTE users as well.
2. The ARGs have then been uploaded to a repository able to assign the publication with a PID. For the CENDARI guides, a choice was made to use TextGrid as the platform of choice (see section 3.4.3 for a discussion of the reasoning behind this decision). This did represent a compromise, as TextGrid (like most other options we evaluated) offers only a handle format rather than a full DOI as their persistent identifier. Unfortunately this format is slightly less robust than the DOI format, but does provide the necessary trace that the purpose required. Future users will have a

---

[26] https://www.clarin.eu/sites/default/files/wg2-2-pid-doc-v4.pdf
[27] Daan Broeder, Malte Dreyer, Marc Kemps-Snijders, Andreas Witt, Marc Kupietz, Peter Wittenburg
Clarin, Persistent ands Unique Identifier Report, 2008.

better option available through the DARIAH ERIC infrastructure, but unfortunately this service will not be available until after the end of the CENDARI project in January 2016.

3.4.3. Deposit with an open research repository

The ARGs available in PDF will be uploaded[28] to the TextGrid Digital Library Platform[29]. This is a simple, open process possible available involving the creation a new project and adding of metadata related to the creation of the document. The platform will make the ARGs freely available and visible to generic and specialist search engines. The TextGrid metadata are indexed by Google, so they can be easily retrieved with a simple search on any browser (and found through the specialist search environment of Google Scholar as well). TextGrid has however some limitations: even though it represents a valuable publishing platform, it is not one of the main platforms used by modern and medieval historians to find bibliographical resources. Its purpose for CENDARI is to provide a stable digital home for the PDF format ARGs that is registered and that supplies PIDs, two things we could not do, for example, by putting the PDFs only on the project website.

Of particular benefit in the choice of TextGrid is that fact that the TextGrid PID handles will be transformed into DOI automatically by DARIAH-DE in collaboration with the Göttingen State and University Library in the first half of 2016.

Another option considered is the deposit within HAL, 'Hyper Articles en Ligne' [30]. HAL is an independent scholarly resource, an online archive of scholarly articles where authors from any academic field can deposit a copy of his/ her work. HAL has been created by the *Centre pour la communication scientifique directe,* which is part of the *French National Centre for Scientific Research* (CNRS). Although directed primarily to the French academic community, HAL has been used by researchers from every country and this is mirrored by the languages of the publications available through the website, and indeed some earlier CENDARI position papers were hosted there.

The CENDARI ARGs will be deposited in HAL after they will receive a DOI from DARIAH - EU during the first half of 2016.

It is possible, however, that some ARGs will appear, for example in the edited volume, in an institutional repository such as the Freie Universität Berlin's or Trinity College Dublin's. Deposit with such repositories is generally based on an author's affiliation with the institutions providing the service, however, so while they would provide many of the same benefits, they would not have been a good or even possible solution in every case for the

---

[28] As soon as the process of transforming the ARGs in PDFs is completed.

[29] https://textgrid.de/en/digitale-bibliothek

[30] https://hal.archives-ouvertes.fr/

CENDARI original ARGs, though they may be very good options for individual later users of the NTE and export workflow.

3.4.4. Peer Review of the ARG as a form of scholarly communication

The fact that the CENDARI ARGs will be able to be found outside of the NTE will not alone guarantee that either the NTE or the guides themselves receive the recognition they deserve as scholarship. For this reason, a further action toward their sustainability has been to submit the guides for review by RIDE.

RIDE is a review journal for digital editions and digital resources, launched by the University of Cologne's Institut für Dokumentologie und Editorik in 2014. The aim of the review journal is to offer a space for the creators of digital editions to have their work reviewed and published in an online journal. This Journal's mission addresses the ongoing tension that sees digital editions being "neglected by the established review institutions". Authors of digital editions usually receive less credit that their counterpart in the world of printed-publications; moreover as digital editions have a greater level of complexity (as they connect resources of different nature) it is difficult for them to be represented in a traditional publication form.

The RIDE platform also aims to review and publish academic publications that connect digital resources such as cultural heritage records and collections. The ARGs can indeed be thought as a hybrid form between digital editions and digital publications, as they connect cultural heritage records and collections described in the CENDARI repository. Moreover the ARGs offer to their users the possibility of applying computational technologies to their notes, such as (manual and automatic) entity recognition and connection of those entities with the CENDARI ontologies on the domain of the WW1 and medieval culture. From this perspective, the CENDARI ARGs match the criteria defined by the RIDE team.

A first contact with the RIDE Journal was made in 2015 to establish their interest: with the completion of the NTE and new CENDARI web presence in 2016 the submission will now go forward under the CENDARI SUSTAIN Working Group (see section 3.6.1 below).

3.4.5. Presentation through articles in a (traditional) historical journal

Ideally, the scholarly material in the Archival Research Guides would appear in traditional historical journals, both for modern and medieval history. This would be the best way of making the widest possible body of scholars aware of their existence and contribution. Unfortunately, the Guides were not designed with this outlet in mind - in many ways, the traditional monograph form is actually antithetical to the aims and objectives of the ARG as a format. In other words, in order to do so, the Guides would need to be adjusted to a linear

format publication, including an effort by the authors of the ARGs to re-envision what was born as hyperactive and interactive.

It is possible that this action for sustaining the ARGs will take place after the end of the CENDARI project: every ARG is indeed authored by specific CENDARI researchers and they will be responsible both for their transformation in a linear and read-only format, as well as in contacting the proposed academic journals. We hope that they will become a part of later work, and be referenced as such in that context.

Sustainability of the Archival Research Guides:

- **Incorporate export formats into the interactive environment**
- **Ensure their wide availability through a trusted research repository as well as through the NTE**
- **Provide each ARG with a Unique and Persistent Identifier**
- **Validate their quality through an appropriate peer review mechanism**
- **Ensure their visibility to traditional scholars via a traditional mainstream journal publication or edition**

3.5. Project Publications and Knowledge: Toolkits, and Knowhow, Management Data and Assets

CENDARI, like any project, has learned a lot along the way. All of this project learning has come together in the final environment, but only a small amount will actually be visible there. But much of this 'hidden' knowledge could be of great value to future projects addressing similar problems, and therefore should be made available wherever possible.

Some of this know-how will already be documented, for example in presentations to CENDARI summer schools, technical documentation, training material, or the process documentation regarding the CENDARI toolkit for the approach to archives. Making this wealth of documentation accessible in an appropriate place and format is crucial, for it tracks the history of the project and its outcomes - information of great potential use for future projects. Moreover, the documentation related to the technical infrastructure and the CENDARI training material form the basis for an effective user experience as well as a guide for the reuse of the CENDARI infrastructure (in toto or single applications) by future projects. Often this kind of material is regarded as less relevant or viewed as ancillary to the official documentation; it should be however considered with the highest priority, as it has the potential to support prospective and future users. Where such documentation exists, it must

be secured somewhere for long-term availability to the DARIAH knowledge base. If properly contextualised and made visible to its successors, such information could itself be reused, bringing greater efficiency to the research infrastructure development process and contributing to the sustained value of the CENDARI investment. In an ideal world, this would mean that the information that helped to build one infrastructure is easily available and therefore less inclined to be forgotten or abandoned.

In other cases, the project has endeavoured, in its final year, to determine a list of areas where such know-how could be lost, and use the very traditional knowledge transfer mechanism of the scholarly article or report to ensure these experiences are shared with the larger community for the long term. In particular, CENDARI has tried to ensure that some of its 'failures' were exposed (such as software components that were tested and rejected): this information in particular could all too easily remain hidden, in spite of its obvious potential utility for successor projects.

The CENDARI approach to harvesting and securing its legacy as a project and as a base of tacit knowledge is described below.

3.5.1. The 'Tacit Knowledge Audit'

One of the most challenging actions emerging from the CENDARI project's sustainability planning process was the commitment to expose the areas of tacit knowledge the project had developed. The Project Management Board (made up of the 9 technical work package leaders) were tasked in early 2015 to undertake a 'tacit knowledge audit' within their work packages, seeking to identify challenges that were met and managed, but for which the substantial progress made toward refining good progress within the project may not have a public face.

Several areas of well-developed tacit knowledge were defined, the two most prominent of which were the project-wide commitment to participatory design and the experience of trying to harvest content from a wide variety of archives and other content holding institutions. The first of these will become the focus of a publication or formal conference paper in the course of 2016-17. The latter topic was captured in the form of a 'white paper' report, which was released as a printed publication at the time of the CENDARI launch (also available for download from the project website: http://www.cendari.eu/thematic-research-guides/white-book-archives) and in the form of a CENDARI ARG.

Other areas of broad innovation had already been released, such as the unique approach the project adopted to federating research data, and the conceptual underpinnings of the ARG as a micro-publication. Both of these appeared as peer reviewed or pre-print publications within the lifetime of the project. A final area of significant progress identified was the interaction between CENDARI, DARIAH and other related research infrastructure projects, which is largely the topic of this document, the CENDARI Sustainability Plan.

3.5.2. Documentation for End Users and Training Material

The CENDARI team produced a wealth of training material both for its testers and potential users: among them are for example, are videos, use cases, tools documentation and user manuals. The CENDARI team in the last years has reflected on both how these training materials could be best presented to users of the final platform and disseminated outside of the project.

For the former audience, it was decided the contextual integration of the resources into the final CENDARI website would bring the most value. The website offers users entry points to 'Search the Archives,' 'Working Spaces' and 'Thematic Research Guides.' For each of these pathways, both a light informational 'landing page' and a more detailed 'Intro" page is offered. The majority of the training material is offered through these pages, though other and more detailed information is sometimes also available through the pathway 'About Us' > 'User Guides' (http://www.cendari.eu/log/user-guides). The documentation and material available at the following links has partly been developed for the training sessions developed in the context of the "Trusted User Group" monthly sessions, which will be introduced later in this document.

ARCHIVAL DIRECTORY
- Introduction http://www.cendari.eu/search-archives/intro-archival-directory
- User Guide, CENDARI archival Directory
  https://docs.cendari.dariah.eu/user/atom.html

NOTES TAKING ENVIRONMENT
http://www.cendari.eu/working-spaces/intro-working-spaces
- Working with the Notes Taking Environment:
  https://docs.cendari.dariah.eu/user/writing_arg.html
- Introduction Video: https://youtu.be/128eoXTrfVk

ARCHIVAL RESEARCH GUIDES:
- Introduction http://www.cendari.eu/thematic-research-guides/intro-thematic-research-guides

TRAME
- Introduction: https://docs.cendari.dariah.eu/user/trame.html

For the most part, CENDARI's training material was designed specifically to assist users in making the most of their CENDARI experience. Like any infrastructure project, however, we recognise that we may have developed something with a wider value we cannot necessarily see. For this reason, the CENDARI project is connected with both the **PARTHENOS** Cluster

Project (http://www.parthenos-project.eu/) and with **DARIAH Teach** (http://dariah.eu/teach/index.php/2015/05/21/welcome-to-dariahteach/). PARTHENOS brings together a number of research infrastructure projects, including the DARIAH and CLARIN ERICs, but also cognate projects like EHRI, ARIADNE and IPERION-CH. The Cluster has a broad remit upon which to deliver by 2019 (and therefore a broader role in CENDARI's sustainability, see section 3.6 below), but CENDARI team members are particularly involved in the delivery of the training strategy within the cluster.

DARIAH Teach is "developing open-source, high quality, multilingual teaching materials for the digital arts and humanities" and "Its goal is to strengthen alliances and foster innovative teaching and learning practices among members of the DARIAH (Digital Research Infrastructure for the Arts and Humanities) network." [31]

In this context, CENDARI took part in the workshop "Open Humanities" organized by DARIAH Teach in Belgrade in November 2015. DARIAH Teach represents an ideal place where to make available to end users the training developed by CENDARI: as an example the training material on the development of the CENDARI ARGs as forms of enhanced publications can support researchers to open up to new forms of scholarly publication, where the textual element is complemented with primary and secondary sources and linked to ontologies or controlled vocabularies. Alternatively, the CENDARI training material can prepare historians to encode archival institutions and archival collections in ATOM as this represents a very good training for historians who want to learn more on how to encode historical resources in XML EAD and EAG.

DARIAH Teach will run through to June 2017 and the conversation on the possibility to include the CENDARI training material in the DARIAH Teach platform will continue after the official end of the CENDARI project. This task will be followed by the CENDARI SUSTAIN Working Group.

3.5.3. Technical Documentation

The CENDARI **technical documentation** (http://www.cendari.eu/about-us/technical-documentation) includes a user guide to the applications developed in the context of the technical research infrastructure as well as examples of uses cases and answers to frequently asked questions. This document, which is also available in PDF (http://www.cendari.eu/sites/default/files/CENDARI%20technical%20documentation.pdf) provides a picture of the CENDARI set of tools, as of October 2015 and is accessible under a CC-BY 4.0 license.

The document is structured in **four main sections**:

---

[31] http://dariah.eu/teach/index.php/2015/05/21/welcome-to-dariahteach/

- **Introduction**: gives an overview of the components, including back-end applications, data store and user interface.
- **User Guide**: HowTo's for the single applications
- **Administrator Guide**: how the components have been set up and connected
- **Developer**: Actual implementation of the components. When possible, inline source code of the single components has been used to edit this section.

3.5.4. Internal Project Documentation and Management Data

As a part of CENDARI's commitments to both responsible data management and to increasing the capacity for later discovery and application of tacit knowledge generated in the project, the raw management assets of the CENDARI project also need to be accounted for.

These assets reside in a number of different locations. Basic project management has been facilitated through the use of the **Basecamp** platform. Basecamp was a good choice for CENDARI because of its familiarity and flexibility, but unfortunately as it is a subscription-based, external service it is not feasible to use this as a long-term solution. Basecamp will be maintained for a period of several months after project close, in which time the assets will be removed and organised on a separate external drive.

In addition, CENDARI also used a **Confluence** wiki as a shared work space. Most of what is currently housed on Confluence is pre-final or working versions of final documents and deliverables, but it is conceivable that some of these versions and position papers might have later value. As a DARIAH service, Confluence will remain available to the project team members, but for the purpose of good organisation, this material will also be held on the same drive as the Basecamp material. At the end of the day, it will also be the Project Coordinator's responsibility to retain a full record of the project for 20 years (to meet potential audit requirements) so her contact details should remain associated with the project portal in case deeper knowledge is required.

The third current location of working data from the project has been the **JIRA** bug tracking system. In particular, we used this system to log and track our contacts with archives and other content holding institutions, information which could be useful to a potential future CENDARI extension team. This material will be exported into a flat file system for inclusion in the closed project-level documentation.

The final source of information to be included in the closed drive will be the central administration team's **internal files and shared drive**. These assets, including important and unique internal files such as the original and high resolution logos and images, are

recognised for their potential future utility, not to a public audience but to future potential successor projects.

Sustainability of Documentation and Training Material:

- **Training material and technical documentation need to be available readily and in context.**
- **Project documentation is not only important within the project, but a valuable potential resource for external researchers/ users as well**
- **Tacit knowledge and management data are two important resources of a project that are often not considered in terms of overall project sustainability.**

3.6. Communities: People, Networks and Relationships

Perhaps the least easily defined and sustained aspects of the CENDARI project will be the communities it has brought together and contributed to. Collaboration with research communities has been an important aspect of the CENDARI project. As emerged in the previous sections of this document, the entire sustainability of the CENDARI research infrastructure depends on its use by historians, developers and other researchers: without this, the whole research infrastructure has very little reason to be supported and sustained. CENDARI has therefore strongly focused on the community of users from the very beginning of its life, and it's planning to continue to do so even after its funding period.

Many of the groups we would consider continuing our engagement with are those we convened for the purpose of developing the project. During the prototyping phase of our tools, we organised **participatory design workshops**, where historians created video prototypes, mock-ups of research tools and other visualisation features. Based on their needs and research methodologies, the CENDARI team drafted the tools requirements that were necessary to design and develop the research infrastructure. The **CENDARI summer schools, external advisory board** and the **transnational fellowships holders** represented a valuable way to get insight from historians about their changing practices and to have feedback about the usability of the Virtual Research Environment. Additionally, in the last year of activities CENDARI set up a **Trusted User Group**. The CENDARI Trusted User Group served the need of connecting the software developments with the community of modern and medieval historians, who volunteered with the testing of certain functionalities in the Virtual Research Environment. The Trusted User Group met monthly via webcast, and their involvement as a group of users in the development of the CENDARI research infrastructure has been revealed as a valuable tool for the continuous and technical

improvement of its components. The Trusted User Group has also enabled the production of a wealth of valuable training material: this was developed in view of the testing sessions, but it nonetheless constitutes a valuable training material for other researchers interested in our activities. [32]

CENDARI developed a number of highly productive external links to networks and projects as well, however. In terms of its **users**, the International Society for First World War Studies, the Leeds International Medieval Congress, and the COST Action IS 1005 all represented groups of our users with whom we had informal but productive relationships. Ensuring these groups stay engaged and aware of CENDARI is one key to ensuring these groups can continue to act as gateways for our users. We also interacted with a huge range of cultural heritage institutions and other **content holders**: being able to continue to interact with these key partners, and indeed to be able to ingest new data as it comes available is another key network implication for the post-funding stage of the project. Among these, large aggregators such as Europeana and APEx stand out as collectors we would like to see consider CENDARI as still a viable conduit to key audiences. Finally, the many **parallel projects** with which we have engaged, including but by no means limited to DARIAH, have formed a very useful network for sharing ideas and supporting each other. CENDARI and EHRI, for example, have been working together already to support a more productive dialogue between research infrastructure projects and content providers. This kind of activity will continue as well under the umbrella of the Horizon 2020 Cluster project PARTHENOS.

3.6.1. Basic Continuity of Communications

The established CENDARI website URL (www.cendari.eu) and email address (info@cendari@eu) will remain as the central points of access to and communication with the project. Basic information about how to use the site and request an account is also included on the 'Terms of Service' page on the CENDARI website.

3.6.2. DARIAH Working Group "CENDARI Sustain"

Knowing the power and potential of the CENDARI networks for supporting sustainability of its outcomes and delivering on that potential are two very different things, however. For this reason, the Project Coordinator and core partners of the CENDARI project have proposed to form the **CENDARI Sustain - DARIAH Working Group**.

---

[32] This point is also discussed in the section 3.5.2

The CENDARI SUSTAIN Working Group is a crucial element of the CENDARI project Sustainability Plan. In order to ensure that the user communities and tools that have been built during the lifetime of the project continue to have a context for their maintenance and development, a core of the CENDARI leaders will contribute to the WG over the course of three years (until February 2019) with the intention of seeking further funding for active development at the end of that period.

The actions of the CENDARI Sustain Working Group are twofold: on the one hand it will make sure that the CENDARI users will be granted access to the CENDARI Research Infrastructure as well as to the main CENDARI website. On the other hand it commits to maintain and to extend the community of users that CENDARI has created in the last years of activity.

To achieve these objectives, the CENDARI-SUSTAIN Working Group will seek to achieve the following:

- Maintain and manage contact with the CENDARI user communities (medieval, modern, Trusted User Group) and any new communities wanting to come forward to utilise/further develop the CENDARI core platform
    - maintain website and social media presence
    - act as point of contact for finished project
    - deepen engagement with current users (medieval and modern)
    - liaise with DARIAH DE regarding approval of new users
    - promote CENDARI with potential new user groups and work to enable their reuse of project outcomes
    - propose and run outreach events under the DARIAH umbrella and potentially with funding from DARIAH small grants schemes
- Sustain engagement with cultural heritage institutions to better understand their needs and continue to develop the CENDARI data resources
- Maintain contact with DARIAH DE (VCC1) regarding use of the CENDARI platform and any changes to user/data management policies or procedures that may be required
- Maintain a point of contact with the developers responsible for the CENDARI tools suite (NTE, API, Repository, ingest modes etc.) to ensure reliability in the provision of services
- Continue work started with VCC3 within the active phase of the project to align and federate data across DARIAH associated projects
- Promote the project results at scientific and other meetings, as well seeking to extend the development of the infrastructure and its communities
- Liaise with other projects seeking to link their sustainability plans to DARIAH in the manner of CENDARI

The aforementioned tasks related to these two strands of activities will be contextualized in the following sections, and in particular in the section related to the Sustainability of the CENDARI Website and the technical infrastructure, as well as the sustainability of the community of users.

Sustainability of the User Community:

- **Maintain a consistent central communication point, even after active project close**
- **Recognise the collecting points available for users - both those developed within the project and those with external leadership**
- **Identify a small group of people that can dedicate time to strengthen the community of users, even after the end of the funding period.**
- **Provide to the end users instruments or forms to contact the project's team, to report bugs and usage of tools**
- **Make use of DARIAH mechanisms to give this loose confederation a platform for continuing to work together informally**

---

4. Recommendations for Future Projects

Much of what has been written above is specific to the CENDARI project and its particular assets, but many of the mechanisms for sustainability planning and assurance can be extrapolated into good practice guidelines for future DARIAH-affiliated projects. At a minimum, we would recommend the following:

1. Projects should make sustainability a part of their initial funding proposals, including at a minimum a sustainability meeting at the project mid-point and a sustainability report as a deliverable.

2. Knowledge sharing across the DARIAH affiliated projects should become the norm, not just at the level of project heads, but across the cognate skills areas within projects. Mapping project structures against each other should occur as a matter of course. Projects should name representatives to all relevant DARIAH VCCs, and a project-focussed sidebar event should become a part of the annual DARIAH GVCC schedule, and should include project proposers as well as funded projects.

3. Projects should use common platforms, such as GitHub, to share their results, making the project developments available by multiple modes and in multiple places.

4. Projects should, wherever possible, apply known and accepted standards in their work. These may be drawn from a number of relevant communities (TEI for scholarly editions, EAD or ISAD for archival material, etc.) but maintaining reference to a larger context for the organisation of data should be viewed as an essential element of its potential reusability. In the future, the DARIAH survival kit and helpdesk should be developed and deployed to assist with this.

5. Project publications should cover scientific, technical and management aspects of the project, with different audiences in mind for different pieces.

6. Every project should have a data management plan, explaining what data it will collect or create and where it will be left, including as well any services the project will create and how they will be sustained. This plan should include consideration of data created by end users: the prerequisite for a strong and durable community is indeed that all its data are accessible in the long-term. Users should not hesitate to work in an environment for fear that they might at some point lose access to them.

7. Projects should guarantee the availability of their basic technical formats for 3 or 5 years. It does not seem reasonable to expect or support a longer window for a project not under active development, as within that time it will need to be updated anyway.

8. Where projects continue to prove their utility to new and old user communities within and beyond this basic window, successor funding projects to focus on new user groups or new functionalities should naturally come to be sought, and DARIAH should encourage and support this in order to keep useful and 'evolving and involving.'