THE CENDARI WHITE BOOK OF ARCHIVES

Data Exchange Recommendations

for Cultural Heritage Institutions and Infrastructure Projects

CENDARI
COLLABORATIVE EUROPEAN DIGITAL ARCHIVE INFRASTRUCTURE

THE CENDARI WHITE BOOK OF ARCHIVES

AUTHOR(S)	Jakub Beneš, Nataša Bulatović, Jennifer Edmond, Milica Knežević, Jörg Lehmann, Francesca Morselli, Andrei Zamoiski
COLLABORATOR(S)	
THEMES	Archives, Data Exchange, Metadata, Repository, Data Ingestion, Cultural Heritage Institutions
PERIOD	19th century - 21st century

CENDARI is funded by the European Commission's 7th Framework Programme for Research

TABLE OF CONTENTS

The CENDARI White Book of Archives

EXECUTIVE SUMMARY

Over the course of its four year project timeline, the CENDARI project has collected archival descriptions and metadata in various formats from a broad range of cultural heritage institutions. These data were drawn together in a single repository and are being stored there. The repository contains curated data which has been manually established by the CENDARI team as well as data acquired from small, 'hidden' archives in spreadsheet format or from big aggregators with advanced data exchange tools in place.

While the acquisition and curation of heterogeneous data in a single repository presents a technical challenge in itself, the ingestion of data into the CENDARI repository also opens up the possibility to process and index them through data extraction, entity recognition, semantic enhancement and other transformations. In this way the CENDARI project was able to act as a bridge between cultural heritage institutions and historical researchers, insofar as it drew together holdings from a broad range of institutions and enabled the browsing of this heterogeneous content within a single search space.

This paper describes a broad range of ways in which the CENDARI project acquired data from cultural heritage institutions as well as the necessary technical background. In exemplifying diverse data creation or acquisition strategies, multiple formats and technical solutions, assets and drawbacks of a repository, this "White Book" aims at providing guidance and advice as well as best practices for archivists and cultural heritage institutions collaborating or planning to collaborate with infrastructure projects.

THE CENDARI WHITE BOOK OF ARCHIVES

INTRODUCTION

There is a wealth of historical material stored in Europe's cultural heritage institutions. Thousands of records relevant for research are held within every one; the physical artefacts preserved and organised so as to expose their provenance. Each record group can be consulted via analogue inventories and other finding aids, which are fully accessible within a given institution and often fully accessible online as well. The development of these virtual resources has occurred in concert with changes in historical methodology, where the pursuit of transnational topics and application of techniques such as data mining require and privilege digital sources that are fully exposed and widely available.

In general, libraries and museums are at a much more advanced stage than archives in sharing and presenting their holdings to the broader public. These institutions have long since formed clusters to exchange data on the material they hold. More recently archives have started to apply similar methods and to build collaborations and networks in order to promote the development of their digital presence. Many institutions are still in the process of digitizing their analogue catalogues and inventories and are far from completing this task. From the historian's perspective, therefore, that landscape of provision can be quite uneven: while some institutions provide excellent digital access, others may display only some unstructured information in PDF or Word format on their websites. This variability in provision has its roots in the lack of resources faced by many cultural heritage institutions, as well as in institutional cultures well-adapted to analogue processes and hierarchies, which may be different from their digital or virtual era equivalents. These restrictions can hold the institutions back from gaining the full benefit of the 'digital turn' in historical scholarship, and in particular from being able to fully participate in research infrastructure projects like CENDARI.

CENDARI – the Collaborative European Digital Archival Research Infrastructure – began its work in 2012, charged with a mandate to 'integrate digital archival resources for medieval and modern history'. In order to deliver on CENDARI's case studies of World War I and Medieval Culture, the CENDARI team identified and contacted more than 250 cultural heritage institutions whose collections were deemed pivotal enablers for research, and priority collections to expose in the CENDARI Virtual Research Environment (VRE). Apart from desk research, the CENDARI team contacted cultural heritage institutions to establish a basis for data exchange, i.e. collection descriptions already existing in digital formats and displayed on individual websites of the institutions. The aim was to bring them together in one single repository – the CENDARI repository – thus facilitating the enhanced processing and enrichment of the data. At the same time, these contacting activities, including Skype conferences, phone calls, letters and on-site visits, were used to trace less visible archival information relevant to the CENDARI case studies, and to provide digital descriptions of collections where they didn't yet exist.

Because the cultural hertitage institutions involved are major research archives, museums and libraries, they all had some digital presence or ongoing digitisation activities. This

presence did not guarantee the full display of information on all relevant collections and holdings, nor – and more importantly – did the existence of digital finding aids and collections mean that these resources could be accessed as 'open data' (that is, without restrictions for their reuse) by a search engine from outside of the institution's own closed data environment, e.g. via an application programming interface (API). The omnipresence of such difficulties challenged the CENDARI team to develop a robust approach to data acquisition and data management, satisfying the needs of both the institutions holding the data and the potential future users of CENDARI's archival descriptions.

It is the purpose of this document – the CENDARI White Book of Archives – to document this process and the mechanisms deployed by the project to bring diverse archival sources together. The CENDARI team does so in the strong belief that the historical research ecosystem requires better pathways for the open exchange of data between cultural heritage institutions and the digital infrastructures that will increasingly mediate between them and their advanced users. The CENDARI team does so also in the hope that future projects will be able to benefit from the experiences gained, and build upon them toward this vision.

TWO HISTORICAL REALMS, TWO WAYS OF AGGREGATING CONTENT

Although both of the CENDARI case studies revolve around the study of history, the project team quickly learned how different the two cases were, not only in terms of how research questions were designed and pursued, but also in the distribution and digital preparation level of the relevant source material. Archival institutions relevant for the CENDARI case studies were therefore identified according to the following sets of criteria:

World War I archives selection criteria:	Medieval culture archives selection criteria:
• All countries that participated in the First World War are included. • Special attention is given to records in Eastern and South East Europe, in order to highlight so-called 'hidden archives', which do not have any digital representation of their sources. • Archival institutions with significant holdings are described, with priority given to central national archives, national military archives, national war museums etc. • Archives that possess a large selection of digitized finding aids.	• Archives in most Western countries, including the USA and Australia. • Special attention has been paid to archives in Eastern and South East Europe. • Collections of great interest to the academic community, and which are frequently cited in scholarly publications. • Archival institutions that digitized a large part of their collections.

Altogether more than 1,200 archival institutions were added to CENDARI's repository, using an international standard format, the Encoded Archival Guide (EAG).

In a second step, more granular descriptions of specific relevant archival holdings were produced by CENDARI team members. Given the large number of relevant institutions identified and the breadth of their individual holdings, the project adopted a strategy to prioritise 1) collections most relevant for the international research community and 2) collections most relevant for the Archival Research Guides (ARGs) to be produced by the project. In this way, the project captured both the key 'backbone' collections, which would provide an overall context for other material, and those 'hidden' archives that would be better exposed through the thematic guides. In this way, the CENDARI description strategy served the focussed needs of the Archival Research Guides, while also devoting special attention and resources to archives with little or no digital presence for their collections, to the level of not even having websites with basic information on the archive itself, often in the new European member states of Eastern Central Europe.

The most important holdings were identified and described in the international standard format Encoded Archival Description (EAD). More than 2,700 archival descriptions were created manually, chosen according to a transnational perspective and representing nearly all European languages while at the same time providing translations into English. These manually created archival descriptions can be regarded as "golden data" of a high quality, carefully curated by CENDARI team members and leveraging the benefits of drawing together dispersed information in a common archival repository. Although uniformly structured and designed for reusability, this information comes from a variety of sources and approaches: gathered by desk research or received from the archivists themselves, (e.g. through on-site visits to take up the descriptions available in repositories, guidebooks and finding aids), and even by close investigation of the material itself. The advantage of such an approach is that not only does it make archival material visible in the CENDARI infrastructure and thus presents sources of which an individual researcher would not have become aware, but also that it points toward valuable and relevant collections which may not be evident from the formal description created by the archive. Furthermore, the translation into English opens up the visibility of these holdings further still. The disadvantages of this approach clearly lie in the fact that manual creation of archival descriptions is extremely time-consuming and in the incapability of archival descriptions to be fully standardized, for example in regard to common labelling, or in the fact that they can follow the interests of the individual researcher. As such, the CENDARI project team risked introducing knowledge frameworks that could as easily disguise as they could expose sources to potential users, but the team felt this risk was minimised due to the use of linked data technologies as well as an evolving approach to data integration, which are at the core of the project infrastructure. In this way, nearly infinite future extension of the records is enabled which can incorporate further (even conflicting) information about the collections in question.

This process of manual creation of archival descriptions also served as a model for future users of the Virtual Research Environment, since users will be able to establish archival descriptions themselves in the final environment. The software used for this – called AtoM or Access to Memory (https://www.accesstomemory.org/en/) – is user-friendly and controls

the quality of the descriptions through the inclusion of mandatory fields, facilitating export of the content in a standardized format, i.e. Encoded Archival Description (EAD).

Manual establishment of archival descriptions was accompanied by data acquisition from data providers and institutions holding relevant data which satisfied the CENDARI selection criteria listed above. Data acquired during this process varied from structured data offered via an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) or another kind of technical API, up to data dumps in various formats (e.g. EAD, RDF) sent directly to the technical team or transferred via an FTP interface. At the other end of the technical-sharing spectrum, the technical team was only provided with "links", i.e. URLs to web pages where some information has been published on the web site or in pdf documents, and thus had to scrape and further process the data.

There were some differences between the medieval and the modern cases with respect to this workflow. The medievalists were fortunate to be able to build upon previous work with the project partners to aggregate relevant data; this has been done primarily in the form of a metasearch engine called TRAME. Because a vast majority of medieval manuscripts are held in particular (and often well-developed) library repositories, a vast amount more data was available via OAI-PMH and could be harvested. This meant that far fewer descriptions needed to be created manually than in the modern case. Because of the high technical bar they were able to start from, the medieval team also invested in the development of further tools to enhance the data they collected. Examples were a scraper tool to register data visible to a human user but not sharable from a technical side; and the 'triplification' of database resources, enabling them to be deployed beneath and discovered alongside collection descriptions.

LEGAL FRAMEWORK

By establishing a central repository, CENDARI aimed to protect the rights derived from the cultural heritage institutions as owners of the records, and from a technical point of view to guarantee the creation and preservation of stable links as well as a sustainable storage of this information. The repository stores not only data established by CENDARIs partners, but also all the archival descriptions created manually in the archival directory AtoM, which are transferred into the repository on an automatic basis. The repository contains no digitized material, but only references to it, if available. The CENDARI repository is based on the CKAN open source software (http://ckan.org).

Sharing of data is a sensitive issue, however, even where institutions are ready and able to share. To enable this, CENDARI developed a legal framework for the sharing and reuse of data that consists of a Data Exchange Agreement and Data License, based upon the Creative Commons license CC-BY 3.0 (http://creativecommons.org/licenses/by/3.0/ – unless otherwise specified by the data owner), and a Memorandum of Understanding (MoU) with the DARIAH-ERIC to guarantee the long term maintenance of the data. These documents formed the basis for establishing contacts with important cultural heritage institutions, and were further supported by a suite of internal and external communications tools, such as a set of FAQs for institutions and an internal flow chart mapping different approaches for different types of partners (see appendix).

With some holders of content, no formal agreement for sharing was required: Europeana (http://www.europeana.eu/), for example, makes data available for reuse with no further agreement other than an application for an API key. In other cases, a clear statement of CENDARI's policies was enough to satisfy the content holder. As a final category, many institutions required a formal data licence agreement, signed by representatives of CENDARI (for the short term) and DARIAH (for the long term). Where used, the data licence was adapted to the individual needs of the archives, for example where a specific licensing agreement was requested.

Having a set of excellent tools to facilitate data sharing did not make the negotiation with archives simple, however. Institutions under resource pressures and with primarily national mandates are often ill-equipped to respond. Over time, the CENDARI team realised that such agreements requires up to three different institutional specialists to judge the merits and costs of sharing data: someone who knows the content itself, someone who knows the technical infrastructure, and someone in a position of authority to approve the relationship. This complexity was often exacerbated by change processes ongoing in the institutions and the fact that the CENDARI project was not able to invest significantly in assisting institutions to prepare their data in house; as a result the process of developing and signing a data licence agreement was often delayed. Perhaps the project's most critical instrument for data recruitment, therefore, was patience and perseverance: that said, no sharing would have been possible without the legal and social contracts the CENDARI project entered into with the providers.

CENDARI Data Sharing Agreement and Data License

The CENDARI Data Sharing Agreement is a 'light-touch' statement of the commitment CENDARI makes to using and sustaining any data it holds. It describes the goals and philosophy of the project, its relation to the DARIAH ERIC, and canonises the project commitment to a Creative Commons 'by attribution' licensing framework and willingness to work with providers with different requirements. It was intended to inform potential partners of our intentions and commitments, without adopting the language, formality or overhead of the license.

The CENDARI Data Licence is a far more rigorous document, intended to provide a further level of assurance and commitment to collaborating institutions. It is loosely based on the Europeana Data Licence, but shows some minor differences and one major one, namely the commitment to a CC-BY licensing regime. The Data Licence establishes a legal, agreed framework between the cultural heritage institution and the CENDARI project.

The first section of the document, "Provision of data," establishes that the data have to be relevant for the CENDARI case studies in order to be included (i.e. pertaining to World War One and Medieval Culture) and that in case the data are supplied by a third party the data provider needs to clear every authorization before sharing the data with CENDARI.

The second section "Use of Data" sets the creative commons licence under which the data provided are made available to the researchers working in the CENDARI infrastructure: the selected licence is the CC-BY attribution licence even though the provider can request a different licence (i.e. CC0, CC-BY-SA, etc.).

The articles “Terms” and “Termination” establish the validity and the rules regulating the termination of the contract. The last section, entitled “Liability and Notice of Take-Down” sets the conditions by which the data provider or CENDARI might request to remove the data from the CENDARI portal.

These three headings provided the project with a firm basis for ensuring clear communication and agreement between providers and the project where something more robust than the Data Sharing Agreement was required. However, the Data Licence did require support from a few further documents as well and which are described below.

DARIAH Memorandum of Understanding (MoU)

It would not have been credible for the CENDARI project, funded for four years from 2012 to 2016, to commit to an offer to maintain and preserve data for the long term. The Data Sharing Agreement and Licence are therefore underwritten by an agreement between CENDARI and DARIAH-EU, guaranteeing that the terms of the Licence would be upheld by DARIAH after the end of the CENDARI project. This memorandum protects the data providers as it ensures technical support by DARIAH-EU and the sustainability of the content previously shared with CENDARI.

Letter of Collaboration

The letter of collaboration represents a “lighter” version of the CENDARI Data Licence and underlines that a contact has been established or is being created between an archival institution and the CENDARI project. The letter of collaboration can precede or replace the Data Licence.
Frequently Asked Questions
The Frequently Asked Question Document was designed as a support for archival institutions considering collaboration with the CENDARI project. In particular, the FAQ document answers providers’ concerns related to:

- the CENDARI users
- the benefit that cultural heritage institutions can have by sharing their content through CENDARI
- preferred data formats for the CENDARI project
- the type of content that CENDARI is preferably aggregating
- data licensing conditions
- technical standards and software in use by the CENDARI project

This information was made available both as a web resource (http://www.cendari.eu/research/libraries-archives-2/faqs-for-cultural-heritage-institutions/) and as an attractive printed flyer.

CENDARI Technical Checklist

This document serves multiple purposes: it is a communication tool ensuring that both the Data Provider and the CENDARI project understand clearly what the exchange will involve and require; it is a reference for the Data Provider in relation to the recommended

data formats and data delivery methods; and it is an internal communication tool for the CENDARI team to ensure collection and technical specialists share enough and correct information about the details of a forthcoming data exchange (e.g. the language of the described records/collections; metadata format; data provision method; type of supported database; additional documentation and any requested or required technical assistance). It has been an interesting experience that this technical checklist was used only rarely as in the example of the American JDC archives which is described below.

CENDARI Workflow Model

This internal document established a number of common models for cultural heritage institutions and clarified what would be expected and/or sought from engagement with them. It enabled the CENDARI team to understand clearly what questions to ask of potential data providers, and how to use that initial questioning to communicate clearly (internally and externally) regarding what their collaboration with the CENDARI project might result in.

JIRA

Communications with institutions often lasted many months between first contact and final conclusion. In this time, memories of exact agreements of next steps sometimes faded, and indeed members of the CENDARI team changed from time to time.

To support the contact with the archives, and keep the state of communication up to date, the JIRA Issue tracker service provided by DARIAH was used. As JIRA is a common bug-reporting tool for technical developers, a separate project, customized to fit the Data Acquisition Workflow was created.

Whenever a new archive (or data provider) was considered as being of interest for CENDARI, researchers would create a new JIRA issue for that archive (data provider). In cases where specific collections or multiple data interfaces were identified, the researchers created subtasks with more specific information.

JIRA entries were also used to track contact names, outcomes of discussions, dates of communication, name of team member last in contact with the archive, status of the data sharing agreement sign-up, the technical checklist and any other relevant technical information.

This workflow allowed the team to track the harvesting process and to share information between the CENDARI archival and the technical teams. As soon as data had been acquired from the providers, they were ingested into the CENDARI repository and thus made available for browsing by historians, archivists and all the users of the CENDARI Note Taking Environment.

BRINGING THE METADATA TOGETHER IN ONE REPOSITORY

It could be argued that CENDARI's federation of data represents a duplication, as some datasets exist both in the institutions' own repositories as well as in the CENDARI repository. The risk of this approach is that the data in the CENDARI repository could become obsolete, because they are "detached" from their original repository and are therefore not subject to updates from the original institution.

However, this risk is minimal, as archival descriptions do not get outdated very fast: once established in a comprehensive way and according to international standards, archives do not tend to substantially revise the digital descriptions of their holdings. This is especially true for older records and collections, since accruals (for example, administrative records) cannot be expected for this time period. An exception to this rule might be less formal collections, such as documents of the history of everyday life (e.g. "Alltagsgeschichte" in Germany), including items like war diaries and personal memories.

In order to manage the risk of data becoming obsolete, CENDARI has established a DARIAH Working Group dedicated to the project's sustainability, as well as a good practice toolkit that is available to cognate projects and to the larger digital humanities community. While the sustainability plan outlines maintenance of the final infrastructure, its knowledge, and communities from a variety of perspectives (like the Portal, Archival Research Guides, Tools etc.), sustainability of the data is one of the most complex and detailed sections of the plan, providing for a model in which the data environment can be secured but not only in a static format for the medium to long term. This includes the project Memorandum of Understanding with DARIAH (described above) as a central component, but by no means the only one.

If the risks of this approach are minimal, the potential benefits are great. The primary reasons for establishing what could be seen as just 'one more silo' are threefold:

1. Data silos, particularly in the form of national institutions, place limitations on historical research. In an age when more and more historians seek to take transnational approaches, the research community urgently requires resources that enable the visibility of research sources at a transnational level.

2. CENDARI has aggregated heterogeneous data formats (the so-called "Data soup" described in the next section), which represents an innovative way to harmonize the archival landscape in a virtual space. Using computational techniques to enable a more flexible knowledge representation and curation strategy, this allows reuse of the original digital assets from the archives as they are, without either making substantial additional requirements of them (related to data format, transformations, granularity, completeness etc.) or forcing historians to adapt their work habits to a virtual hangover from the fragmented analogue landscape of resource description standards (for archives, for library data, for textual data, for structured data, etc.). The CENDARI data model and the compilation of heterogeneous data in it presents a pragmatic solution in between the often-praised ideal of linked open data and the main emphasis being laid by cultural heritage institutions on collecting, preserving, ordering, administering and protecting. The pooling in a single repository allows for a balance between curated "golden data" and "big data" ingested from aggregators.

3. Finally, the ingestion of data into the CENDARI repository opens up the possibility to process and index them through data extraction, entity recognition, semantic enhancement and other transformations (for example, translation), and to build an unified knowledge base encompassing acquired resources, external (ontologies) and user generated knowledge. From the viewpoint of historians, this means that when working in the CENDARI Note Taking Environment, they will find a much more 'intelligent' search functionality, facilitating retrieval not only of all datasets and archival descriptions containing the name of, for example a certain person, but also other facts directly connected or inferred from that person. In terms of impact on research methodologies, this technological feature – as well as the fact that data from many cultural heritage institutions are brought together in one repository, indexed and made far faster and easier to search as a single body of data – represents a considerable asset.

In these ways, the CENDARI repository represents a unique response to the current structure of the field of cultural heritage institutions and their policies of data creation and curation. This field can be seen as being in the middle of a process of transculturation, and the collection of data from 'hidden archives' and created manually, from small archives which provided their descriptions in spreadsheet format and from big aggregators with advanced data exchange tools in place, bears witness to this process. The change in the culture in how cultural heritage institutions present their holdings and share data becomes just as obvious as the change in the techniques of how data are being made accessible and being explored.

SOURCES AND NATURE OF CENDARIS 'DATA SOUP'

In its initial user needs assessment, the CENDARI team determined that the data of interest to the future users are highly heterogeneous, representing a wide variety of sources, structures, media types, licences and granularity of described content. The decision to federate this content into a single repository required the project to adopt (and indeed construct) a system with the ability to manage such a variety of data types and standards, and it led to the coinage of the term 'data soup', a hearty mixture of objects, descriptions, sources and XML formats, database exports, PDF files and RDF-formatted data respectively. Some aspects of this mix are described in more detail in the following sections.

Data Providers

CENDARI has worked directly with a number of archives and institutions, large and small, to be able to represent their content in the CENDARI repository. This ingest was coordinated between the historical researchers and the technical experts within the project, including for example 500 collections from Italian archives, 2,116 collection from the Czech

Archives Administration (mapped to EAD from xml-files), as well as 25,000 files from the German Bundesarchiv in EAD format. Further ingest is expected before the end of the project from cooperating archival institutions in Poland, Italy, Belarus, Russia, Slovenia, the Slovak and the Czech Republic, from United Kingdom, Austrian provincial archives and the Swiss ICRC archives.

In order to promote the prominence of CENDARI's activities and support this process, the CENDARI project has also cooperated with other European projects who liaise with archives as well as other institutions in regard to the First World War: In the first place, with the project "1914-1918-online. International Encyclopaedia of the First World War", the project "Europeana collections 1914-1918," the "European Film Gateway" as well as The European Library and the Wellcome Trust.

Categories of Data and Ingestion Options

Harvested (aggregated) data

Collections descriptions encoded manually by the CENDARI team are compliant with Encoded Archival Description (EAD), the archival standard used in almost every European and also many non-European countries, which is approved by the International Council of Archives. Compliance with this standard is ensured by the CENDARI Archival Directory tool AtoM.

However, descriptions harvested from the data providers were sometimes based on other standards or formats not yet enclosed within EAD, like the Text Encoding Initiative (TEI) format, the Europeana Data Model (EDM), the Metadata Encoding and Transmission Standard (METS), the Metadata Object Description Schema (MODS), RDF-based formats etc. Furthermore, data heterogeneity was manifested in the languages and granularity of descriptions, their quality and licences for further reuse or processing.

Another way to understand the complexity of the aggregation process is from the perspective not of the standard used by the institution to describe their collections, but how advanced their technical systems were overall, and how CENDARI was able to respond in terms of a data acquisition, which can be largely divided into two major categories. The first one applied where an institution had an application programming interface (API) or other open mechanism for data sharing; the second applied when an institution offered data directly exported from its local repository. Within each of the two categories, the CENDARI team found that the data harvesting processes were not completely uniform but came in different flavours.

The types of APIs provided by institutions were found to differ significantly. They were either in accordance with a standard for metadata exchange (for example OAI-PMH) or developed as custom interfaces. While the latter usually offered richer search and filtering options for the selection of relevant data, each of them supported a custom set of functionalities and specialized operations. Therefore, the first step in data acquisition from institutions with custom APIs was an analysis and examination of the interface in order to understand how to filter out records relevant for the CENDARI case studies on World War One and Medieval Culture and how to extract descriptions from API responses. Documentation availability and quality enabled this type of analysis to be carried out efficiently and relatively rapidly. The CENDARI team noted that analysis and investigation of poorly documented APIs may be very time-consuming. In that case, establishing contact with institutions' technical support team and gaining their assistance was found to be of great value. On the other hand, the methods for data acquisition from institutions that follow the standardized provision protocol were uniform and once the procedure for data ingest was established, it could be easily reapplied for new institutions providing the same type of interface.

Both standard and institution-specific interfaces encountered by the CENDARI team were either open or closed. Open APIs are publicly available to anyone, usually after registration in order to obtain an authentication key. Access to closed APIs is more tightly controlled, more restricted, and requires the explicit approval of the institution. Regardless of the interface type and implementation specifics, it is a common practice that there is an upper limit for the number of calls an API would accept per client within a defined time window (e.g. the number of requests per minute or per day). These restrictions exist to prevent exploitation and system overload, however, in some cases they may significantly slow down the process of data harvesting. This problem usually can be solved if the API rate limits can be adjusted per client, but significant relaxations of the restrictions is often given only to partners, collaborating institutions, and, in general, clients whose work is recognized as significant, relevant and valuable.

There are many cases when it is not possible to harvest data from institutions via an interface. This may happen, for example, where no interface is implemented or there is an interface but it is private and cannot be accessed from the "outside". For these institutions, CENDARI offered several other possibilities for data exchange. Where the quantity and size of data allowed for it, content was sent directly via email to the relevant CENDARI contact person. In cases where this was not possible, or if the content-providing institution did not permit it, data was also uploaded using WebDAV to the dedicated CENDARI server. While the descriptions harvested via APIs were almost always in a standard metadata format, content received in the aforementioned way was sometimes quite diverse when it came to structure and format (e.g. spreadsheets, exports from local databases, PDF or Word files etc.). Extractions of actual descriptions from such content required additional pre-processing and transformations in order to ensure its usability in the CENDARI virtual research environment. Also, since such content was usually created as an export from the institutions' internal software system, descriptions received in this way, although they followed some metadata standard, tended to be very large in size. The direct ingest of such descriptions in the form that they were received in could make them hard to manipulate and they could not be shown to the end user in a suitable and elegant way. In addition, some data received in this way were not relevant for the CENDARI case studies of World War One and Medieval Culture. Thus, the CENDARI team applied methods of filtering and transformations in order to extract relevant, standalone pieces of information (e.g. item or file level descriptions) from very large and complex descriptions; partly in order to make them easier to handle, partly to extract only relevant content, while still preserving the information that was originally provided.

The data integration

Given the wide range of data sources and paths of entry into the CENDARI system, a flexible technical approach was called for. Instead of applying a classical integration approach and defining a common description format, the CENDARI repository was designed to allow for the 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 accepted by a repository. In addition, it requires substantial intellectual effort and technical work invested upfront in defining and implementing translation rules, such as whether the "common-denominator" or "union-all" principle is used. By contrast, the method adopted by the CENDARI project was 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 was not avoided. However, the upfront efforts were lower and the system allows for incremental integration over time.

Dataspaces – the top level organization for CENDARI Data

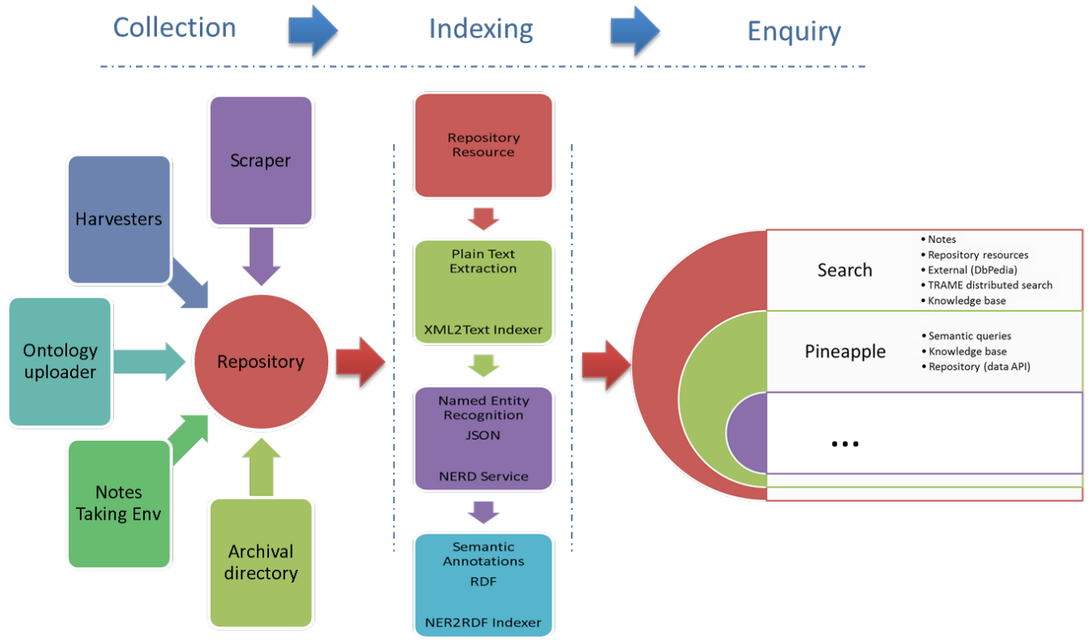
Data providers imposed different restrictions on data which they shared with CENDARI in terms of licence or allowed visibility for data. Taking the technical aspects into account, even in cases when various data providers shared their data through e.g. an OAI-PMH interface, there were various settings which had to be used by the harvesting component in order to harvest an exact selection of data or to produce the desired data structure in the CENDARI repository.

Furthermore, at the core of the whole integration endeavour was the data provenance. Since, for researchers, the CENDARI data soup should provide trusted and credible resources, it was and is necessary to keep provenance records from the moment data entered the CENDARI repository.

To enable support for these requirements, all data in CENDARI were organised into Dataspaces. A Dataspace holds access permissions and delineates data coming from various providers via data harvesting. Thus, for every unit of data the information about who is the "authority" behind it is well known.

Integration of Dataspaces

The following process was developed during the development of CENDARI. All data collected during the harvesting processes are sent to the appropriate dataspace in the CENDARI repository.



As soon as new data (or updated data) enters the repository, the process of data indexing starts for each new file created in the repository. This process is managed by the CENDARI data integration platform, which implements various indexers. Each indexer component decides if it should index the content or not, based on its definition and the format of the incoming data. The result of each processing step is saved and can be retrieved at any time via the CENDARI data API. Furthermore, it may itself become further input to another indexing step.

For example, when an EAD XML file is saved to the repository, the following transformations are performed:

- transform selected EAD fields from EAD XML file to a plain text file
- transform the selected EAD fields to an RDF file, in accordance with the CENDARI Archival Ontology (CAO), and send it to the CENDARI knowledge base (Virtuoso triple store)
- transform the EAD file to a JSON format, which in addition includes the plain text file content generated in the first step. This resulting JSON format is sent to the CENDARI Elastic search service, thus searchable through the CENDARI Faceted search tool along with other CENDARI contents.

In a parallel process, the plain text result is sent to the CENDARI NERD Service (Named Entity Recognition and Disambiguation Service), which returns annotations about the entities recognised from the provided text. These may be Persons, Dates, Organizations, Places etc. The NERD output is through an additional indexer transformed into CENDARI ontology format and sent to the CENDARI Knowledge base.

The CENDARI Data integration platform was built as an extensible and inclusive integration platform, where additional indexers can be invoked and defined if needed, for example to better process data in some special format, or from a particular dataspace. Thus, beyond feeding the CENDARI Faceted search service with data, and ensuring extension of the knowledge base, it additionally enables semantic enrichment and alignment of its data, thus building a rich enquiry environment for historical research.

BEST PRACTICES

Introduction

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 size, amount of digitised records and catalogues, and technical solutions has required great flexibility from the CENDARI team. On the other hand, the variety of cultural heritage institutions that the CENDARI team has been in contact with represents one of the biggest value-adding resources of the project, which aimed to guarantee a wide representation of institutions, according to their geographic coverage, type of collections preserved and data describing the analogue collections.

The following examples describe in more detail the broad range of processes that the CENDARI team has put in place to exchange data with cultural heritage institutions. The CENDARI team considers the cases chosen and described as best practices; they should thus serve as models harmonizing the needs of future infrastructure projects like CENDARI and those of cultural heritage institutions.

Pan-European Aggregators

Europeana and The European Library

The cooperation between Europeana, The European Library (TEL) and CENDARI shows how the project was able to engage with a digitally well-prepared pan-European aggregator. This collaboration has meant much more than a collaboration at the level of data exchange. The European Library (http://www.theeuropeanlibrary.org/) is indeed among the CENDARI project partners, and has been a key source of expertise concerning the aggregation of digital European library content and data/ontologies modelling. Europeana (http://www.europeana.eu/) – deeply connected to The European Library, both in its historical development and for its physical location – is the well-known European Digital Library, aggregating digitised cultural heritage and the related metadata from museums, archives, galleries and libraries (this last is aggregated via The European Library).

For the purpose of data ingestion from Europeana into the CENDARI repository, colleagues from Europeana compiled a list of 49 collections containing records related to World War One. Approximately 470,000 records were selected for ingestion to CENDARI from Europeana and The European Library. All of the records related to the period of World War One (1914-1918), and this timeframe provided for the main selection criterion.

Within Europeana, the selected content has been aggregated in the context of the following projects:

- *Europeana Newspapers:* The aim of the project is to create full-text versions of approximately 10 million newspaper pages. Additionally, it allows users to quickly search for specific articles, people and locations mentioned in the newspapers.

In the context of CENDARI, Europeana Newspapers has provided newspaper collections from the following institutions:

- Royal Library of the Netherlands
- National Library of Luxembourg
- National Library of Poland
- National and University Library of Slovenia
- University of Belgrade
- Tessman Library
- National Library of Serbia
- Hamburg State Library
- National Library of Latvia
- National Library of France
- National Library of Wales
- National Library of Estonia
- Staatsbibliothek zu Berlin – Preußischer Kulturbesitz
- Österreichische Nationalbibliothek
- National and University Library in Zagreb
- National and University Library of Iceland

- *Europeana Collections 1914-1918:* The aim of the project is to gather digital collections from National Libraries and other partners in eight countries that had a prominent role during World War One.

In the context of CENDARI, Europeana Collections 1914-1918 has provided relevant collections from the following institutions:

- Royal Library of Belgium
- National Library of Rome
- Istituto Centrale per il Catalogo Unico (Italy)
- Bibliotheque Nationale de France
- Austrian National Library
- Staatsbibliothek zu Berlin – Preußischer Kulturbesitz
- National and University Library of Strasbourg
- Contemporary International Documentation Library (France)
- National Library of Serbia
- Central National Library of Florence

- *European Film Gateway (EFG) 1914-1918:* was a digitisation project focusing on films and non-film material from and related to World War One. It started on 15 February 2012 and ran for two years until 15 February 2014.

In the context of CENDARI, Europeana Film Gateway 1914-1918 has provided relevant collections from the following institutions:

- Imperial War Museum (UK)
- Eye Film Institute (NL)
- Estonian Film Archive (EE)
- Bundesarchiv (DE)
- Danske Filminstitut (DK)
- Centre nationale du cinéma et de l'image animée (FR)
- Istituto Luce (IT)
- Vintage Films Ltd. (UK)
- Filmarchiv Austria (AT)
- MaNDA – Hungarian National Digital Archive and Film Institute (HU)
- Arhiva Naţională de Filme (Romania)
- Institut Valencia de L'Audiovisual (ES)
- Filmoteca Espanola (ES)
- Národní filmový archiv (CZ)
- Deutsches Filminstitut DIF (DE)
- Scottish Screen Archive at National LIbrary of Scotland (UK)
- Landesfilmsammlung BW (DE)
- Kansallinen Audiovisuaalinen Arkisto (FI)
- Cineteca del Friuli (IT)
- Museo Nazionale del Cinema (IT)
- Fondazione Cineteca Italiana (IT)
- Österreichisches Filmmuseum (AT)
- Jugoslovenska Kinoteka
- Deutsche Kinemathek (DE)
- Filmoteka Narodowa (PL)
- Cinemathèque Royale Belgique
- Cineteca del Comune di Bologna

- *Europeana Photography* is a three year project (February 2012 to February 2015) uniting some of the most prestigious photographic collections from archives, public libraries, museums and photo-agencies covering 100 years of photography – from 1839 to the beginning of the Second World War in 1939.

In the context of CENDARI, Europeana Photography has provided relevant collections from the following institutions:

- Parisienne de Photographie (FR)
- Topfoto.co.uk (UK)

- Apart from the aforementioned Europeana Projects, a number of datasets relevant for the study of World War One have been aggregated by Europeana from single cultural heritage institutions.

A number of datasets have been harvested from Europeana and The European Library by CENDARI, and they originate from the following institutions and projects:

- Zielonogórska Biblioteka Cyfrowa (PL)
- Jewish Museum London
- Erfgoedplus.be (BE)
- Culture Grid (UK)
- Public Library Varna (Bulgaria)
- The National Library of Scotland (UK)
- The National Library of the Netherlands (NL)
- Stadtgeschichtliches Museum Leipzig (DE)
- National Library of Denmark (DK)
- National Library of Rome
- Manchester Regiment (UK)
- Trinity College Dublin (IE)
- Ghent University Library
- Russian State Library
- Nationaal Archief (NL)
- DDB Deutsche Digitale Bibliothek (DE)
- International Institute of Social History (NL)
- The Great War Archive (UK)
- Lithuanian State Historical Archives

Europeana and The European Library model their data according to the Europeana Data Model (EDM), a semantic data model developed by the Europeana R&D team. EDM allows for excellent interoperability and facilitates the inclusion of a great level of detail. At the same time it is a "malleable" standard, well re-adaptable for mappings by libraries, archives and museums. Europeana and The European Library support the retrieval of EDM records where each record corresponds to one cultural heritage institution. In addition to EDM format, The European Library offers records in Dublin Core as well as Linked Open Data (LOD) versions of the records.

Data aggregated from Europeana were harvested by CENDARI via the Europeana REST API (http://labs.europeana.eu/api/introduction) in an automated fashion using a dedicated script. The harvesting process for each collection began with the gathering of unique identifiers of the records from the selected collection that was to be ingested. Since a collection may contain not only records related to World War One, but may also cover other subjects, it was necessary to apply an additional filter in order to get relevant records only. For this purpose the subject filter *"world war I" OR "1914-1918"* was added to the search query. After the relevant records had been identified, the corresponding metadata were fetched via the API and uploaded to the CENDARI repository.

The Europeana API is well documented and easy to use. Thanks to the Europeana API technical team there is a responsive forum for sharing ideas and suggestions, solving issues and answering questions related to the API. It should be noted that there is a standard request limit of 10,000 API calls per day, but it can be significantly increased for a single API key in specific cases. It may also be worth noting and of interest for other API users that, although it is not currently specified in the Europeana API documentation, besides the JSON equivalent of an EDM, the original EDM record can be obtained as well.

The European Library OpenSearch API (http://www.theeuropeanlibrary.org/confluence/download/attachments/8880494/TheEuropeanLibrary_API_V2+0.pdf) provides access to metadata from around 3.5 million newspaper issues and bibliographic records from around 90 million catalogue records from national and research libraries in 48 countries. 17 newspaper collections covering the World War One period were identified and selected for acquisition by the CENDARI project. The selection was based on the specific time range, since subject based filtering is not supported for the newspapers data set. The harvesting process went similarly to the previously described ingestion from Europeana: after the identification of records that belonged to the selected collections, the corresponding metadata in Dublin Core, EDM and LOD were ingested by CENDARI.

National Aggregators

Archives Hub UK

The collaboration with Archives Hub UK is a typical example of data exchange with a national aggregator which provides an up-to-date technical infrastructure. Archives Hub UK is an online portal to 220 UK archives that generally do not fall under the purview of the National Archives or other state archives. It aggregates content from a range of institutions, including universities, firms, and non-profit organizations. A CENDARI team member got in contact with Archives Hub UK, represented by its service manager. Based on this personal contact and shared approaches toward interoperability and open access, the Archives Hub agreed to share metadata records relevant to World War One. A first obstacle to transferring data to CENDARI was the lack of an agreement on open access among Archives Hub's content providers. This was resolved with content providers agreeing to an open licence (CC BY). Archives Hub UK offers several interfaces for metadata sharing: OAI-PMH, SRU and Z39.50. Since OAI-PMH interface was not fully implemented and functional, SRU was the most natural choice since the CENDARI team was already familiar with it and had good experience in data harvesting via SRU from other providers. Records related to World War One were selected based on subject criteria "First World War" OR "1914-1918". The search query resulted in 1,193 collection descriptions. Archives Hub UK provides records in EAD and Dublin Core. For ingestion, EAD records were selected by the CENDARI team as they were much richer.

National Archives/ Libraries

Czech Department of Archives Administration

Data exchange with the Czech Department of Archives Administration serves as an example of a one-off data acquisition based on personal negotiations. In March 2015, the central Czech Archives Administration provided CENDARI with a one-time data dump of 2,116

XML metadata records of fonds and collections from archives across the Czech Republic relevant to World War One. The records were from the central state authorities, the political and financial administration, and the military – especially the Austro-Hungarian army, the Czechoslovak legions, and selected important individuals in the Czechoslovak resistance. This major data ingest was the outcome of several months of collaboration between the CENDARI project and the Ministry of the Interior of the Czech Republic, which is responsible for state archives.

The Archives Administration was not interested in a sustained collaboration because it would require time and resources unavailable to the agency. Yet its representatives did immediately promise a one-off data dump, in the interest of advancing scholarly research and gaining exposure for Czech archival materials. Although there was a delay in receiving this data since the Archives Administration needed to complete its inventory of the Czech archival holdings, in the end a package of metadata records was prepared for CENDARI. The selection was made simply by aggregating all records from the chosen types of archives (administrative and military) that fell within the period 1914-1918. The XML schema used by the Czech Archives Administration did not map directly onto EAD, so the CENDARI team translated the elements used into those that appear in the CENDARI EAD schema. Because the Czech data providers were interested in publicising the collaboration with CENDARI and wanted something to exhibit, it was decided to make the Czech records visible and searchable in the AtoM repository.

Istituto Centrale per gli Archivi

The case of the Istituto Centrale per gli Archivi (the Italian Central Institute for Archives) serves as an example for an institution providing Linked Open Data (LOD) available for reuse. The Istituto Centrale per gli Archivi (ICAR) is the instrument for the study and the applied research of the General Direction of the Archives ("Direzione Generale degli Archivi"). It is responsible for the management, maintenance and development of the Italian archival Information Technology system. In collaboration with the SAN (Sistema Archivistico Nazionale, the National Archival System), ICAR has established an experimental service by transforming and making accessible to the public Linked Open Data, corresponding to 128 datasets and more than five million RDF triplets.

The process of accessing the datasets was straightforward for the CENDARI team. The data are available via the link http://dati.san.beniculturali.it/dataset/ and are browsable according to a broad range of criteria. With the help of the CENDARI tech team, the SPARQL endpoint of the ICAR database was interrogated with regard to records containing the keywords "Prima Guerra" or "Grande Guerra". From this refinement 444 files were obtained, each corresponding to a collection, which were then saved in XML-RDF.

The LOD initiative of the Istituto Centrale degli Archivi represents an invaluable opportunity for the reuse of such archival data by researchers and third party aggregators. It allows the Italian Central Institute for Archives to act as an example to similar institutions which still struggle when dealing with actions related to openness and data interoperability.

In the Italian case, the LOD initiative draws on the guidelines for the "Valorisation of the Public Digitized Heritage" ("Valorizzazione del Patrimonio Informativo Pubblico", http://www.agid.gov.it/sites/default/files/linee_guida/patrimoniopubblicolg2014_v0.7finale.pdf) released by the Agency for Digital Italy ("Agenzia per l'Italia Digitale") by the Presidency of the Council of the Ministers ("Presidenza del Consiglio dei Ministri").

The German Bundesarchiv

The German Bundesarchiv (German Federal Archives) provides for an example where users do not need deeper technical knowledge to receive access to its archival descriptions and finding aids. The German Bundesarchiv has the legal responsibility of permanently preserving the German federal archival documents and making them available for use. This includes documents (files, papers, cartographic records, pictures, posters, films, sound recordings and machine-readable data) arising from the central institutions of the Holy Roman Empire (1495-1806), the German Confederation (1815-1866), the German Reich (1867/71-1945), the occupation zones (1945-1949), the German Democratic Republic (1949-1990) and the Federal Republic of Germany (since 1949). In 2014, the German Bundesarchiv established a portal (https://www.ersterweltkrieg.bundesarchiv.de/) presenting its records on World War One. A part of these records have been digitised. Upon request by the CENDARI project, the German Bundesarchiv pointed towards their open data repository (https://open-data.bundesarchiv.de/) which is freely accessible, and where a good part of the archival descriptions of the holdings of the Bundesarchiv are available in EAD format. The files stored there are being continually updated and complemented, and they are arranged according to their signatures. A researcher familiar with the signatures can simply download each file containing shorter or longer descriptions of each unit. The files available at the open data repository contain the indexing data; digitised objects are not enclosed. All the data available are free from legal restrictions (e.g. for personal data), and all data provided there can be reused. The German Bundesarchiv is being financed via public money; it is submitted to the German Federal Informationsfreiheitsgesetz which grants any person an unconditional legal right of access to official information by federal agencies.

From the open data repository, approximately 160 files were chosen. As a rule, every file contains the records of an institution of the German Reich or for military units such as infantry regiments. If these records were relevant for the history of World War One, they were chosen and ingested as a whole into the CENDARI data repository. If they contained, for the most part, records not relevant for World War One, as in the case of the fond of the German Foreign Office which spans from 1867 to 1945, records which had been produced between 1914 and 1921 (i.e. including the peace negotiations led by the German Foreign Office) were filtered and each put into a separate file in the CENDARI repository, all of them containing a link to the research system of the Bundesarchiv which is called 'Invenio'. Thus around 25,000 item descriptions from the German Bundesarchiv have been incorporated into CENDARI. This way the German Bundesarchiv enables users to evaluate, present and enrich the available data in multiple ways.

The National Archives of Estonia (Rahvusarhiiv)

The case of the National Archives of Estonia serves as a bright example for a small archive which provides comprehensive access to its archival descriptions and finding aids as well as to its digitised content. The National Archives of Estonia preserve the administrative records of the Republic of Estonia established in early 1918. However, the archival holdings include large collections on World War One, the emergence of the new Estonian state, records on the Russian Revolution, and on the war with the Bolsheviks, which are mostly in Estonian. There is a large number of mostly untouched collections in German and Russian, including records of the "Baltic Regiment" and diaries of German soldiers. Finding aids are

The CENDARI White Book of Archives

online (none of them treated with Optical Character Recognition) as well as a substantial part of the archival documents themselves.

CENDARI team members communicated with the deputy director of the Digital Archives of the National Archives of Estonia. The National Archives of Estonia sent an English overview of the fonds in their holdings relevant to World War One. The archival descriptions had been established by an Estonian archivist and were manually added to the CENDARI repository.

Since December 2014 the National Archives of Estonia have provided all archival descriptions as open data available in their open data repository; all information presented there is in Estonian (http://opendata.ra.ee/). The National Archives of Estonia do not see the need for a Data Exchange Agreement as in general archival descriptions are under the CC0 licence; users are free to use these.

All descriptions are presented in two formats:

- RDF: Each zip file contains all the RDF descriptions for one archival fond. The title of the zip file accords to the fond number; after the fonds of interest have been chosen, they can be downloaded
- apeEAD: In the section "Koondfailid" in the lower-left corner the EAD files are available as one zip file per archival institution

There are several zipped folders containing the files of Estonian archives:

- Ajalooarhiiv (Historical Archive) EAA.zip
- Riigiarhiiv (The National Archives) ERA.zip
- Filmiarhiiv (The Film Archive) EFA.zip
- Maa-arhiivid (Land or regional Archive)s MA.zip
- Tallinna Linnaarhiiv (Tallinn City Archives) TLA.zip

Both sets also include references to digitised images in case they are available. The licence for the images is CC-BY-SA if there is the need to reuse them. In terms of digitised content, users can have a look at the SAAGA website (http://www.ra.ee/dgs/explorer.php) containing the digitised archival sources of the Raahvusarhiiv. This environment provides a topic-based entry for most of the digitised content. They also include data about 1914-1920 (e.g. lists of recruits until 1917, documents of Estonian government offices starting from 1916, military records 1917-1940 etc.). If a user wants to view some images, a registration is required but one can also log in through Facebook, LinkedIn, etc. The links provided in the EAD or RDF files also point to this environment.

Local Archives/ Libraries

Museo Storico Italiano della Guerra di Rovereto

The case of the Museo Storico Italiano della Guerra di Rovereto shows how the CENDARI project engaged with "hidden" archives and their collections. The historical archive of the Museo Storico is the repository of collections (manuscripts, documents, maps and audio-visual recordings) related to the history of wars, from the modern era until the 21st century. In the last twenty years the archive increased its collaboration with the Museo Storico

in Trento with the aim of preserving the manuscripts and documents of popular writings of national and local interest. Despite having a rich collection of diaries and letters (most of them from the period of World War One), the archive of the museum in Rovereto had a very basic (and lacking) cataloguing of this collection. The chief archivist was contacted in order to get acquainted with the collection named "Diari e Memorie", and to learn whether he thought there would be certain diaries worthwhile to be further analysed by CENDARI.

The chief archivist promptly pointed to the "Luigi Speranza Collection": according to him, this collection hasn't been given enough attention yet and was a good candidate for an in-depth research work to be included in the CENDARI Archival Research Guide on Private Memories of the First World War. Luigi Speranza was a craftsman from the little town of Lavis, in the province of Trento, Italy. On the first of January, 1915, Speranza left his village to work as a militarised workman for the "Genio Militare" (a military corps that created the infrastructures necessary for the war, such as roads, buildings, trenches) for the Regio Esercito Italiano (the Royal Italian Army).

The Speranza collection comprises four diaries for which an EAD description has been created manually in the CENDARI Archival Directory AtoM (https://archives.cendari.dariah.eu/index.php/diari-di-luigi-speranza). Particular attention was given to references of places, dates and events, and they were highlighted in the Archival Research Guide as entities: those entities will then be visible in the visualization section of the Note Taking Environment, in maps and dates-histograms. Noting down places and dates in Speranza's way through the war was very important for the work in CENDARI as it will allow historians to follow the movements of militaries and ordinary people in the north east of Italy, an area on the border between Italy and the Austria-Hungary empire, where crucial battles were fought during World War One.

In addition an interview was set up with the chief archivist at the Museo Italiano della Guerra, focusing on the importance of including personal memories in historical research. The interview can be found within the Archival Research Guide on Private Memories at https://notes.cendari.dariah.eu/cendari/ARG_Private_Memories_of_the_First_World_War/notes/770/.

American JDC archives

The exchange with the American Jewish Joint Distribution Committee (JDC) Archives serves an example of how data acquisition with respect to a highly specialized research question has been undertaken. The American JDC archives were contacted by the CENDARI team as an important institutional archive with rich and unique collections. The archive itself is located in two centres, one at the headquarters of the organisation in New York and the second in Jerusalem. The Archives of the New York office preserves one of the most significant collections in the world for the study of modern Jewish history. Among its holdings the records of the New York office for the period 1914-1921 are of especially great importance. The rich data collected by the JDC in the crucial years of World War One and afterwards highlight the life and history of Jewish communities all over Eastern Europe. Among the rich collections are the organizational records of JDC on its activities in different places where Jewish population suffered from the war, deportations, diseases and pogroms.

The JDC archives have shared their data on file level for the 1914-1918 collection with the Europeana project. Nevertheless, the CENDARI team decided that it would be more useful not to harvest JDC archives data straight from Europeana in EDM format, but to receive the data directly from the Archives in EAD format, which is an international standard format and provides for richer archival data than EDM.

The Archives of the American Jewish Joint Distribution Committee contributed file-level metadata for the World War One collections (1914-1918 and 1919-1921) to the CENDARI project. After the formal agreement had been signed by both parties, data were transferred by uploading the EAD files to the CENDARI project via WebDAV. After being uploaded to the CENDARI repository the files were additionally processed by the CENDARI EAD indexer service. This service extracts different levels within an archival inventory (collection, subcollection, item, file, etc.) and creates corresponding semantic connections between them, thus allowing for effective search within an archival inventory provided in EAD format.

In supporting the CENDARI project, the JDC archives aim to spread information about their rich holdings in the research community and to scholars who express interest in the history of World War One.

CONCLUSION

In exchanging data with cultural heritage institutions, the CENDARI project decided to set up a repository and to store the data collected there. Such a solution is not mandatory; another possible solution would be that cultural heritage institutions set up their own open data repositories which can be accessed by an interface (API). This would have an advantage for cultural heritage institutions in that they would be able to have their data updated and be in control of who has access; for infrastructure projects there would be no need to set up their own repository. A lack of resources on the side of cultural heritage institutions, which may result in the incapability to maintain such a repository could be compensated for by a federated solution, such as the national archives portals which are in existence in Germany and Austria. These archives portals would manage the technical as well as the legal aspects on behalf of the individual institution.

Data can be regarded as the gold of the 21st century. Digital descriptions of archival collections – though they may seem to be 'only' of historical value and therefore of limited interest – open up a range of interesting possibilities: If, for example, an archival description contains information on the correspondence of one person with others, this information can be extracted and brought together with the archival descriptions of the correspondent's partners, and social network graphs can be created out of these data. This is a typical example of the reuse of data with digital methods: The networks created on the basis of the information provided in the archival description goes far beyond what an individual researcher or even a research group would be able to investigate; thus interpersonal networks become visible as a yet under-researched resource, compared to the history of institutions or companies that have often been the object of investigation by

The CENDARI White Book of Archives

historians. In the U.S., the project “Social Network and Archival Context (SNAC)” (http://socialarchive.iath.virginia.edu/) pursues such an approach as well as the German Kalliope Union Catalogue.

Accustomed to the classical task of storing, preserving, ordering and administering cultural heritage, and bound to the physical and local presence of the historical material, archives, museums and libraries now open up the visibility of their treasures by exposing digital descriptions of these objects or even digitisations of the objects themselves. This process has only begun in the recent past, and libraries and museums are much more advanced in sharing and presenting their holdings to the broader public than archives. Sooner or later this process will bring a differentiation, interoperability and standardization of archival information as well as a comprehensive coverage of information on what is actually stored in these institutions. But currently the limitations in regard to the availability of resources, be it manpower, technical knowledge, and proficiency, or be it in technical terms (formats, APIs, repositories) can be perceived often and nearly everywhere.

It is advisable for future infrastructure projects to win cultural heritage institutions as well as national or transnational aggregators as partners from the stage of the first draft of the project, and to provide for funding for them in order to relieve them from the most urgent scarcity of resources. Designed in this way, future projects could unlock synergies and underline the benefit which both sides could draw from such a cooperation: Legal and technical advice for cultural heritage institutions on how best to present their holdings, shared data and common development of ideas enhancing reuse of data and inspiring research for the infrastructure project. For an effective workflow and a well-functioning division of labour within the project, ample documentation in contemporary tools like JIRA and Confluence is essential. These core points can be seen as ‘lessons learned’ by the CENDARI project in regard to data management. Nevertheless, one has always to take into account the current realities and its pitfalls, which may be regarded as a phase of transition. The diverse data creation or acquisition strategies, the multiple formats and technical solutions, the pros and cons in regard to the establishment of a repository which have been described extensively above – all these factors bear witness to the fact that the archives and museums, more than the libraries, are in the middle of a process of adaptation to the demands of the 21st century.

In light of the experiences made, the CENDARI project recommends that cultural heritage institutions use formats conforming to international standards as well as an appropriate registration software supporting these standard formats while establishing descriptions of their holdings and records. This provides for the basis for future data reuse, but also the reuse of further data processing and transformations. Researchers and infrastructure projects like CENDARI will be able to build upon this basis.

We can expect that in the upcoming years more archival institutions, especially in Eastern Europe, will provide open access to the description of their sources. The significance of the CENDARI project lies with the experience already gained, and it could be of great value for cultural heritage institutions. Such projects and the visible results will encourage different institutions to become more actively involved in international or regional cooperation using modern technologies and new approaches.

APPENDIX

CENDARI/DARIAH Data Exchange Agreement

Section 1: Parties to the Agreement

Name of Organisation:	DARIAH
Address:	DARIAH-EU Coordination Office Göttingen Centre for the Digital Humanities (GCDH) Papendiek 16 - Heyne Haus 37073 Göttingen Germany
Phone:	+49 (0) 551 39 20476
URL:	www.dariah.eu
Name of the authorized person:	Laurent Romary
Title/role in the organization:	Director
Work Phone:	+49 (0) 551 39 20476
Work email:	laurent.romary@inria.fr

Hereinafter known as the "**CENDARI/DARIAH**"

And

Name of Organisation:
Address:
Phone:
URL:
Name of the authorized person:
Title/role in the organization:
Work Phone:
Work email:
Hereinafter known as the "**Data Provider**"

ARTICLE 1. DEFINITIONS

CC-BY, Creative Commons Attribution 3.0: Refers to the licensing framework as published at: http://creativecommons.org/licenses/by/3.0/. This license allows the sharing and reuse of work so long as it is attributed to the author or licensor in the manner s/he specifies.
CC-0, Universal Public Domain Dedication 1.0: Refers to the licensing framework published at: http://creativecommons.org/publicdomain/zero/1.0/. This license allows the sharing and reuse of work without attribution or permission.
CENDARI is the **Collaborative European Digital Archive Infrastructure**, a fixed-term research infrastructure project with the goal of integrating digital archival resources for medieval and modern history. Although it will create a digital environment for historical research, the emphasis in this is on the federation of existing content and the enhancement of it with digital tools. It is therefore distinct from a digital library project, as its goals are to enhance, rather than create and sustain, digital resources.
CENDARI/DARIAH refers to the combined goals and acceptance of responsibility for digital content in the CENDARI digital resource, as per the MOU of August 2013 (see Annex 4).
Content: A physical or digital object, typically held by the Data Provider or by a provider of the Data Provider.
DARIAH, the **Digital Research Infrastructure for the Arts and Humanities**, brings together the state-of-the-art in digital arts and humanities activities of its member countries and works with research communities via a network of affiliated projects. To coordinate these activities and help secure long-term sustainability for digital arts and humanities research in Europe, DARIAH is being established as a European legal entity or ERIC (European Research Infrastructure Consortium). Because of the unique relationship between CENDARI and DARIAH, any long-term commitments entered into by CENDARI in terms of content management and agreed usage restrictions will be underwritten by DARIAH, allowing these agreements to have a potential duration beyond the temporal scope of CENDARI (which will complete its work in 2016). This relationship has been codified in the MOU to the effect, signed between DARIAH and CENDARI in August 2013 (See Annex 4).
Data: One or more interconnected digital objects.
Data Provider: Cultural Heritage Institution – archive, museum or library – which provides digital content in the form of standardized or unstandardized Data to CENDARI/DARIAH.
Digital Object: a single electronic file from the Archive and Library domain (e.g. image, Data set, audio-visual or audio resource).
Intellectual Property Rights: Intellectual property rights include, but are not limited to, copyrights, related (or neighbouring) rights and Database rights.
Metadata: Textual information (including hyperlinks) that may serve to identify, discover, interpret and/or manage Content. In most cases, metadata will constitute an instance of standardized Data (see definition below).
Third Party: Any natural or legal person who is not party to this Agreement.
Standardized Data: Data provided to CENDARI/DARIAH in a common metadata format and encoding used in digital archives and libraries, both human and machine readable. Examples of standardized Data formats are: EAD, MARC, ESE, METS, RDF, et al. Examples of standardized Data document encoding/file formats are XML or CSV.
Unstandardized Data: Data provided to CENDARI/DARIAH in non-standardized formats or encoding for digital archives and library environments, which are therefore not readily interoperable with other Data in the system. (e.g. .doc files, .pdf, excel files, image files and image previews, etc.).

ARTICLE 2. PROVISION OF DATA TO CENDARI/DARIAH[1]

1. The Data Provider shall decide in consultation with the CENDARI/DARIAH staff which content from within their digital holdings is appropriate for release through CENDARI/DARIAH. This may include partial content related to some collections.[2]
2. The Data Provider must make best efforts to provide CENDARI/DARIAH with correct information on the intellectual property rights of the digital content including the identification of the digital content that is within the public domain as having this status.

1 For the definition of standardized and unstandardized Data please refer to Article 1, Definitions.
2 For a list of the CENDARI preferred Data format and delivery methods, please see annex 1, 2 and 3.

3. In as far as the Data Provider has provided or will provide CENDARI/DARIAH with standardized or unstandardized Data that it has aggregated from Third Parties or that otherwise originate from Third Parties, the Data Provider shall ensure that these Third Parties have authorized the Data Provider to authorize CENDARI/DARIAH to make that Data publicly available in accordance with paragraph 2.2 of this article.

4. The Data Provider may request the correction, update or removal of their Data from the CENDARI/DARIAH digital environment and repository. Removal requests will be honored within 30 (thirty) days. Other required adjustments may take longer to deliver depending upon their complexity, but CENDARI/DARIAH will undertake to contact the Data Provider to agree a specific time frame and course of action within 30 (thirty) days.

ARTICLE 3. USE OF DATA

5. Under the condition that the requirements of Article 2 are met, CENDARI/DARIAH shall include the Data provided by the Data Provider in the repository held by CENDARI/DARIAH and shall make these available as a part of its digital environment. It does not undertake to provide any bespoke or unique method of access for any individual collection or for any Data Providers' Data.

6. Under the condition that the requirements of Article 2 are met, CENDARI/DARIAH will normally make publicly available all Data provided by the Data Provider under the terms of the Creative Commons CC-BY 3.0 license agreement, and is hereby authorized by the Data provider to do so. Where this is not a reasonable license for the Data shared with CENDARI/DARIAH, the Data Provider may request that CENDARI/DARIAH apply an alternate licensing framework to all or some of the Data provided. Exact specifications of alternate rights arrangements for specific content are specified in the Collection level information that appears in Annex 1 to this document. This information will be associated with the provided Data in the CENDARI/DARIAH repository and will be marked appropriately in the CENDARI/DARIAH digital environment.

7. Where appropriate, Data Providers may agree to provide their Data to CENDARI/DARIAH under a CC-0 Universal 1.0 - Public Domain Dedication, either because the Data are dedicated to the public domain or to provide interoperability with the Europeana Licensing Framework.

8. When making available Data or any parts thereof under the terms of the CC BY 3.0, CENDARI/DARIAH will provide a standard reference and link to the CENDARI/DARIAH Data Use Guidelines with the CC-BY 3.0 Attribution License guidelines.

9. When CENDARI/DARIAH publishes on the CENDARI/DARIAH digital environment Data that can be (in whole or in part) attributed to the Data Provider, CENDARI/DARIAH is obliged to give attribution to the Data Provider and to the party or parties referred to by the Data Provider.

10. In the event that CENDARI/DARIAH publishes a translation or transcription or any human transformation (including user annotations and saved references) of Data provided

by the Data Provider, CENDARI/DARIAH shall identify the translation or transcription as such. Provided Data may also be subjected to active forms of semantic enrichment. Where such transformation have been made public, the results of any such transformations will be offered to the Data Provider for their own use as well as appearing within CENDARI/DARIAH's digital environment.

11. CENDARI/DARIAH will clearly identify the rights framework under which Data within its environment is to be used, and inform users of their rights and responsibilities. CENDARI/DARIAH cannot be held responsible, however, for how these Data are been used by third parties outside of the CENDARI system.

12. The Data Provider recognizes its function to support academics and researchers to annotate (e.g. comment, explanation) the Data and the digital objects and to save a reference in the personal virtual research space.

ARTICLE 4. TERM

13. This agreement enters into force as of the date of the signature of the parties.

14. The agreement shall end on the 31st December following the Effective Date. The Agreement will be renewed automatically for a period of one year every 1st January, unless terminated by one of the parties, by written notice received by the other party ultimately on 30 September of that year. It is the intention of the CENDARI project and the DARIAH ERIC that all CENDARI activities, resources and agreements will be turned over for management from CENDARI to DARIAH on or before 1 February 2016. Should this migration cause any changes in the terms of this agreement, the Data Provider signing this agreement will be made aware of these changes and any implications for its Data available through CENDARI/DARIAH by 1st December 2015.

ARTICLE 5. LIABILITY AND NOTICE OF TAKE DOWN

15. The Data provider must make best efforts to ensure that performance by CENDARI/DARIAH of articles 2, 3 and 4 do not constitute an unlawful act towards a third party, including but not limited to:

a. A violation of Intellectual Property Rights of a Third Party
b. An infringement of personality, privacy, publicity or other rights; or
c. An infringement of public order or morality (hate speech, obscenity, etc.)

16. In the event that performance by CENDARI/DARIAH of articles 2 and 3 constitutes and unlawful act towards a Third Party, CENDARI/DARIAH shall assist the Data Provider in limiting the negative consequences of such unlawful act, however without accepting any liability. In the performance of this obligation, CENDARI/DARIAH shall use the notice and take down procedure described in paragraph 3 of this article.

17. In the event that a Data Provider or a Third Party notifies CENDARI/DARIAH that it is of the opinion that performance by CENDARI/DARIAH of articles 2, 3 and 4 constitutes an unlawful act towards any party, CENDARI/DARIAH shall within 5 working days decide

whether it considers the notice (i) void of grounds, (ii) readily awardable or (iii) subject to debate, and CENDARI/DARIAH shall perform the following:

a. In the event that CENDARI/DARIAH considers the notice void of grounds, it shall inform the notifying party accordingly.
b. In the event that CENDARI/DARIAH considers the notice readily awardable, it shall take all required measures to end the unlawful state. CENDARI/DARIAH shall inform both the notifying party and the Data Provider of its decision.
c. In the event that CENDARI/DARIAH considers the notice subject to debate, it shall inform the notifying party of this decision and allow the Data Provider to provide its views on the opinion within five (5) working days from the date that CENDARI/DARIAH has forwarded the opinion to the Data Provider. Upon receipt of the views of the Data Provider, CENDARI/DARIAH shall decide within five (5) working days whether measures are required to end an unlawful state. CENDARI/DARIAH may decide to request the notifying party and, subsequently, the Data Provider for further views.

18. Both parties shall hold the other party free and harmless of any action, recourse or claims made by any Third Party due to non-observance of its obligations under this agreement.

ARTICLE 6. TERMINATION

19. Either party may terminate this agreement at any time on the material breach or repeated other breaches by the other party of any obligation on its part under this agreement, by serving a written notice on the other party identifying the nature of the breach. The termination will become effective thirty (30) days after the receipt of the written notice, unless during the relevant period of thirty (30) days the defaulting party remedies the breach.

20. This agreement may be terminated by either party on written notice if the other party becomes insolvent or bankrupt, if the Data Provider's project ends or if the Data Provider withdraws or ceases operations. The termination will become effective thirty (30) days after the receipt of written notice.

21. Upon termination of this agreement, CENDARI/DARIAH shall only be obliged to remove the Data provided by the Data Provider if the Data Provider requests CENDARI/DARIAH to remove them. Removal shall happen no later than thirty (30) days after such a request has been received by CENDARI/DARIAH.

22. In case of withdrawal, any transformation or semantic enrichment applied to the original Data as result of the work of the CENDARI technical team or its users (as described in Article 3.6) will be maintained.

23. Termination of this agreement does not affect any prior valid agreement made by either party with Third Parties.

ARTICLE 7. MISCELLANOUS

24. If any term of this agreement is held by a court of competent jurisdiction to be invalid or unenforceable, then this agreement, including all of the remaining terms, will remain in full force and effect as if such invalid or unenforceable term had never been included.

25. This agreement may be supplemented, amended or modified only by the mutual agreement of the parties. Any modification proposed by CENDARI/DARIAH must be notified to the Data Provider in writing. The Data Provider shall be allowed at least two months from the date of reception of the notice to accept the new agreement. If the modifications are not accepted by the Data Provider in writing within the allowed period, the modifications are presumed to have been rejected. If the proposed modifications are rejected by the Data Provider, CENDARI/DARIAH has the right to terminate this agreement as of 31 December of any year, with a one month notice.

26. This agreement is drawn up in English, which language shall govern all documents, notices, meetings, arbitral proceedings and processes relative thereto.

27. All disputes arising out of or in connection with this agreement, which cannot be solved amicably, shall be referred to the conflict resolution process of the DARIAH ERIC for mediation. The outcome of the mediation process will be binding on the parties.

Signed by both parties:

Date:	Date:
Data Provider:	CENDARI/DARIAH

ANNEX 1: TEMPLATE OF THE CENDARI CHECKLIST FOR COLLABORATIONS WITH CONTENT HOLDING INSTITUTIONS.

This is an example of the checklist that one of our collaborators or researchers will ask you to fill in. Please note that you should fill in one form for every Dataset you will submit to CENDARI.

Basic Details

Institution	
Institutional Contact	
CENDARI Contact	

Content Details and Standards applied

Description of Fond (s) / Collection (s) of interest (title and signature/shelfmark)	
Approx. size	
Finding Aids (Digital/Analogue)	
Digital Objects?	

Archival Networks and Project

What networks are you active in (in particular digital libraries/archives like Europeana/ APEX)?	

Technical Details

OAI-PMH interface available?	
Link to Data provision interface	
Data format	
Cataloguing or other metaData standards applied (eg METS, MARC2)?	

Licensing Details

Other questions (Y/N)

May we include you on our public network map?	
Is there any other information we can provide you, or ways in which we might collaborate further?	

ANNEX 2. RECOMMENDED DATA STANDARDS

1. DC Dublin Core
2. ISAD International Standard
3. MARC Machine Readable Cataloguing
4. ESE Europeana Semantic Element
5. EDM Europeana Data Model
6. MODS Metadata Object Description Schema
7. EAD Encoded Archival Description
8. TEI Text Encoding Initiative
9. METS MetaData Encoding and Transmission Standard

ANNEX 3. RECOMMENDED DATA PROVISION METHODS

1. Harvest via OAI-PMH
2. Delivery via FTP
3. Delivery via API

ANNEX 4. DARIAH – EU AND CENDARI MEMORANDUM OF UNDERSTANDING (MOU)

DARIAH, the **Digital Research Infrastructure for the Arts and Humanities** aims to enhance and support digitally-enabled research and teaching across the humanities and arts. DARIAH- brings together the state-of-the-art digital arts and humanities activities of its member countries and works with research communities via a network of affiliated projects. To coordinate these activities and help secure long-term sustainability for digital arts and humanities research in Europe, DARIAH is being established as a European legal entity or ERIC (European Research Infrastructure Consortium).
Research projects within the humanities that have received national or European funding and whose work programme comprises of an important move towards using digital methods are a core stakeholder group for DARIAH activities. In addition to practice benefits such as access to the DARIAH technical environment (e.g. virtual machines, long-term archiving, single-sign on, collaboration space) and expertise in data modelling, standards for (meta-) data interoperability and virtual research environments, DARIAH is able to offer sustainability of research data, results and publications beyond the lifetime of the project.
CENDARI (Collaborative European Digital Archive Infrastructure) is a fixed-term research infrastructure project with the goal of integrating digital archival resources for medieval and modern history. The project brings information and computer scientists together with leading historians and existing historical research infrastructures (archives, libraries and other digital projects) to improve the conditions for historical scholarship in Europe through active reflection of and considered response to the impact of the digital age on scholarly and archival practice.
In order to fulfil its goals, CENDARI will be required to conclude basic agreements with content owning institutions (such as archives, libraries and museums) so as to be able to assure these bodies of the parameters CENDARI will apply for the responsible use of the data these institutions share with CENDARI. As CENDARI has a fixed term and no status as a legal entity, however, the project defers this responsibility to the DARIAH legal entity, which will sign on behalf of the project, and ensure that the terms of the agreements concluded will continue to be observed in the period after the CENDARI project ends in 2016.
The parameters of use will be agreed and encoded in the CENDARI data licensing agreement, which will clearly outline how the data will be managed, and what the roles of CENDARI, DARIAH and the content owner will be in this relationship.
With this Memorandum of Understanding we formally recognise CENDARI's status as a DARIAH Affiliated Project and thereby authorise DARIAH's legal entity, the DARIAH ERIC, to sign content-sharing agreements negotiated by CENDARI on behalf of the project.

CENDARI Frequently Asked Questions for Cultural Heritage Institutions

How is CENDARI different from other digital repositories?

CENDARI's purpose is to bring content together and make it more usable for advanced historical research. Our development emphases are on enhancing discovery, both of collections and of patterns within collections, rather than on creating and maintaining actual resources. CENDARI is not for the long-term collection and preservation of digitized content. CENDARI builds on the work already done by cultural heritage institutions to add value to already digital finding aids and assets, focussing also on the 'hidden' collections, which may be less well represented in the digital landscape. Unlike a typical digital repository, we don't have strict requirements for data ingestion. Cataloguing documents such as collection descriptions and finding aids are essential instruments for historical research. Therefore we are keen to provide tools for gathering, enhancing and sharing this documentation with the CENDARI end users, within a powerful enquiry environment.

Who are CENDARI's users?

CENDARI is a research infrastructure for medieval and modern historians. Although the final environment may be of interest to other user groups, our two case study areas for the period of 2012-2016 are medieval culture and the First World War. After the end of this period, we plan to hand over our work to DARIAH, where it may be accessible to future user groups for application in other areas of interests.

What are the benefits for your institution?

Cultural heritage institutions are the starting point for most historical research: while research practices are changing within the digital landscape, CENDARI aims to accommodate pre-existing research methodologies. In other words, we wish to enhance the existing relationship between archives, libraries and researchers by providing them with the tools to discover, organise and enrich their data. By sharing your data with the CENDARI project, you will join a network of European cultural heritage institutions whose content is connected within a powerful research infrastructure, making it more easily discoverable by historians and researchers. In particular, the benefits for your institution are:

- Visibility and searchability of collections
- Targeted access to the key user group (historians and collections experts)
- Data enrichment which will be shared back with your institution
- Transnational and multilingual functions, user annotations, data mining, data visualizations, connections with ontologies, etc.
- Inclusion in a community of scholarly practice developing technical tools and standards fit for use in cultural heritage collections
- Opportunities to test the technologies we are implementing and observe how they work with your content.
- Participatory design workshops that help define requirements for new tools and services

As members of our user group, you will also have access to training events and summer schools, which may be of general interest to you and your staff.

What format does my data need to have to be contributed to CENDARI?

CENDARI can accommodate the following data provision scenarios: If your data are digitally available in common meta-data standards for archives and libraries, we can harvest it remotely either via API, OAI/PMH, or direct deposit. If your data are not available in common metadata standards for archives and libraries, CENDARI can still integrate, and add value to, the collection descriptions, finding aids and other cataloguing material your institution can provide. Where the records of interest are not covered in existing finding aids, we can assist you with the creation of high level descriptions which will allow your content to be seen in the environment along- side other similar collections. We also have some capacity for technical consultancy, where this can be mutually of benefit.

What kind of content is Cendari looking for?

As the CENDARI pilot areas are medieval culture and First World War, we are looking for data and content related to these two area of studies. We highly value the data and content that are less visible in the digital environment. Examples of relevant data and content related to the First World War case study are:

- Military records
- Administrative records
- Personal collections, diaries, correspondences
- Photographs, films, posters, pamphlets

Examples of relevant data and content related to the medieval studies are:

- Manuscripts
- Incunabula
- Illuminations

What are the licensing conditions for the data shared with CENDARI?

CENDARI advocates CC-BY as a licence for use by historians and collections experts. We can, however, accommodate other licence formats and restrictions if required. The Data Sharing Agreement gives an over- view of the licensing framework in relation to our activities and target users. We are able to provide a full signed Data Licence Agreement if required. This agreement is underwritten by DARIAH, which has a permanent existence and can guarantee management of your content beyond the period of CENDARI's funded activity.

What technical standards and software is CENDARI using?

CENDARI takes a Service-Oriented approach meaning that software components can be tailored effectively for different use scenarios. The Virtual Research Environment (VRE) is being developed as a standards-oriented infrastructure and aims to: support the research processes of historians; foster collaborations; and assist researchers through visualisations and analysis of digital files. Our basic data model is based upon EAG (Institution level)/EAD (collection level)/MODS (item level). Resources will be linked with ontologies (containing rich domain knowledge), in order to provide a flexible, rich and multi-relational historical classification scheme. For data harvesting and storage there are several options,

like CKAN, allowing us to harvest resources via OAI-PMH repository and API; ingest various formats; internally organise datasets; connect datasets with the providing institution.

For data harvesting and data storage, we are looking at a number of possibilities, including CKAN, which allows to:

- harvest resources via OAI-PMH repository
- harvest resources via API
- ingest various data formats
- internally organise data and datasets
- record the dataset review history
- connect datasets with the providing Institution

We will also have a strong component of RDF linked data resources.

terface, there were various settings which had to be used by the harvesting component in order to harvest an exact selection of data or to produce the desired data structure in the CENDARI repository.

Furthermore, at the core of the whole integration endeavour was the data provenance. Since, for researchers, the CENDARI data soup should provide trusted and credible resources, it was and is necessary to keep provenance records from the moment data entered the CENDARI repository.

To enable support for these requirements, all data in CENDARI were organised into Dataspaces. A Dataspace holds access permissions and delineates data coming from various providers via data harvesting. Thus, for every unit of data the information about who is the "authority" behind it is well known.

Integration of Dataspaces

The following process was developed during the development of CENDARI. All data collected during the harvesting processes are sent to the appropriate dataspace in the CENDARI repository.

CENDARI Technical Checklist

Data Provider data description table

Reference Number:

Data Type (Finding Aid, Images, Full Text):

Short Description of Topic:

Licensing Alternative to CC-BY (if required):

Technical contact:

Institutional contact:

Cendari contact:

Approximate No of Records (or Size)	*If possible, provide approximate number of records at present, or the size of the data (if known)*	
Data available through aggregators	*Check if data is already provided to data aggregators like Europeana, TEL, ...*	• Europeana • TEL
Available metadata or container formats	*If known, provide information which metadata formats are used e.g. oai_dc, MARC21, MODS, DC, DCQ (Dublin Core Qualified), METS, TEI, or custom. List all which are used.Mark preferred formats for harvesting with "*". Provide additional comments if needed.*	

Content language	*Circle all languages applicable for the content.If not in the list, simply add it.*	• English • German • French • Italian • Polish • Serbian • Russian • Czech • Latin • Spanish • _____ • _____
Content encoding	*Circle the writing system in which the content is encoded. If not in the list, simply add it.*	• Latin • Cyrillic • _____ • _____
Data provision method	*Provide basic information how data will be provided to Cendari (circle option below, or comment how). Note that the method chosen also affects the frequency of updates.* *For 1) (provide here service URL and user credentials if needed). If possible, avoid 2).* *For methods 3) and 4) authenticated web interface will be provided for data upload.*	1. Data can be downloaded (or harvested) by Cendari via a web service (API). (provide here service URL and user credentials if needed) 2. Data will be sent to a Cendari technical contact 3. Data will be uploaded to a Cendari service in own format 4. Data will be uploaded to Cendari service in a Cendari format
Database	*Circle the storage/ database type (mostly applicable when method of data provision 2, 3 or 4 is used) used or name it if not in the list.*	• MySQL • SQLServer • PostgreSQL • Oracle • Filemaker • Access • ExistDB • _____ • _____

Documentation and assistance	*Circle option applicable for assistance to Cendari team during data acquisition and initial processing. Options 2 and 3 are very useful,so please try to agree on this whether or not other options are applicable.*	• technical contact person can assist Cendari team • structure schema can be provided • example annotated records can be provided • additional technical documentation available (please provide link)
Comments	*Comments or any other useful information not included in the table.*	

CENDARI Data Ingest Workflow