COLLABORATIVE EUROPEAN DIGITAL ARCHIVE INFRASTRUCTURE

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

Deliverable No. :	5.1
Title of Deliverable:	Archive Directory
Date of Posting to Basecamp/Confluence for Partner Review:	February 2014
Date of Deliverable:	August 2014
WP No.:	5
Lead Beneficiary:	FUB
Authors and Contributors (Name and email address):	Sheila Anderson sheila.anderson@kcl.ac.uk Jakub Benes j.benes@bham.ac.uk Pavlina Bobic pavlina.bobic@gmail.com Anna Bohn anna.bohn@fu-berlin.de Andrea Buchner a.buchner@bham.ac.uk Valentine Charles valentine.charles@europeana.eu Emiliano Degl'Innocenti emiliano@sismelfirenze.it Jonathan Gumz j.e.gumz@bham.ac.uk Oliver Janz oliver.janz@fu-berlin.de Milica Knezevic knezevic.milica@gmail.com Alexander Meyer alexander.meyer@inria.fr Francesca Morselli MorsellF@tcd.ie Aleksandra Pawliczek a.pawliczek@fu-berlin.de Klaus Richter k.richter@bham.ac.uk Zdenek Uhlir Zdenek.Uhlir@nkp.cz
Revision No.	1
Dissemination Level:	PU
Nature of Deliverable:	R = report
Abstract:	This report focuses on the development and creation of the CENDARI Archive Directory – Investigation and description of archives (Deliverable No 5.1). It explains the single steps which

	were necessary to establish a common workflow on the one hand and the methodological and substantial backbone of the displayed information on archives and holdings relevant for research on the First World War (WW1) and medieval manuscripts (MM) on the other. It also outlines the activities leading to building a network of cooperation with cultural heritage institutions and other national and international projects on digital representation of historical material referring to these two pilot topics. As described in the CENDARI Deliverable 5.1: Archive Directory, the project activities of WP5 consisted of researching relevant information for researchers working on topics of WW1 and MM, agreeing on formats and repositories to store the information, testing tools, contacting institutions to ensure their cooperation and to access their repositories, disseminating information on CENDARI in scholarly and archival circles and based thereupon, preparing the framework for the CENDARI Archival Research Guides (Deliverable 5.2).

Table of Contents

Introduction

The objectives of CENDARI Work Package 5 (WP5) focus on enhancing the accessibility and visibility of unique historical archives and collections across Europe through the creation of an electronic Directory of sources for medieval and modern history. It is also our aim to create research guides to pivotal and to less visible ("hidden") sources for medieval and modern history in order to improve their virtual and analogue usability. This also requires the engagement of a broad network of archives, libraries, research institutions and other resource holders for medieval and modern history in the development of the CENDARI infrastructure to support comparative historical approaches.

This report focuses on the progress of the activities of WP5 towards these objectives and outlines details for each task, summarizing WP5 working papers and workflows, building single steps in the process of conceptualising the shape and content of the CENDARI Archive Directory. The Directory is a shared repository populated by all members of the WP5 team, containing results of research on Medieval Manuscripts (MM) and the First World War (WW1) and building the basis for virtual research in these two domains.

This report on the Archive Directory covers the following sections:

- selection of relevant institutions and holdings;
- development of common standards and concepts to describe institutions and holdings in accordance with other CENDARI work packages;
- developing common workflows and choosing common tools;
- contacting institutions and aggregators with relevant holdings to start cooperation, including also elaboration of a data sharing agreement.

The dissemination work – presentations and conference papers - also belong to this last task of the work package. The report on work in progress on Archival Research Guides will be presented separately, but as it also refers to the content of the Archive Directory, some parts of it are presented here in order to illustrate the interdependency between both deliverables.

This report will focus on the CENDARI global vision, but when needed will also give details on domain specific (MM and WW1) details. For the sake of clarity we use the CENDARI DOW terminology to describe the WP5 deliverables (Archival Directory, Archival Research Guides, etc.), indifferently for WW1 and the Medieval domain, explaining in the text the most relevant differences and extensions.

Archive Directory: General Outline

For historical research, any historical source needs indications on the context in which it was created and distributed. Moreover, it requires the information of its history – its use, storage and authenticity. Usually, the cultural heritage institutions (archives, libraries and museums) possess the legal authority to give account on the credibility and reliability of source material they are responsible for. Thus, their catalogues and finding aids contain the necessary information on the subject of single holdings and record groups, integrating essential explanations on how to use any given inventory and how to interpret its contents. This quality, however, mostly refers to the working methods of archival institutions, less so to those of libraries and museums. In particular cases, like the medieval culture field of research, a part of this context - needed to fully describe and understand the *item* (i.e. the subject of the research process) - is provided by data coming also from research institutions.

Then again, libraries and museums are much more advanced in sharing and presenting their holdings to the broader public. These institutions more often form clusters and exchange data on the material they are holding. This also refers to the digital presentation of this material. More recently, however, archives have started to apply similar methods and to build collaborations and networks in order to promote the development of their digital presence. This ongoing process will sooner or later bring a differentiation, exchangeability and standardization of archival information as well as a comprehensive coverage of information on what is actually stored in these institutions. Being built on two seemingly detached pilot cases (MM and WW1), the CENDARI project is, to the contrary, a means of cross-fertilization between different communities (e.g. libraries, archives and research institutions) and research domains: the progresses made in the medieval digital ecosystem concerning Linked Open Data, interoperability, data integration and reuse - for example - helped the WW1 side to extend the list of potential tools, methods and standards to address their needs. Vice versa, the medieval domain was fostered to increase the level of interoperability with the archival community, traditionally less developed than the one with libraries and research centers.

At the moment, however, any research on archival holdings is to some extent confined to existing available information, especially in the digital domain, which depends on the granularity and the extensiveness of the displayed data. The digital presence does not allow for any final statements as to what exact resources and in what amount is to be found where. Hence, bearing in mind that the process of digitisation and the increasing digital availability of information on historical (re)sources is still going on it is necessary to adopt – as far as possible – the existing standards of describing (digital) archival data while at the same time enhancing the digital information by including other, existing analogue evidence on the historical material. Thus, a certain degree of interoperability and exchangeability – also sustainability – can be ensured. Once more, aware of the differences between the two digital ecosystems, the WP5 team worked to shape our processes taking advantage of the most advanced achievements in each domain (MM and WW1) to improve both and allowing the final digital ecosystem to be more consistent and less fragmented.

Consequently, the results of the initial assessment and evaluation process formed part of the process of developing the criteria for the shape and content of the Archive Directory. The WP5 team extended the research activities from desktop research to on-site research and, above all, to contacting the cultural heritage institutions in order to ensure their cooperation and willingness to share their digital content with CENDARI.

The two CENDARI pilot areas, Medieval Manuscripts and the First World War, both refer to a great number of historical sources for individual research. In almost every national archive or library, but also in regional, municipal and private institutions and societies, manifold material relevant to research on these two domains can be found. The MM domain also embraces a vast number of ecclesiastical organisations that are holding relevant sources for understanding a number of historical, social and intellectual facts during the Middle Ages.

Academic researchers and the wide community of historians researching the two pilot domains are the target group of users for CENDARI, and thus for selecting relevant information to be included in the Archive Directory. Towards this common goal and considering the different habits and hermeneutic methods of each research domain, WP5 developed focused approaches to meet the requirements of medieval and WW1 historians in terms of (re)sources selection. For instance, the number of existing materials differs significantly for research of medieval and modern times. The growth in literacy and bureaucracy as well as increased sophistication in regard to political and administrative practice produced an immense amount of material for WW1 research, while the amount of available medieval manuscripts, although undoubtedly significant in itself, is more easily manageable. This requires, on the one hand, a more granular approach to medieval sources – items (i.e. manuscripts) rather than collections – and, on the other hand, among the community of medieval experts to an increased practice of sharing data, comments and repositories on single medieval manuscripts, facilitated by the relatively clear and uncomplicated legal status of most of the medieval material.

In comparison, historians of the WW1 domain expect less granular information in a digital environment (even though they may wish for it) and focus on a more general approach to sources, meaning collections' and holdings' rather than item's descriptions, contained in content holding institutions rather than (non) existing web databases. Furthermore, due to the fact that medieval manuscripts are often (though not exclusively) stored in libraries while modern historical records mostly remain in archival institutions, the research for each domain focused accordingly on libraries and archives, respectively, to some extent corresponding with each other, as national research institutions proved relevant for both CENDARI research domains, the medieval as well as the modern.

Taking this into account, CENDARI has developed a *map* (i.e. the Archive Directory) of the memory institutions, collections and sources relevant for doing research in the two research domains of MM and WW1, built on top of shared principles but extended - when needed - to fit to the different research communities. For example, in the case of the MM domain, the Archive Directory comprises more than 300.000 shelfmarks of medieval manuscripts -

needed by scholars to investigate the medieval textual transmission. This was prepared by WP5 as a domain-driven extension of the shared Archive Directory.

Selection Criteria for Archival Institutions, WW1

The CENDARI Archive Directory is more than a simple list of archive and library addresses and their relevant holdings. It covers, in a representative manner, different types of institutions with archival holdings (archives, libraries, research institutes and museums) in all European and many non-European areas important for historical research of WW1. The Archive Directory is the backbone of the CENDARI research infrastructure on the content level. Any further information on holdings, collections and items will be built upon the information contained in the Archive Directory.

For historical research aiming at transnational and comparative studies, all source material originated by the belligerent countries of the First World War play a key role, in regard to their military, diplomatic and political actions and thus to the records of the central administration and military leadership. In addition, the requirements of the "*histoire totale*" and also of local or regional history of the War must not be neglected. The focus of WW1 studies has changed over the course of time – aspects of colonial or gender history, of memory and commemoration, of the history of everyday life ("*Alltagsgeschichte*") on the home front have been added to the topics of political processes of decision making or military developments on the Western and Eastern fronts. And this process has not ended yet. Accordingly, the relevant information on WW1 topics is vast and almost endless, given the fact, that private material, usually stored in family attics, has also become of interest for historians and has become accessible via web portals such as "Europeana14-18"[1].

This vast amount of archival records, stored all over the world in many different institutions, allows for a comparative yet cumbersome approach - due to the multilingual aspect but also due to the sheer volume of relevant material. Just as a researcher can only proceed with his/her work by selecting the most relevant papers (or pamphlets or photographs) for his or her research topic, CENDARI has also adopted selection criteria for organising and displaying information, based on well-reflected considerations to guide the priorities of the work progress.

At the level of institutions with archival holdings, priority lists have been developed to capture the diversity of the collected material with regard to geographical range, type of institution, on-topic relevance and status of digitisation or accessibility.

To ensure that CENDARI proves helpful to the community of historians, the selection of relevant institutions was defined as a core for any research on WW1. Therefore, institutions in former front regions, in capital cities, in heavy industry regions were given priority, just as armament firms´ archives were considered of bigger relevance than textile firms´ archives, etc. Military archives and museums were considered pivotal for inclusion in the CENDARI

[1] http://www.europeana1914-1918.eu

Archive Directory, just as national (and State) archives, libraries and museums in all belligerent countries were prioritized over those in neutral and non-belligerent countries.

Additionally, a representative coverage of all administrative levels was achieved, including municipal, regional, provincial institutions (e.g. “Laender”, départements, etc.) whilst also incorporating non-state institutions like church archives, archives of the Red Cross, of universities and non-governmental societies or War Graves Commissions.

WP5 placed special emphasis on archival institutions in East and South East Europe, referring to the main objective of the CENDARI project. Those institutions and their holdings often remain neglected due to the already mentioned aspect of language, as information about relevant material is only as perceptible as the language is readable.

In this context, CENDARI gave special focus to the so-called “hidden” archives - smaller and less explored institutions. Thus, the definition of “hidden archives” has been seen as part of the process of selecting relevant information in regard to lesser known institutions and also to holdings and collections which are not being accessed widely by historians due to the perceived lack of visibility.

Hardly any public archival institution can be considered “hidden” in the sense of their existence or presentation not being known to the scientific community. Almost every institution presents itself in a clearly visible way on the web, adding information on its history, holdings and publications.[2] For researchers, most relevant institutions themselves are well-known. However, the rules of the “digital era” establish new conditions of visibility. A homepage displaying address and office hours is no longer sufficient; more and more, the critical determinant of visibility is the quantity and quality of the information provided about the material stored in the institution. This determines the perceptibility of the material and thus the research on a given topic in a new way – as implied in the principle “on the Internet - in the world”. An institution’s visibility therefore depends on the intensity of its digital representation and the accessibility it provides to all available information on its holdings or items. Even some national archives, virtual key players in the field of First World War Studies can thus be considered “hidden” if they do not provide accessible, structured information on their collections and holdings, which still occurs to a great extent. On the other hand some minimal information with little granularity (e.g. nothing but the title of a collection or a record group) does not allow for much evidence either. It merely indicates the existence of potentially relevant material and has to be investigated more intensely in order to allow insight into its content: in order, thus, to become “unhidden”.

Moreover, the visibility of archival material relevant for WW1 and MM research also depends on the aspect of multilingualism. Information provided in the language of the country concerned is only as visible as the language is spoken. While Latin prevails dominantly in

[2] Some exceptions, like the Dashnak Archives in Boston, prove the rule. The archives of the Armenian Revolutionary Federation (Dashnaktsutiun) do not have any digital representation and are thus absolutely hidden within the digital space, even though they apparently are open to public since December 2013.

most medieval manuscripts, although strongly supported by numerous vernacular idioms, the content of modern historical records is much more multilingual and requires polyglotism, especially when the material deals with multinational conflicts and series of events. Apart from that, English has become the "*lingua franca*" of the Internet, so that meta-information provided English is also considered more "visible", while information in many other languages, especially Eastern and South Eastern European languages, remains less "visible" in a global context. A certain significant imbalance arises from the fact that historical material produced by the Western and the Eastern belligerent countries, respectively, is often treated differently, due to the geographical extension of the respective languages.

Selection Criteria for WW1 Institutions:

- Geographical Range: Eastern and South East Europe; all Belligerent Countries, Front Regions
- Type of Institution: Public and Private Archives, Libraries, Museums; state and non-state Institutions
- Relevance of Holdings: Representative Range from different Administrative Levels - National, Regional, Provincial, Municipal Institutions; Military, Diplomatic and Political Records
- On-topic Relevance: based on Selection of Topics for Archival Research Guides
- Status of Digitisation or Accessibility: "Hidden" - less visible and less accessible – Institutions, Repositories and Finding Aids.

Selection Criteria for Archival Institutions, MM

As stated previously, the CENDARI Archive Directory is the backbone for the contents of the CENDARI research infrastructure; this means that at the end of the establishment process there will be a unified repository, without any WW1 and/or MM silos; moreover, the information in the Archive Directory will constitute the raw material for further extensions and enrichments. On top of it a number of by-products will be elaborated according to specific needs as in the case of the manuscripts' shelfmarks list. The Archival Directory and its extensions will also be used as a foundation to support other services (e.g. the Medieval manuscripts map search tool), etc.

Although research in medieval culture is mostly related to manuscripts and their transmission, the relevance of archives in a state-of-the-art research infrastructure is evident: cutting-edge research in medieval culture is carried out using sources coming both from libraries and archives, making this distinction too constraining to be kept.

In this perspective WP5 shaped our activities towards completeness and relevance for the research community. In terms of completeness, we selected the institutions to include in the Archive Directory using traditional scholarly reference tools, such as (but not limited to) the *Iter Italicum* by Paul Oskar Kristeller and the *Latin Manuscript Books Before 1600: A List of the Printed Catalogues and Unpublished Inventories of Extant Collections* by Paul Oskar

Kristeller and Sigrid Krämer. We also used a number of scholarly digital repertoires such as the databases and authority lists in the *Mirabile* platform and in the *Integrated Archive for the Middle Ages* provided by the *Società Internazionale per lo Studio del Medioevo Latino* and the *Fondazione Ezio Franceschini*. The *National Czech Library* together with the *Università di Cassino* and other international partners (including the Italian Ministry for Cultural Activities, the French *Institut de Recherche et d'Histoire des Textes*, the Belgian *Bulletin Codicologique Scriptorium*, etc.) provided additional support in this activity. The spatial coverage of the medieval section of the Archive Directory includes institutions in Europe, USA, North Africa, Western Asia and Australia.

In terms of scientific relevance the WP5 team considered both the presence of sources (manuscripts etc.) ranging from VI to XV centuries as well as the extent (measuring both quality and quantity) of the related scientific production. We analyzed the bibliographical data coming from the Bibliographical Bulletin *Medioevo Latino*, covering the last 30 years of scholarly publications (both journals and books) written in English, Spanish, Portuguese, Italian, French, German, etc. (some 300.000 records related to many disciplines in the field of medieval studies, selected and reviewed by an international editorial staff).

From a methodological point of view, given that medieval historians generally undertake research based on the single item rather than focused on entire (contemporary) holdings and/or collections[3], the activity of identifying relevant institutions to include in the Archive Directory was a bottom-up process, starting from the cumulative shelf marks list. Analysing the secondary literature produced by the scientific community on single manuscripts over three decades, WP5 established a list of the most and least studied items: this merely quantitative data, together with the critical evaluations expressed in the abstracts (coming from the *Medioevo Latino database*) represented the parameter for including an item in the list of relevant sources. Since every (relevant) item is currently part of a holding or a collection preserved in a given memory institution, for each manuscript or document considered, we populated the Archive Directory with the related institution (library or archive) and the relevant collection.

Selection Criteria for MM Institutions:

- Geographical Range: virtually any country with institutions holding medieval manuscripts and documents (EU, USA and North Africa) have been considered;
- Type of Institution: Public and Private Archives, Libraries, Museums (including Ecclesiastical Libraries and Archives); state and non-state Institutions, other private collections;
- Relevance of Holdings: written sources for the reconstruction of administrative, political, religious, and cultural aspects of medieval and early modern era

[3] An important exception to this rule is represented by medieval libraries and collections held by institutions (i.e. monasteries, etc.) and/or individuals (e.g.: the dispersed library of San Giacomo della Marca) is not described here, since we are dealing with contemporary collections and institutions.

- On-topic Relevance: based on Selection of Topics for Archival Research Guides and on scholarly bibliography (last 30 years)
- Status of Digitisation or Accessibility: both "Hidden" (less visible and less accessible) and well known Institutions and Repositories

Selection Criteria for Archival Holdings and Collections, WW1

There is a great amount of material stored in cultural heritage institutions and hundreds of holdings and record groups relevant for WW1 research can be easily traced within almost each and every one of them, their physical content stored and arranged in paper trail. Each record group can be consulted via existing analogue inventories and other finding aids. Those finding aids are fully accessible within a given institution and often are also fully accessible online. However, many institutions are still in the process of digitizing their analogue catalogues and inventories and are far from completing this task. Some of them prefer to display at least some unstructured information in PDF or Word format, others give access to digital databases which are sometimes more, sometimes less well searchable.

Here again, selection criteria have had to be developed in order to capture the wide range of existing material and also to connect CENDARI research to the current developments within the community of historians of the First World War, as assessed by the work of CENDARI WP4.

The CENDARI institutions charged with this research (FUB, UOB, TCD and TEL) ensured good relations with the community via close cooperation with the project "1914-1918 online. International encyclopedia of the First World", an international project assembling almost all historians working in the field of WW1 studies.[4]

At the same time, results of activities of the other CENDARI work packages were also taken into account and informed the reflections of WP5, for example the outcomes of the participatory design workshops of WP8 and WP9 and Deliverable 4.2 the Domain Use Cases. The work on defining ontologies (MS6) for the knowledge framework (which will be reported by WP6) proved very useful for spotting the existing classifications and vocabularies which mirror the work of the community of historians of WW1 research. Here, WP5 implemented the taxonomy and classification system of the already mentioned "1914-1918-online" encyclopedia.

Finally, the research on relevant holdings and collections was closely related to the development of the framework for CENDARI Archival Research Guides (ARGs), the second deliverable of WP5.

In the process of identifying major archival collections that are essential for a functioning research infrastructure that will be used by the research community, some 48 European institutions were identified as pivotal for research on WW1. Because these are major

[4] http://www.1914-1918-online.net

research archives, they all have a digital presence. However, this presence does not guarantee the full display of information on all relevant collections and holdings and further investigation is needed. In the process of research on relevant material, WP5 realised that providing information will partly be possible via desk research and a combination of emails and Skype calls. In order to provide the CENDARI infrastructure with the data necessary for research, data sharing agreements with these archives will be signed in order to allow for automatic ingest of relevant material, wherever such ingest is possible.

In regard to the state of digitisation, the WP5 team is working to complement these data agreements with research and manual enhancement of the existing information, particularly in relation to the transfer of analogue information into digital formats. This is especially relevant for the “hidden archives” or “hidden collections” which are well accessible inside an institution but not similarly easily accessible in the virtual space. Many collections are not or only poorly represented in digital repositories and thus crucial information is lacking. CENDARI places an emphasis on the “hidden archives” in Central, Eastern and South East Europe, defining “hidden” in this case not only as archives with poor digital presence but also those institutions whose digital information is either not visible or not searchable in the format it is now, and can be enhanced e.g. by methods of natural language processing. In this context, digitally available collections and holdings which have been often neglected by traditional historical research practices, like film, photograph or newspapers collections, are also given special attention.

While developing the framework for the ARGs (see below), WP5 agreed on a number of basic requirements for the Guides, referring to the selection of relevant material from a methodological and historiographical perspective. At their core, the ARGs will emanate out of historiographical reflections on where current research in WW1 studies is going. This is essential in order to provide credibility for the CENDARI research infrastructure as a whole with the historians who will be using the infrastructure. As they are developed, the ARGs will also provide an analytical guide to further collection selection. WP5 will create lists of archives and collections for each guide in rank order of importance. Some of the archives will have a digital presence. The WP5 team will approach these archives and institutions with the goal of sealing a data sharing agreement allowing CENDARI to automatically ingest the relevant collection information. In cases where there is no digital information available, the WP5 team will complement the information ourselves.

Thus, the selection of relevant material for the domain of WW1 studies is strongly linked to current research within this domain and simultaneously to the work progress on the CENDARI ARGs which will deal thematically with crucial research questions within this domain. At the same time, the ongoing process of selecting and enhancing information for the ARGs will give credit to the immense amount of relevant material on the one hand, and this will be further continued by the community of historians who will work within the CENDARI infrastructure on the other, using tools provided by CENDARI WP9 and tested by CENDARI WP5 to annotate or comment as well as upload and process their own research notes and materials.

Selection Criteria for Collections and Holdings on WW1:

- Current Research Questions Valid for the WW1 Community of Historians
- Relevance for Topics chosen for ARGs
- "Hidden Archives" in the sense of not digitally accessible Information
- Different (often Neglected) Media Types – Film, Photographs, Newspapers, Audio files

Selection Criteria for Archival Holdings and Collections, MM

As already described, the process of establishment of the Archive Directory has been articulate and iterative, moving along the three pillars of items, collections and institutions, back and forth, requiring domain experts to validate the data gathered via automatic or manual procedures.

Starting from the initial list of the most relevant collections, selected using the principles explained above, for each collection the WP5 team provided a description, focusing on a set of priorities, as agreed in several WP5 meetings. Combining expert editorial work with information coming both from digital and printed scholarly reference tools (see above), institutional websites and digital collections, we elaborated detailed descriptions covering a number of historical (establishment and development), physical (extent, supports etc.) and scientific (distinctive traits such as in the case of the *homeliaries in beneventan* script preserved in Monte Cassino) elements.

As in the case of the WW1 domain, we constantly requested feedback directly from the research community, in order to avoid missing relevant items, which were possibly not covered by our survey. Starting with the CENDARI Participatory Design Session for medievalists (Florence, 25th January, 2013) WP5 regularly hosted a number of workshops and roundtables to discuss typologies of sources in the Middle Ages (Entertainment literature in the Latin Middle Ages. Florence, 22nd March, 2013[5]; The Scientific miscellany. Description, edition and comment of medieval miscellanies of scientific texts, Florence, 22nd-23rd October, 2013[6]; Medieval Anonymous Texts and Digital Research Infrastructures, Florence, 24th May, 2013[7]) involving both international scholars and Ph.D. students.

Furthermore in June 2013, CENDARI organised, in collaboration with the IS1005 COST Action Medieval Europe - Medieval Cultures and Technological Resources[8], the Summer School on Historical Sources and Transnational Approaches to European History (Florence,

[5] http://www.sismelfirenze.it/index.php/it/convegni/item/271-la-letteratura-di-intrattenimento-nel-medioevo-latino
[6] http://www.sismelfirenze.it/index.php/it/convegni/item/292-la-miscellanea-scientifica
[7] http://www.sismelfirenze.it/index.php/it/convegni/item/282-medieval-anonymous-texts-and-digital-research-infra-structures
[8] http://www.cost.eu/domains_actions/isch/Actions/IS1005

22nd – 26th July, 2013[9]) with attendees from different communities (libraries, archives, scholars) and research domains (WW1 and Medieval history).

Selection Criteria for Collections and Holdings on MM:

- Most relevant Research Questions for the MM Community of Historians, based on recent bibliography (last 30 years)
- Relevance for Topics chosen for ARGs
- "Hidden Archives" (i.e.: collections without a digital access, and/or less studied collection, regardless of the availability of a digital fingerprint)
- Typology of sources (i.e.: miscellaneous, scientific or illuminated manuscripts)

Archive Directory: Contents and Numbers

The backbone of the CENDARI infrastructure, the Archive Directory currently contains information on **more than one thousand institutions** relevant for research on Medieval Manuscripts and on the First World War, with more than 450 considered important for medieval research and more than 670 as part of the WW1 network. All European countries are covered, with emphasis placed on those countries which refer to the selection criteria as described above. Furthermore, institutions in many non-European countries were also included in the research. This institutional coverage is intended to build a matrix which will serve as a reference level from which to derive more granular information on archival material relevant for each pilot domain. Moreover, this matrix will also function as a link facilitating the attachment of further material and information in a hierarchical, yet connected and thus intuitive manner. Especially for the domain of WW1 research, institutions play an important role not only for locating archival holdings and collections, but also for connecting them to each other, giving credit to the sometimes very intricate history of wars and lootings, of relocations and fragmentations of historical materials. And in this way institutions also form an important piece of information in order to virtually reconstruct the sometimes illegitimate (and thus difficult to be traced) location and history of important archival material, which is strongly intertwined with the history of countries and nations.

The research by WP5, is now being pursued on the more granular level of holdings and collections and, above all, single historical documents. As it is not currently realistic to expect the entire volume of material for modern history (i.e. single letters, brochures, pamphlets or films) to be presented in a digital format, the starting point for any WW1 research is generally at the level of collections and holdings, classified within inventories, catalogues and repositories, many of them researchable online. They lead the way to the material an individual researcher is usually looking for, and this guidance improves as more information is extracted from their contents, which improves the possible search functionalities.

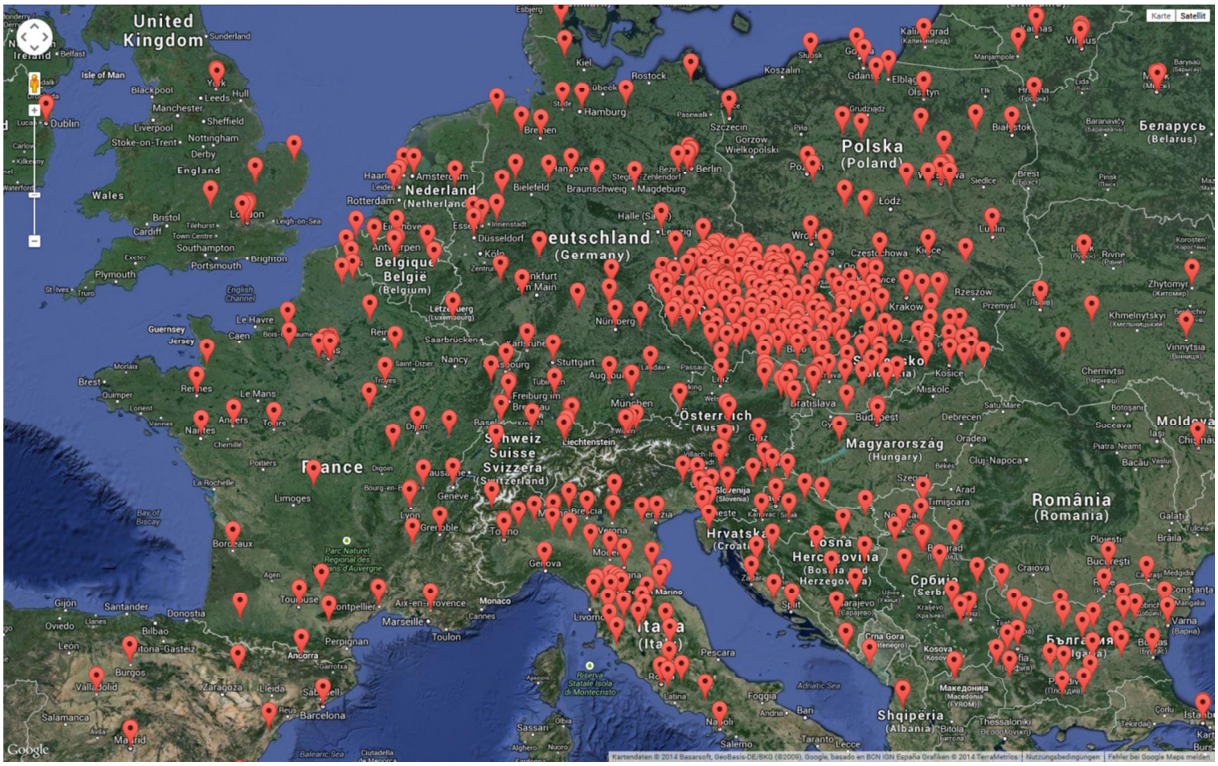
For the medieval domain, the starting point for research is, to a greater extent one single item (manuscript), collections being often rather arbitrarily formed without particular attention

[9] http://www.cendari.eu/research/summer-school-2013

to the aspect of provenance – which, by contrast, is the guiding motif for the preservation and classification of modern historical records. Thus, much more detailed information is needed for the MM research domain and it is often provided in an advanced stage in the digital environment.

Therefore, the WW1 domain research focuses on describing relevant collections and holdings in a substantial way and integrating digital finding aids, where possible. Descriptions going beyond the minimal content accessible online are being conducted on site, using the existing analogue material. They are also being complemented by additional material – published and not published – referring to these collections, e.g. pointing to selections of printed sources, of secondary publications, papers, objects within digital portals, etc. All this will at the same time be inserted and processed, complemented and enriched within the CENDARI Archival Research Guides.

The following image shows the current coverage of European institutions within the CENDARI Archive Directory.

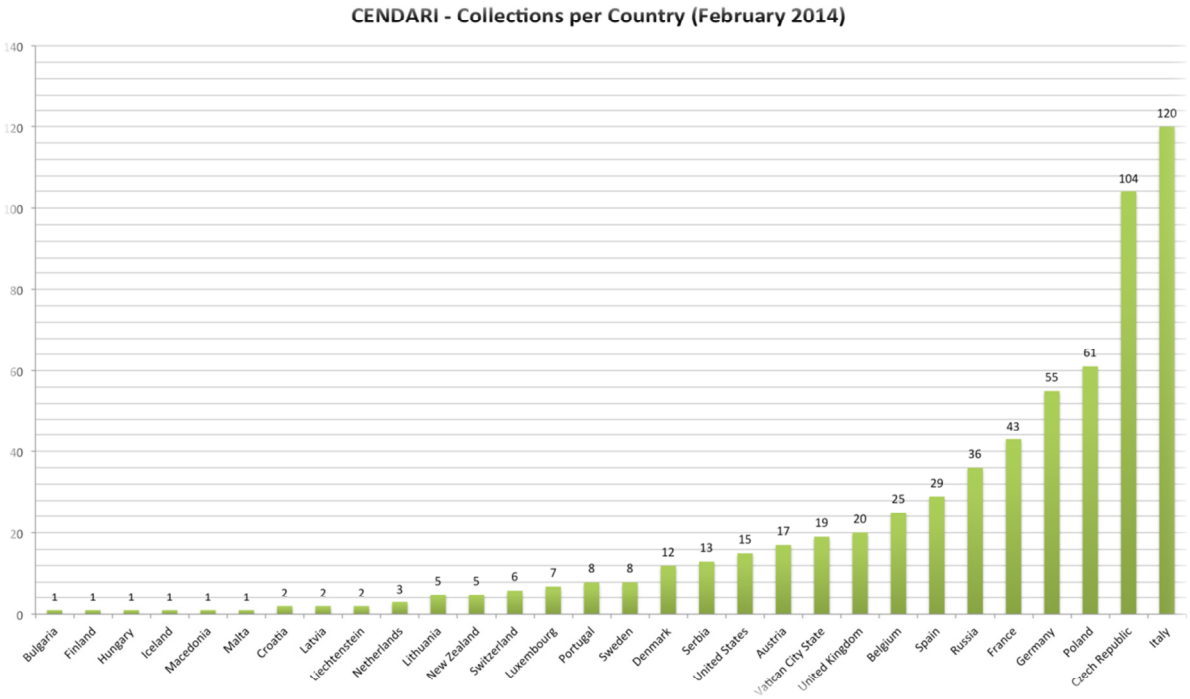


European Institutions in the CENDARI Archive Directory (Europe)

Some 800 collections and holdings descriptions for archives, libraries, museums and ecclesiastical organisations have been manually encoded by the WP5 team so far – ca. 400

for each domain, respectively. This process will continue and will also be carried on by future CENDARI users. At the same time, several hundreds of files representing descriptions and finding aids from different institutions (e.g. BDIC Bibliothèque de Documentation Contemporaine, IISH International Institute of Social History, Bibliothek für Zeitgeschichte, Universitäts-bibliothek Heidelberg, Czech National Library) have been ingested into the CENDARI repository. Furthermore, data have been inserted or processed from aggregators, portals and other European projects (e.g. The European Library, JISC and the Wellcome Trust, TRAME).

The following chart shows the number of (manually) encoded collections for WW1 and MM domain so far, stored in the CENDARI Archive Directory.



Collections descriptions encoded manually by WP5 correspond with EAD – Encoded Archival Description - the archival standard used in almost every European and also non-European country, which is approved by the International Council of Archives. However, the descriptions within the Archive Directory are also based on other standards not yet enclosed within the EAD, like TEI (Text Encoding Initiative). For different media types, different standards are applied (e.g. MODS – Metadata Object Description Schema, METS - Metadata Encoding and Transmission Standard, etc.). They will all be compatible with the emerging CENDARI research infrastructure and are simultaneously exchangeable with the

systems of archival institutions, which will allow the institutions to ingest back enriched data from CENDARI.

Contacts to Institutions and to Other Projects

In order not to duplicate the already existing digital information on relevant holdings for WW1 and MM research, WP5 has decided to put strong emphasis on establishing good contacts with cultural heritage institutions in order to allow for automated ingest of digitally available data. WP5 has developed a data sharing agreement in collaboration with partners from WP2 and WP9 and this forms the basis for establishing contacts with important institutions.

A list of 48 “priority archives” - archival institutions, libraries and museums - has been compiled by WP5, including all European institutions considered as essential content providers to any infrastructure dedicated to WW1 and MM studies. Contacts with some of those institutions have already been established by the WP5 team and data ingests performed. Other contacts have been initiated and are currently being developed.

The important process of trust building process is time consuming and labor-intensive but the WP5 team has observed a great interest on the part of cultural heritage institutions in the development of CENDARI. The process of developing and signing a data sharing agreement can often be delayed by the internal administrative structures of the institutions but it has proved fruitful for WP5 to engage the archival community in developing common strategies and CENDARI has benefitted from the synergy effects which arise. Cultural heritage institutions are currently working on their own strategies to gain more digital visibility and present their holdings to the public and are therefore interested in establishing networks and cooperating with projects like CENDARI as well as other potential users.

In order to promote the prominence of CENDARI´s activities and support this process, WP5 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 already mentioned project “1914-1918 online”, the European project Archives Portal Europe (APEx)[10], the project Europeana collections 1914-1918[11] and European Film Gateway 1914[12].

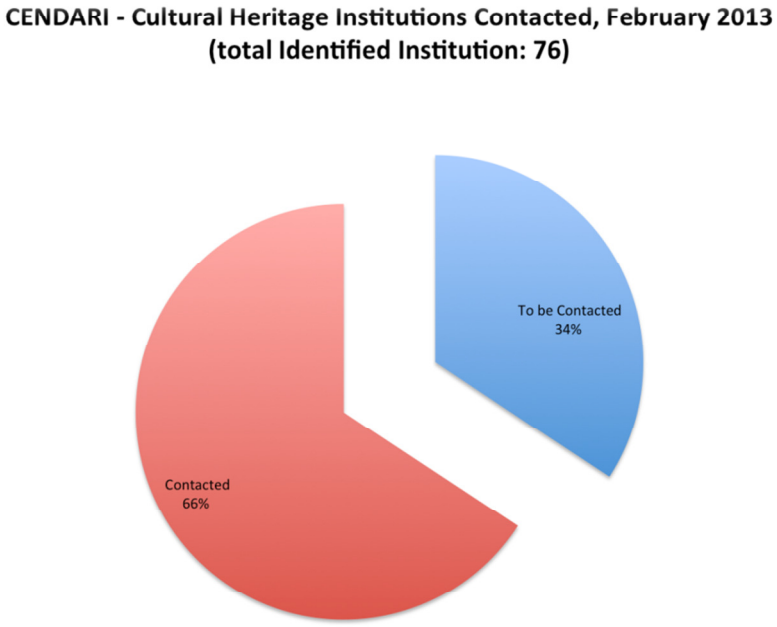
A similar set of activities has been established by the WP5 team focused on the MM domain to steadily connect the development of the project to the scholars and users communities. Permanent connections and joint efforts are already in place with the IS1005 COST Action. This network represents a group of more than 260 researchers coming from 39 leading institutions (archives, libraries, universities and research centers) in 24 countries across the EU. Within this framework WP5 organized a number of workshops and seminars, as well as contributing to the joint CENDARI/COST Summer School (see above) in July 2013. In

[10] http://www.apex-project.eu
[11] http://www.europeana-collections-1914-1918.eu
[12] http://project.efg1914.eu

addition WP5 had a number of presentations in national[13] and international workshops and conferences.[14] - Another set of relevant connections was established with the Italian branch of DARIAH.EU, the Medieval Electronic Scholarly Alliance (USA) and the BIBLISSIMA Project (France), in order to ensure interoperability with the most important actors in the medieval digital ecosystem.

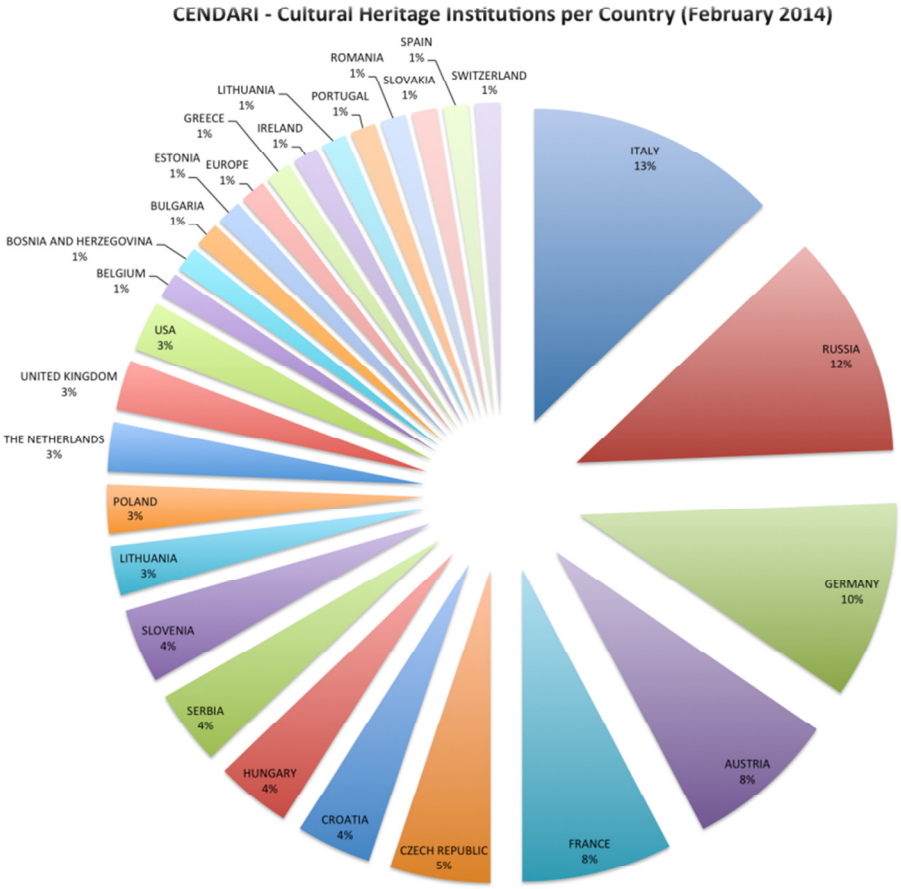
Furthermore, WP5 has presented CENDARI at numerous international conferences and workshops (e.g. APEx Conference, June 2013[15], "Unlocking Sources" Conference, January 2014[16]). Visits to archives and personal contacts with the responsible archivists, combined with on-site research on archival sources within the institutions, are being steadily continued.

The following chart shows the total proportion between institutions already contacted or shortly to be contacted by CENDARI, derived from the list of the priority institutions and institutions with holdings pivotal for the creation of the Archival Research Guides.



[13] 2nd AIUCD Annual Conference 2013 on Collaborative Research Practices and Shared Infrastructures for Humanities Computing. Padova 11-12 December, 2013 (http://aiucd2013.dei.unipd.it/)
[14] International Medieval Congress 2013. Leeds 4 July, 2013 (http://goo.gl/VLYlrK)
[15] http://www.apex-project.eu/index.php/events/dublin-conference, Dublin June 26-28, 2013.
[16] http://www.europeana-collections-1914-1918.eu/unlocking-sources, Berlin January 30-31, 2014.

The following chart shows cultural heritage institutions, organized by country, that are considered priority institutions in accordance with the selection criteria discussed above. These institutions have already been or are currently being approached by the WP5 team in order to establish a mutually fruitful cooperation.



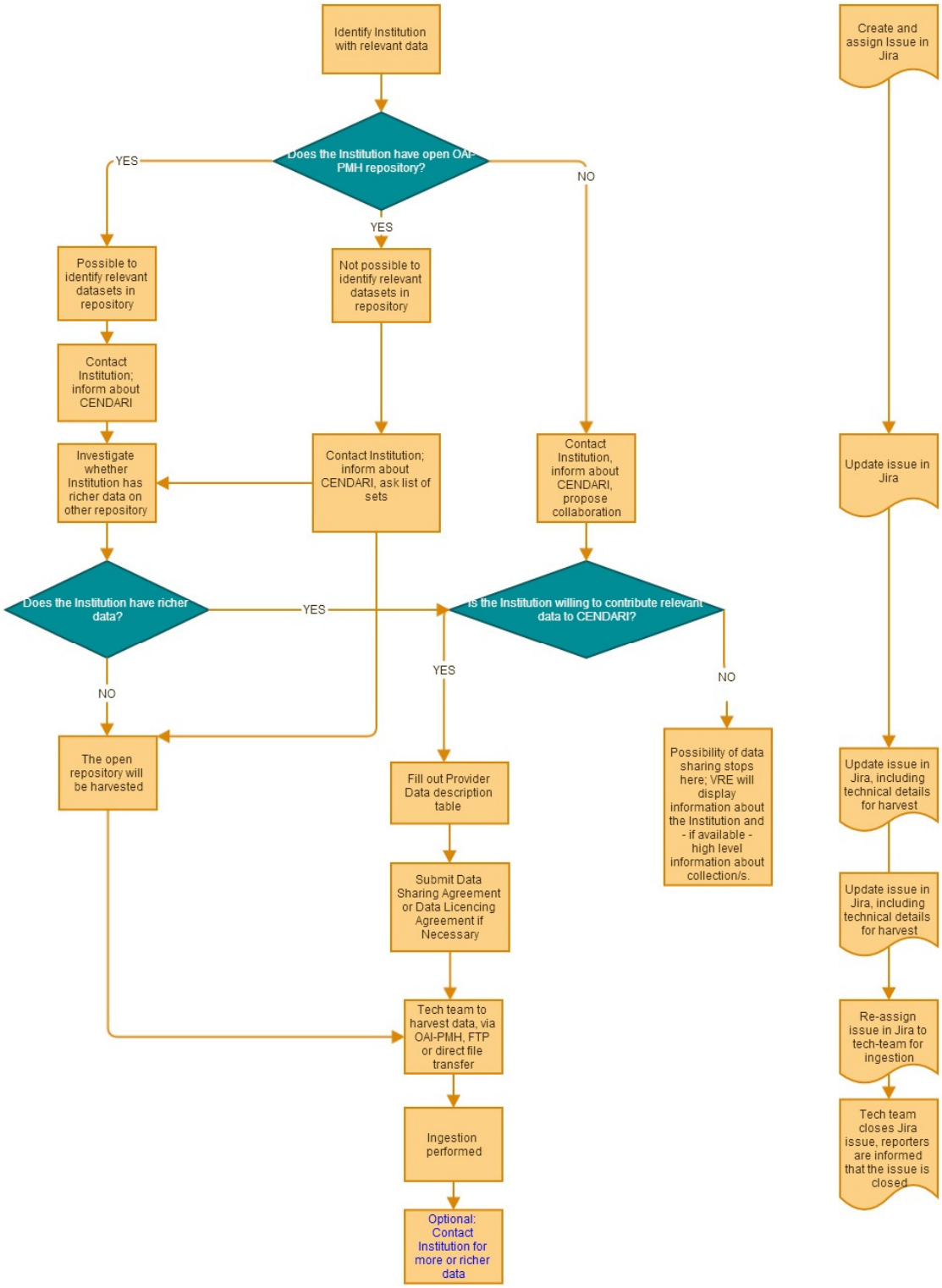
Altogether, WP5 has developed three scenarios for the process of data acquisition from cultural heritage institutions. The first one applies where an institution has an existing open interface (e.g. OAI-PMH); the second applies to an institution with an existing interface which is not open (e.g. FTP); the third applies to institutions where there are no interfaces, and the data can be ingested in a structured or a non-structured format. The WP5 team focused on the medieval domain is working to improve this third scenario by using metasearch tools and intelligent data collection, based on the TRAME technology, developed by SISMEL.

Generally speaking, WP5 advocates the strategy of a broad automated ingest of all relevant information. However, manual encoding of information that does not exist in a digital format is part of the workflow of WP5. Manual encoding is intended to be confined to non-digital

information only, in order to broaden the research space and not narrow CENDARI to a cabinet of curiosities.

The following flow chart shows the three possible scenarios for approaching the cultural heritage institutions established as a common ground from which to start.

In addition, a set of initial information – FAQs – was created by WP5, in cooperation with WP7 and WP2 in order to coordinate CENDARI´s activities in this field. It offers an institution a quick overview of the activities, goals and objectives of CENDARI, and answers questions of a technical nature as well as inquiries about special aspects of WP5's work, which we have encountered repeatedly when liaising with archivists and librarians.



Flow chart: Contacting Institutions

Standards and Tools

Together with the CENDARI partners from WP7, WP9 and WP6, the WP5 team agreed on using common workflows to ensure interoperability and compatibility of the provided archival information. A common repository guarantees the continued exchange of information on the work being carried out while simultaneously ensuring a process of ongoing quality control. The rules governing the first part of creation of the CENDARI Archive Directory were based on existing standards used by cultural heritage institutions to describe and encode information. This first part referred, in the first instance, to the description of institutions with archival holdings – their location, their contacts, their history and their relevant holdings. As this information was collected manually, it was ensured to be compatible with the common broadly applied standard, the Encoded Archival Guide (EAG). A selection of relevant EAG elements was shortlisted by WP5 and implemented into a customized EAG (CENDARI) model schema by the WP6 team.

A selection of appropriate tools was made in collaboration with all the concerned work packages and after testing a number of different software options and existing tools. Working with the WP7 team, a common workflow was then established, based on a combination of the following software: Oxygen xml-editor for encoding, Apache Subversion for versioning and eXtensible Text Framework (*XTF*) for accessing the digital content.

While it was possible and necessary to encode information on more than 1000 institutions manually, this workflow could not be maintained for the more granular level of describing holdings, record groups, collections or single items and artifacts. Moreover, as in many cases the relevant digital information already exists and does not need to be repeated. It is therefore the task of WP5 to make sure that data will be incorporated into the CENDARI infrastructure via an automated ingest of source information in order to broaden the Archive Directory further and ensure the usability of existing data.

The process of manual encoding of information (in a native xml format, considered the most appropriate and standardised format to confirm exchangeability and sustainability) has been confined to the level of institutional descriptions. However, according to what has been said above, information which is not currently available in a digital format still needs to be manually encoded by the WP5 team, in the course of the ongoing and forthcoming creation of the CENDARI Archival Research Guides. This becomes even more apparent where CENDARI is to become a collaborative working space for the community of historians

At the moment, some 800 collections have been manually encoded for both domains and the descriptions are being stored in the common SVN repository, both for the MM and the WW1 parts of the research infrastructure. The level of granularity differs from one description to another due to different levels of existing information and specific research needs. Furthermore, several hundred collections descriptions have been automatically ingested into the Archive Directory. Both processes will continue, covering the two possible approaches to archival institutions: the ingestion of digitally available information and the provision of information with no digital representation.

Archival Research Guides - Guide Framework

The framework for developing the Archival Research Guides (CENDARI project milestone MS4) was agreed on by the CENDARI partners in several meetings. The Guides are intended, in the first place, to facilitate the process of researching relevant historical material to a given research topic. They are conceptualised as thematic approaches to several exemplary and paradigmatic areas of research in each of the pilot domains, based on currently leading research questions and will enable research methods in the CENDARI virtual research environment.

The ARGs will facilitate the retrieval of collections and complement holdings descriptions within the CENDARI Archive Directory, providing historiographical context and methodological support to researchers. They will allow for a virtual composition and access to comparable holdings across national and institutional borders, bestowing information on dispersed and relocated material belonging to the same historical context. They are intended to cover big thematic areas of historical research and they will also serve as "showcases" for user-generated content which may also be created in a guide format.

The ARGs are a core activity of WP5, in which many activities of different partners will come together. They are intended to guide users to different contents and also to the application of virtual tools and systems, thus being an enhancement to the traditional methods of historical research. Furthermore, they should make the difference between a research on site – i.e. within an institution – and a digital research visible, i.e. stressing upon the fragmented and incomplete access to information facing the fact that not all information is yet available in a digital format.

Hence, the ARGs are being designed as methodological, paradigmatic guides for a virtual research environment. Some of the Guides will have an **emphasis on "guides"** rather than on "research", i.e. how to find closely related but physically dispersed material all over Europe and the World (e.g. The Fall of the Romanov Dynasty; Vernacular Bibles; Chivalric Romance).

Some ARGs will be designed as access points to relevant contemporary research questions of the two pilot domains, **with an emphasis on "research"** rather than on "guides" (e.g. Coercion and Consent in the Belligerent Armies; Prisoners of War and their Return Home; Prophecy and Political Thought in the Medieval Age; Mnemonic Devices in Medieval Latin and Vernacular Cultures).

And finally, the ARGs will act as explanatory translations from the analogue into the digital sphere, **with an emphasis on "archival"** rather than "guides" or "research" and they will relate to the changes in presentation and representation of archival holdings within the CENDARI Archive Directory. They will address questions like: 1) how archives work, classify and organize information (and which information); 2) which aspects of their work change when this information is available online (like transnationality, interexchange, ubiquity of access, etc.) – and which do not change at all (e.g. the physicality of documents which are still mostly stored in institutions); 3) how to find and retrieve paths to locate those physical

objects. (e.g. Parallel Records and Supplementary Documents: Polish Material on WW1; Medieval Libraries: Dispersion and Survival).

However, while respecting these three methodological aspects, the ARGs are intended to cover all three of them in a "hybrid" way, differing in the strategic focus and prioritising some aspects more than others. Nevertheless, they will all **guide** the path for **research** questions to be answered and worked on, reflecting the methods of the **archives, libraries and museums** storing the historical material.

The topics of the ARGs have been and continue to be selected according to a transnational approach, taking into account different multimedia sources. They focus on historical concepts, events or developments and enclose recent historiographical currents. The chosen topics do not claim objectivity, but according to the above mentioned methodological approaches they are perceived as paradigmatic and exemplary collections of information to be enhanced and complemented by researchers through comments, annotations, links etc.

As the development of the CENDARI ARGs is strongly related to the development of the Archive Directory and at the same time related to the current developments within the areas of medieval and WW1 research, the final selection of topics is a process which will be finalised in due time. The following list comprises topics which the team is currently working on:

WW1:

- Prisoners of War and their Return Home
- Coercion and Consent in the Belligerent Armies
- Workers and Workers´ Movements
- Private Memories of the First World War
- Women in the First World War
- The Fall of the Romanov Dynasty
- National Narratives of the War – Memory and Commemoration
- Parallel Records and Supplementary Documents: Polish Material on WW1
- The "Jewish Question" 1914-1918 and after

MM:

- Early Medieval Poetry
- Prophecy and Political Thought
- Medieval Poetry
- Chivalric Romance
- Vernacular Bibles
- Mnemonic Devices in Medieval and Vernacular Cultures
- Medieval Libraries: Dispersion and Survival

The content of the ARGs does not claim to be exhaustive and can, and should, be further enriched by users. The Guides are perceived as a starting point within the CENDARI

research infrastructure for research on a given topic and a framework in which to work. The Guides will comprise collections descriptions of different granularity in more than one language across several institutions. They will strike a balance between collections with in-depth descriptions and digitised finding aids and collections descriptions from “hidden archives”. They will comprise visual material as well as secondary – analogue and digital – resources.

Although the ARGs will differ to a certain extent concerning their focus and methodological target, there are nevertheless several components considered to be essential for both research domains (MM and WW1), in regard to their function as guides to topics and materials, whilst also meeting the requirement for historiographical and editorial context. Thus, some obligatory elements will be incorporated into all of the Guides: introduction, historiographical context (current research), narrative short texts explaining aspects of the topic and relating them to the relevant institutions and the existing archival material. They will also comprise secondary information on the material, linking to digital repositories, bibliographies and digital editions of source material as well as historical publications within the digital and the analogue sphere. The ARGs will also provide space for the community of historian users to comment, annotate, remark and request on topics and details, to tag keywords and add complementary material.

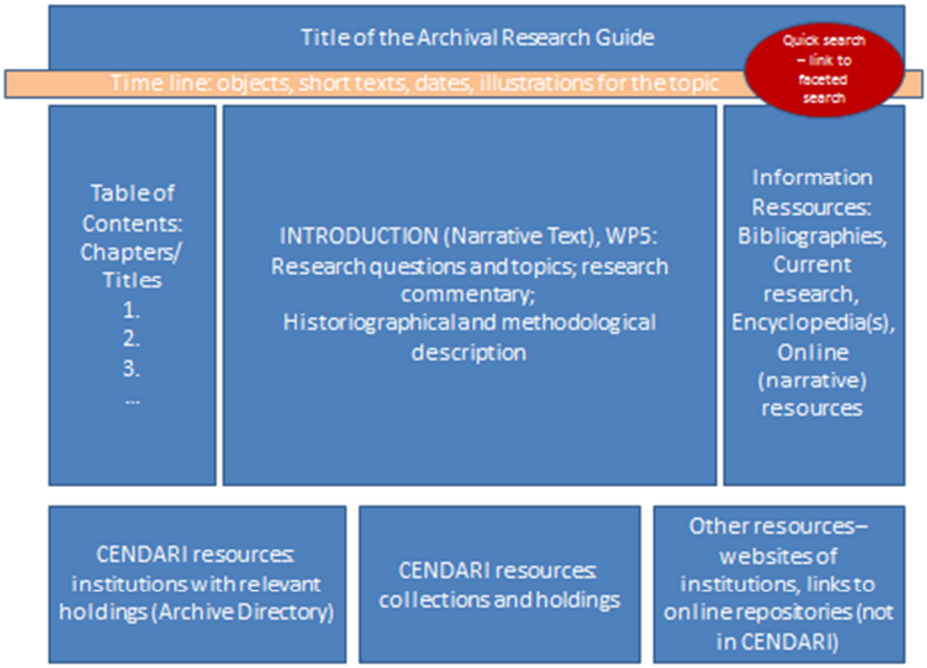
They will thus be built upon and fit into the CENDARI knowledge framework developed by WP6, incorporating ontologies, vocabularies, authority files and other sources of knowledge.

Archival Research Guides building on Archive Directory: Preview

As the Archival Research Guides are not intended to be merely a static digital addition of sources and information, WP5 conceives them as closely related to the development of the CENDARI virtual research environment and thereby also to the development of tools for the management of user-created content. Certain measures will need to be applied in order to clearly distinguish between content generated by different groups of content providers – the distinction being drawn between information provided by public, and thus authoritative, institutions, by the CENDARI team and by the users.

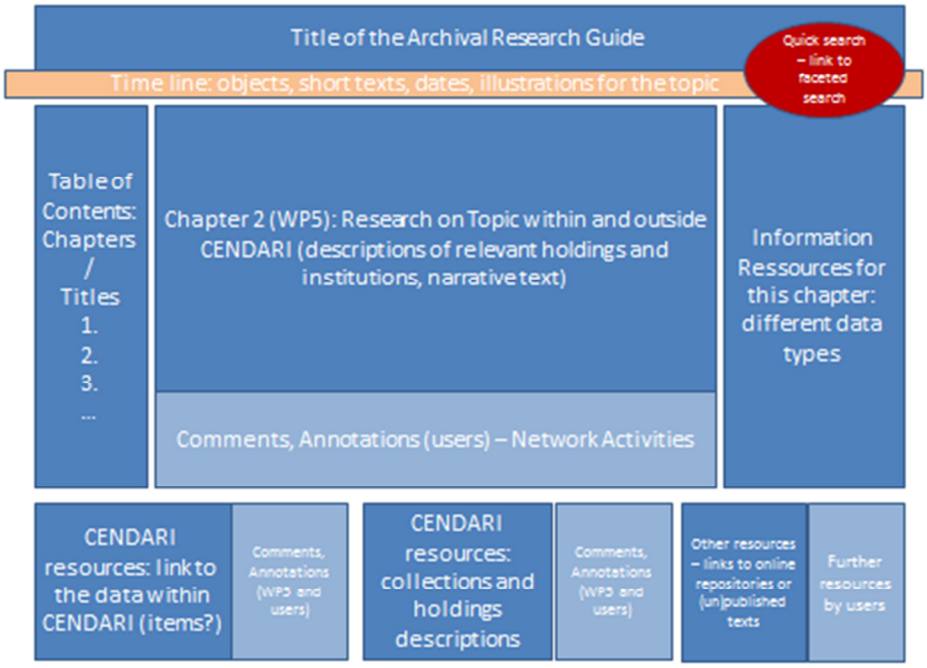
The ARGs are supposed to be handled intuitively; therefore it is essential to develop them as an easy-to-use set of tools and options. Members of the WP5 team have created a number of wireframes and mock-ups, which will be developed to become a dynamic and interactive research space. They include the components listed above and serve as the initial visualisations of requirements for tools and functions, which will be considered and worked on by the CENDARI development team in the coming months. The following drafts are thus intended to give an impression of how the concepts of content and shape for the ARGs can be visualised.

1. Introduction



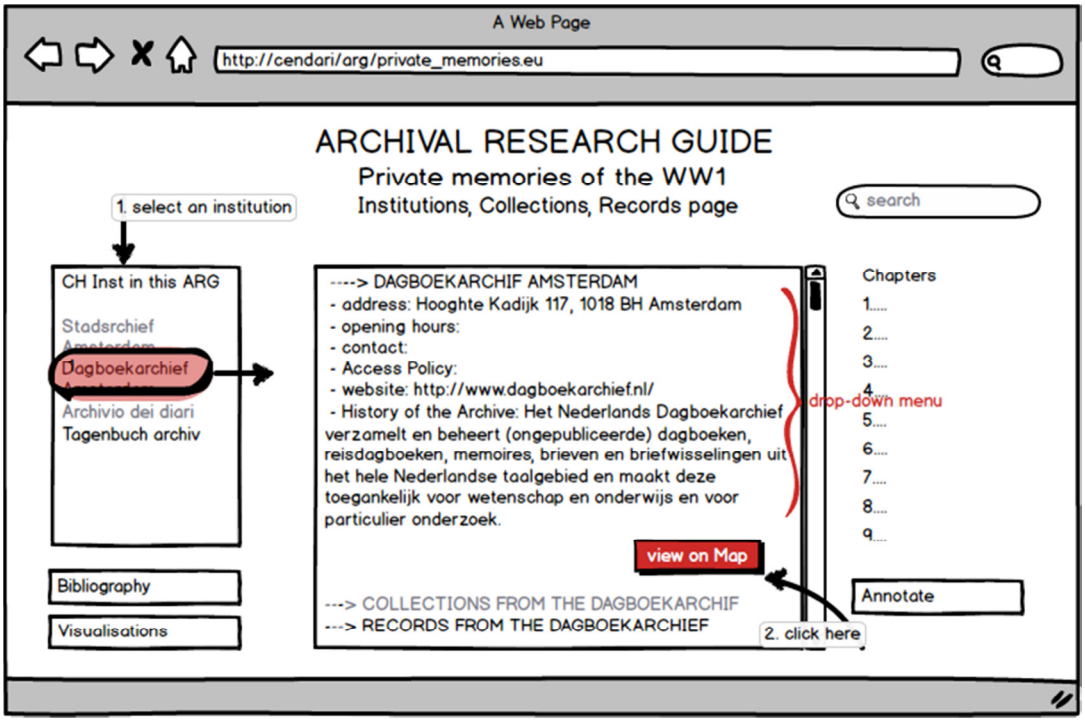
1. Introduction text / Abstract / Methodology
2. Search functionality
3. List of relevant cultural heritage institutions (links to websites)
4. Table of contents (link to chapters)
5. Link to bibliography
6. Link to visualisations (maps, images, etc.)
7. Name of author
8. Names of collaborators
9. Names of other authors (of annotations, comments, etc.); link to annotations provided by researchers

2. Chapter Page



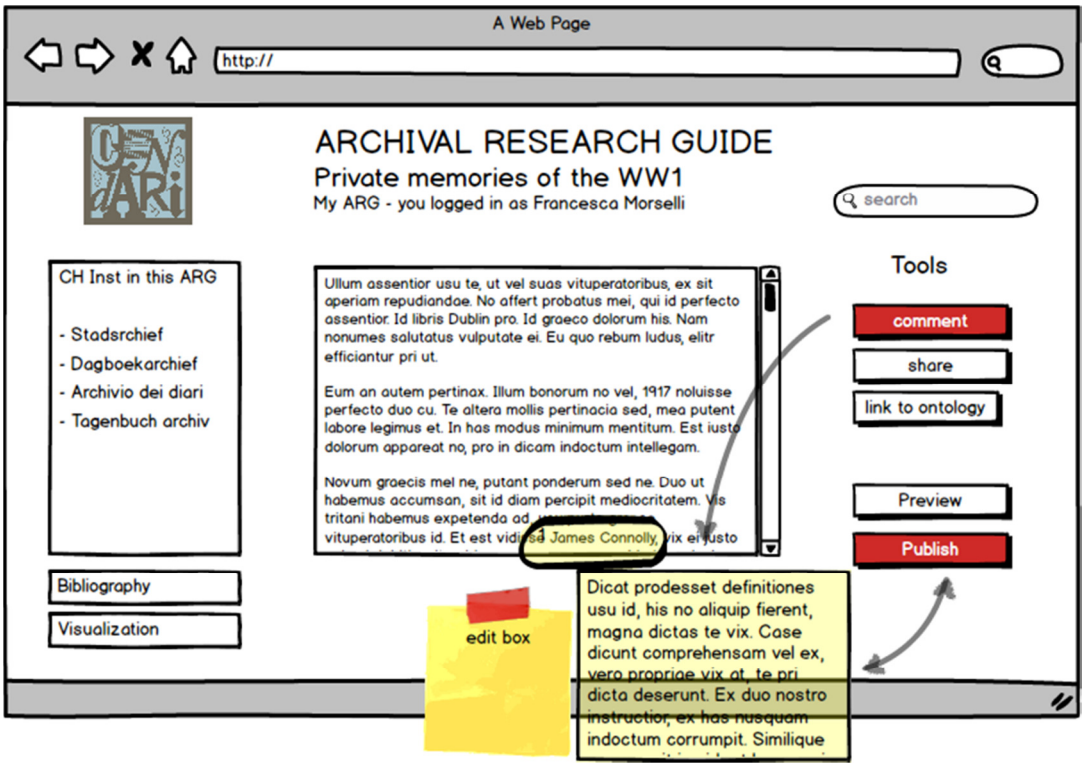
1. Chapter text
2. Search functionality
3. List of relevant cultural heritage institutions (links to websites)
4. Table of contents (link to chapters)
5. Link to bibliography
6. Link to visualisations (maps, images, etc.)
7. Name of author
8. Names of collaborators
9. Names of other authors (of annotations, comments, etc.); link to annotations provided by researchers
10. Button to collections and records relevant for this chapter (links)

3. Institutions, Collections and Record Page



1. List of institutions (for each selected institution the central text will show the information about the institution itself, its collections and records)
2. Main text space: three headers (Institution, Collection, Record) opened up in a drop down menu
3. 'View' on Map button (link to map to locate selected institution)
4. Search functionality
5. Link to chapters
6. Link to bibliography
7. Link to visualisation page

4. Annotation Page



1. Visualisation chapter text
2. Search functionality
3. Link to institutional website
4. Link to bibliography
5. Link to visualisation page
6. Comment Button: once one or more words are selected the user can click on the "Comment" button - an editing box opens and the user can write text in it.
7. Share button (link to share page)
8. Link to ontology button (link to ontology- annotation page)
9. Preview button (shows a preview of the page in the exploration mode, after the editing)
10. Publish button (the user can save the changes) in the public space

Summary and Further Work

The main goal for WP5 activities in the first half of the project was to establish the CENDARI Archive Directory, a matrix to which more in-depth information on archival material relating to the two pilot domains of CENDARI, Medieval Culture and the First World War, will be fixed to. At the same time and in relation to this matrix, a framework for creating the CENDARI Archival Research Guides was conceptualised, leading the way for new methods of historical research in the digital space. This will lead to the creation of an innovative CENDARI research infrastructure for historians, which will go beyond the existing portals and digital repositories and meet their needs for collaborative as well as individual research.

In the course of the work programme, the team has selected relevant institutions and holdings for the two pilot domains. It agreed on common tools to enable collaborative workflows and together with other CENDARI partners decided on standards and formats to store and further process provided information.

As a result, an Archive Directory containing information on more than 1,000 institutions and 800 collections has been created. It will be further populated with data relevant for the two domains via manual encoding as well as via automated ingest of already existing digital information in order to provide a broad content base for researchers' work. This will give them the space to elaborate their own research paths in a powerful environment.

At the same time, the team developed strategies for collecting relevant information and established workflows to fulfill this objective. Three scenarios have been abstracted from the work experience so far: harvesting existing information on archival material from open repositories, contacting prioritised institutions in order to include their data in the CENDARI repository and manually encoding information on relevant material which is not currently available in a digital form.

The amount of existing archival material for both domains makes it necessary to connect the research on the archival material with its exposure within the CENDARI ARGs. The Guides will work as paradigmatic approaches to research on digital information on the level of collections, holdings and single documents or manuscripts. The Guides will be designed to allow researchers to enrich, annotate and comment on them, thus providing a model for users to create their own research paths.

Consequently, the further activities of the WP5 team will combine the enrichment of the Archive Directory with the development of the Archival Research Guides, both on the technical and the conceptual side.