The CENDARI Infrastructure

Nadia Boukhelifa, Mike Bryant, Nataša Bulatović, Jean-Daniel Fekete, Milica Knežević, Jörg Lehmann, David Stuart, Carsten Thiel

The CENDARI Infrastructure

NADIA BOUKHELIFA, INRA
MIKE BRYANT, King's College London
NATAŠA BULATOVIĆ, Max Planck Digital Library
IVAN ČUKIĆ, Faculty of Mathematics, University of Belgrade
JEAN-DANIEL FEKETE, INRIA
MILICA KNEŽEVIĆ, Mathematical Institute of the Serbian Academy of Sciences and Arts
JÖRG LEHMANN, Faculty of Humanities, University of Bern
DAVID STUART, King's College London
CARSTEN THIEL, University of Göttingen

The CENDARI infrastructure is a research-supporting platform designed to provide tools for transnational historical research, focusing on two topics: medieval culture and World War I. It exposes to the end users modern Web-based tools relying on a sophisticated infrastructure to collect, enrich, annotate, and search through large document corpora. Supporting researchers in their daily work is a novel concern for infrastructures. We describe how we gathered requirements through multiple methods to understand historians' needs and derive an abstract workflow to support them. We then outline the tools that we have built, tying their technical descriptions to the user requirements. The main tools are the note-taking environment and its faceted search capabilities; the data integration platform including the Data API, supporting semantic enrichment through entity recognition; and the environment supporting the software development processes throughout the project to keep both technical partners and researchers in the loop. The outcomes are technical together with new resources developed and gathered, and the research workflow that has been described and documented.

CCS Concepts: • **Applied computing** → **Arts and humanities**;

Additional Key Words and Phrases: History, research

The research leading to these results received funding from the European Union Seventh Framework Programme (FP7/2007-2013/FP72007-2011) under grant agreement 284432.

Authors' addresses: N. Boukhelifa, UMR 782 INRA / AgroParisTech, Site de GRIGNON, Bâtiment CBAI, 78850 Thiverval-Grignon, France; email: nadia.boukhelifa@gmail.com; M. Bryant and D. Stuart, King's College London, Strand, London, WC2R 2LS, United Kingdom; emails: michael.bryant@kcl.ac.uk, dp_stuart@hotmail.com; N. Bulatović, Hybris Gmbh, Nymphenburger Straße 86, 80636 München, Deutschland; email: nbulatovic@o2mail.de; I. Čukić, Faculty of Mathematics, University of Belgrade, Studentski Trg 16, 11000 Belgrade, Serbia; email: ivan@math.rs; J.-D. Fekete, INRIA Saclay - Île-de-France, Bat 660, Université Paris-Sud, F91405 ORSAY Cedex, France; email: jean-daniel.fekete@inria.fr; M. Knežević, Mathematical Institute of the Serbian Academy of Sciences and Arts, Kneza Mihaila 36, 11000 Belgrade, Serbia; email: mknezevic@mi.sanu.ac.rs; J. Lehmann, Faculty of Humanities, University of Bern, Länggassstrasse 49, Postfach, 3000 Bern 9, Switzerland; email: joerg.lehmann@germ.unibe.ch; C. Thiel, University of Göttingen, State and University Library, Platz der Göttinger Sieben 1, 37073 Göttingen, Germany; email: thiel@sub.uni-goettingen.de.

https://doi.org/10.1145/3092906

ACM Reference format:
Nadia Boukhelifa, Mike Bryant, Nataša Bulatović, Ivan Čukić, Jean-Daniel Fekete, Milica Knežević, Jörg Lehmann, David Stuart, and Carsten Thiel. 2018. The CENDARI Infrastructure. *ACM J. Comput. Cult. Herit.* 11, 2, Article 8 (April 2018), 20 pages.
https://doi.org/10.1145/3092906

1 INTRODUCTION

The CENDARI infrastructure is the technical result of the CENDARI project (CENDARI 2015), a European infrastructure project funded by the EU for 2012–2016. The infrastructure is designed to provide and support tools for historians and archivists. The tools are Web based, using modern Web-oriented technologies and standards. The CENDARI infrastructure is innovative because it is designed to address multiple scenarios with two types of actors: researchers and cultural heritage institutions providing data. Both benefit from the platform, although the novelty concerns the researchers more. To address researchers' needs, the platform provides a note-taking environment (NTE) to perform the initial steps of historical research, such as gathering sources, writing summaries, elaborating ideas, planning, or transcribing. CENDARI integrates online available resources initially curated by cultural heritage institutions with the explicit goal of integrating them into the infrastructure to support the research process. Researchers from the project visited these institutions and negotiated data sharing agreements. For archivists, an archival directory was implemented, which allows the description of material based on the international Encoded Archival Description (EAD) standard (http://www.loc.gov/ead/index.html). The resulting infrastructure not only provides access to more than 800,000 archival and historical sources but also integrates them into a collection of tools and services developed by the project as a digital resource for supporting historians in their daily work.

The researchers' primary entry point into CENDARI is via the NTE (Figure 1). It enables curation of notes and documents prepared by researchers within various individual research projects. Each project can be shared among colleagues and published once finished. These notes and documents can be linked to the data existing in the CENDARI Repository. This comprises both the linking of entities against standard references such as DBPedia and connection to archival descriptions. A faceted search feature is part of the NTE and provides access to all data and additionally connects to the TRAME service (Trame 2016) for extended search in distributed medieval databases. The repository is based on CKAN (http://ckan.org/), manages the data, and provides a browser-based user interface to access it.

To support the creation of curated directories of institutions holding relevant material for both research domains, and to raise awareness about the "hidden" archives that do not have a digital presence or are less known but relevant, we integrated the AtoM software (https://www.accesstomemory.org), enabling historians and archivists to add new and enrich existing archival descriptions in a standardized way, following the EAD.

Once data is collected, an internal component, called the *Litef Conductor* (Litef), processes it for further semantic enrichment. It then sends the text extracted from the documents to the Elasticsearch (http://www.elastic.co/products/elasticsearch/) search engine, invokes the high-quality Named Entity Recognition and Disambiguation (NERD) service (Lopez 2009), and generates semantic data inferred from several document formats, such as archival descriptions or XML encoded texts.

The connection to the Semantic Web is further extended through ontologies developed specifically for the intended use cases and connected through the knowledge base (KB). Researchers can upload their own ontologies to the repository through the *Ontology Uploader* tool. To explore the semantic data collected, the *Pineapple Resource Browser* provides a search and browse Web interface.

To summarize, the CENDARI technical and scientific contributions are the following:

—The use of participatory design sessions as a method to collect users' needs
—The design of the infrastructure as a research support tool

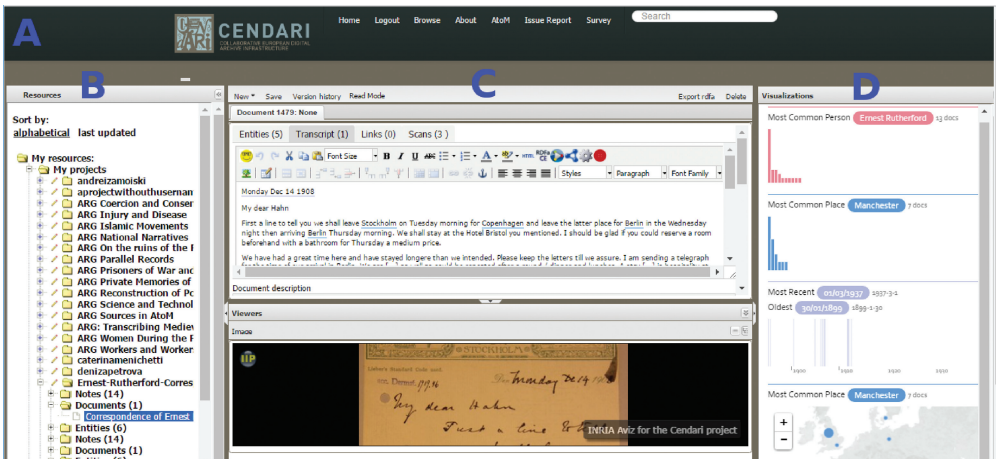

Fig. 1. The integrated NTE for historians: the search and browse panel (A), the resources panel (B), the central panel dedicated to editing, tagging, and resolution (C), and the visualization panel (D).

—The deep integration of multiple resources through a data integration and enrichment process to further support the researchers' workflow.

CENDARI has generated several benefits and outcomes:

—An abstract workflow of how historians and archivists conduct early stages of research
—An integrated system to support the early stages of historical research
—A catalog of hidden archives in Europe (more than 5,000 curated institutional and archival descriptions with special attention given to archives without own digital presence)
—A rich set of archival research guides to help and guide historians by providing transnational access to information (Edmond et al. 2015a)
—Integrated access to heterogeneous data from multiple data sources and providers
—A rich set of resources accessible and searchable from the Web.

2 RELATED WORK

CENDARI is related to several other digital humanities (DH) projects that tackle aspects of the humanities research workflow from either a generalist or topic-specific perspective. The project is directly connected to the Digital Research Infrastructure for the Arts and Humanities (DARIAH) (http://dariah.eu/), harmonizing national DH initiatives across Europe, with a goal to build and sustain an infrastructure that supports and enables the research of humanities scholars in the digital age. From its inception, CENDARI was designed to rely on and ultimately hand over the resulting infrastructure to DARIAH for sustainability and reuse.

The European Holocaust Research Infrastructure (EHRI) project aims to assist the process of conducting transnational and comparative studies of the Holocaust by bringing together in an online portal (https://portal.ehri-project.eu/) archival material across a geographically dispersed and fragmented archival landscape (Blanke and Kristel 2013). Its principal goal is related to data collection, integration, and curation.

The TextGrid Virtuelle Forschungsumgebung für Geisteswissenschaften (https://textgrid.de/) developed a virtual research environment targeted specifically at the humanities and textual research. Its open source components are implemented as a desktop client to work with TEI encoded documents and to create and publish scholarly editions. Following the end of the project, the infrastructure components and technology were integrated into DARIAH.

To understand the needs of historians in the context of transnational history, we have used two key methods: the development of *use cases* and *participatory design sessions*. Participatory design has been adopted by many disciplines where stakeholders cooperate to ensure that the final product meets their needs. For interactive software design, which is the focus of this work, the aim is to benefit from different expertise: designers know about the technology, and users know their workflow and its context. In a similar vein, Muller (2003) describes participatory design as belonging to the in-between domain of end users and technology developers that is characterized by reciprocal learning and the creation of new ideas through negotiation, co-creation, and polyvocal dialogues across and through differences.

Until recently, participatory design was not a common practice in humanities projects. Warwick (2012) provides explanations for this paucity:

> It was often assumed that the resources created in digital humanities would be used by humanities scholars …there was little point asking them what they needed, because they would not know, or their opinion about how a resource functioned, because they would not care. It was also assumed that technical experts were the people who knew what digital resources should look like …their opinion was the one that counted, since they understood the details of programming, databases, XML and website building. The plan, then, was to provide good resources for users, tell them what to do and wait for them to adopt digital humanities methods. DH has since then changed perceptions, and we can now see several projects adopting participatory design to learn about users and their requirements (Mattern et al. 2015; Wessels et al. 2015; Visconti 2016).

3 DESIGN AND DEVELOPMENT PROCESS

Assisting historians and archivists in their research on transnational history is an important but abstract goal. To design relevant tools, CENDARI first had to understand related needs and requirements of historians and archivists. It turned out that the work process of these researchers is not well described in books or research articles, so the initial design phase of the project consisted of obtaining information about the needs and practiced research processes of our target researchers.

3.1 Use Case

Let us start by an example of a historian's workflow, acting as a use case for historical research carried out in CENDARI. It has been developed alongside the *Archival Research Guide* (ARG) *on Science and Technology* written by one of the historians in the project. This ARG examines the shift within the transnational scientific network during and after World War I (WWI), which resulted in the isolation of the German Empire within the international scientific community. The research question is whether this shift can also be found in the correspondence of individual scientists. There was an intense international exchange between physicists before the war, and a particular prominent example is the case of the British scientist Ernest Rutherford, later Nobel Prize winner and president of the British Royal Society.

As a first step, a new project with the name "Ernest Rutherford Correspondence" was created in the NTE. Afterward, available resources were searched through the faceted search, integrated in the NTE, and selected results (consisting of descriptions of archival material) were saved into a note.

The next was a visit to an archive in Berlin, which holds the biggest collection of Rutherford's correspondence in Germany. Several notes were taken there and added to the project space in the NTE. Photographs of the letters were taken, uploaded to the private research space, and later transcribed in the working space, using the provided image viewer.

The transcription of each document was processed with the NERD service, integrated within the NTE. This service produces many results; since the visualization of these entities cannot always be done automatically, the user might have to resolve entities manually, such as by selecting the appropriate place or person from the list

Fig. 2. The three participatory design sessions held with WWI historians, librarians, and medievalists.

of alternatives provided in the resources panel of the NTE. The visualizations on the panel to the right show Rutherford's correspondents in a histogram, the places from where he wrote his letters on a map (e.g., Montreal and Manchester, where he lived from 1907 onward), and a timeline of how Rutherford's German correspondents did not receive any letters from him after 1914. In this way, the research hypothesis—abrupt ending of exchanges between German and Anglo-Saxon scientists from the beginning of WWI onward—is being substantiated.

From the list of resources available in archives and libraries, it can be discovered that Cambridge University Library holds the correspondence and letters of Rutherford and thus the most important and comprehensive part of his heritage. If the researcher is not able to visit Cambridge, for example, because of lack of financial support, the researcher can ask a colleague abroad to contribute to the endeavor by visiting Cambridge and checking the correspondence in the archive. To enable a researcher abroad to contribute to the research project, it can be shared by ticking a box in the project edit page. That does not mean that the project is publicly visible, as all content is inherently private in the NTE. The project is simply being shared with one colleague based in another country. In this manner, collaborative and transnational research becomes possible. Once the material in Cambridge has been explored and partially described, more complex interpretations and the deeper layers of the research question can be pursued—in this case, for example, the correspondence and interaction between Rutherford and Otto Hahn in the early 1930s, and Rutherford's influence on Hahn's discovery of the nuclear fission.

3.2 Participatory Design Workshops

There are many participatory design methods, including sittings, workshops, stories, photography, dramas, and games (Muller 2003). We were inspired by low-fidelity prototyping methods (Beaudouin-Lafon and Mackay 2002) because they provide concrete artifacts that serve as an effective medium for communication between users and designers. In particular, to explore the CENDARI virtual research environment design space, we used brainstorming and video prototyping. Together, these two techniques can help explore new ideas and simulate or experience new technology.

For brainstorming, a group of people, ideally between three and seven in number, is given a topic and a limited amount of time. Brainstorming has two phases: an idea generation phase and an evaluation phase. The aim of the first stage is to generate as many ideas as possible. Here, quantity is more important than quality. In the second stage, ideas are reflected upon and only a few selected for further work (e.g., video prototyping). The selection criteria could be a group vote where each person picks his or her favorite three ideas.

Video prototyping is a collaborative design activity that involves demonstrating ideas for interaction with a system in front of a video camera. Instead of describing the idea in words, participants demonstrate what it would be like to interact with the system. The goal is to be quick and to capture the important features of the system that participants want to be implemented. A video prototype is like a storyboard: participants describe the whole task including the user's motivation and context at the outset and the success of the task at the end. In this way, the video prototype is a useful way to determine the minimum viable system to achieve the task successfully.

We organized three participatory design sessions (Boukhelifa et al. 2015) with three different user groups: WWI historians, medievalists, and archivists and librarians (Figure 2). The aim of these sessions was to understand how

different user groups would want to search, browse, and visualize (if at all) information from archival research. In total, there were 49 participants (14 in the first, 15 in the second, and 20 in the third). Each session was run as a 1-day workshop and was divided into two parts. The morning began with presentations of existing interfaces for access and visualization. To brainstorm productively in the afternoon, participants needed to have a clear idea of the technical possibilities currently available. In the afternoon, participants divided into three to five groups of four and brainstormed ideas for searching, browsing, and visualization functions, and then they created paper and video prototypes for their top three ideas. There were 30 video prototypes in total, each consisting of a brief (30 seconds to 4 minutes) mockup and demonstration of a key function. Everyone then met to watch and discuss the videos.

Findings. These participatory workshops served as an initial communication link between the researchers and the technical developers, and they were an opportunity for researchers to reflect on their processes and tasks, both those they perform using current resources and tools and those they would like to be able to perform. Even though there were different groups of users involved in these workshops, common themes emerged from the discussions and the prototypes. The outcomes of the participatory sessions were threefold: a set of functional requirements common to all user groups, high-level recommendations to the project, and a detailed description of historians' workflow (Section 3.3).

In terms of functional requirements, participants expressed a need for networking facilities (e.g., to share resources or to ask for help), robust search interfaces (e.g., search for different types of documents and entities or by themes such as language, period, or concept), versatile note-taking tools that take into consideration paper-based and digital workflows and support transcription of documents and annotations, and interactive visualizations to conceptualize information in ways that are difficult in text-based forms.

The participatory design workshops were concluded with the following high-level recommendations to CENDARI. The first recommendation suggested combining existing workflows with new digital methods in ways that save researchers time. In particular, notes can be augmented over time, and researchers' willingness to share them might depend on the note-taking stage and their motivation for sharing. With regard to the second recommendation, researchers have data that they do not use after publication or end of their projects. If CENDARI can offer benefits for making this data available with proper citations, such an initiative could encourage researchers to release their data. This could change working practices and bring more transparency to historical research. Indeed, linking the management of research data to publication and presenting tangible benefits for researchers are important factors in attracting contributors. The third recommendation suggested working closely with researchers to develop testbeds early in the project rather than seek feedback at the end of software development. Researchers who are currently working on a project are likely to have useful data and an interest in sharing it. These projects could be implemented as use cases demonstrating the utility of our virtual research environment.

To create a technical system that can adequately support the preceding functional requirements and recommendations, CENDARI's technical experts took into account the information collected from the participatory workshops and used it as the basis for technical development. As part of the follow-up from these workshops, it was decided to write the important functions demonstrated in the video prototypes in the form of use cases. An example of such use cases is described at the beginning of Section 3.

3.3 Historian Research Workflow

Through the discussions and exchanges gathered during the participatory design sessions and standard references (Iggers 2005), CENDARI technical partners identified a workflow that seems to match the work of a broad range of historians in the early phases of their research, although we do not claim that every historian follows it. We summarize it here to refer to it later concerning the tools and mechanisms that the infrastructure supports:

(1) *Research preparation*: All historians start a new project by gathering questions and hypotheses, and possible ways of answering or validating them. This phase is usually informal and carried out using a notebook, either paper based or computer based. During the participatory design sessions, researchers complained that the notes they take are hardly organized, often spread over many computer files or notebook pages.
(2) *Sources selection*: To answer the questions and validate the hypotheses, researchers gather books, articles, and Web-accessible resources, then make a list of primary and secondary sources to read. This stage is repeated each time new information is collected.
(3) *Planning of visits to archives and libraries*: Relevant sources are often accessible only from specific archives or libraries, because the sources are unique; even the catalogs are not always accessible online, or not precise enough to answer questions. Physical visits to the institutions are then necessary.
(4) *Archive and library visit*: Working at archives or libraries involves note taking, transcribing, and collecting scans and photos of documents for later exploitation. Sometimes even more work is involved when archive boxes have never been opened before and some exploration and sorting is needed, hopefully (but not always) improving the catalogs.
(5) *Taking notes*: During their visits in archives and libraries, historians take notes or annotate copies of archival records. These notes and annotations generally follow the main research interests of the researchers, but they also contain side glances to related topics or possible research areas. Most of the time they can be understood as in-depth descriptions of the sources and thus as a complement to the metadata available in finding aids.
(6) *Transcription*: Primary sources consulted are often transcribed, either partially or exhaustively to facilitate their reading but also to enhance their searchability. These transcriptions serve as the basis for literal citations in publications.
(7) *Research refinement and annotation*: From the information gathered, some questions are answered and some hypotheses are validated or invalidated, but new questions arise, as well as new hypotheses. In particular, this list of questions comes repeatedly with regard to the following facts. First, who are the *persons* mentioned in the documents. Some are well known, whereas others are not and yet are frequently mentioned. Understanding who the persons are and why they were involved is a recurring question in the research process. Second, where are the *places* mentioned? Finding them on a map is also a recurring issue. Third, what are the *organizations* mentioned in the documents and their relationship with the persons and places? Fourth, clarifying the temporal order of events is also essential, and dates often appear with a high level of uncertainty.
(8) *Knowledge organization and structuring*: After enough facts have been gathered, historians try to organize them in high-level structures, with causalities, interconnections, or goals related to persons and organizations. This phase consists of taking notes as well, but also in referring to other notes and transcriptions, and by organizing the chunks of information from previous notes in multiple ways, not related to the documents where they come from but rather from higher-level classifications or structures.
(9) *Refinement and writing*: At some point, the information gathered and organized is sufficient for writing an article or a book. All bibliographic references are gathered, as well as the list of documents consulted in archives and libraries, to be referenced in a formal publication.
(10) *Continuation and expansion*: Often a research work is reused later either as a continuation of the initial research or with variations reusing some of the structures (other places, other organizations, other times).
(11) *Collaboration support*: Although collaborations are possible and realized in the physical world, the sharing of gathered material is limited due to the difficulty of copying every physical information between collaborators. In contrast, a Web-based setting allows the sharing of all resources related to a research project. Historians and archivists have expressed the desire, during all of the participatory design sessions, to experiment on digital-based collaborations.

Archivists do not follow the same workflow, but they share some of the same concerns, particularly the need to identify persons, places, organizations, and temporal order, as this information is essential to their cataloging activity. They also have a higher-level engagement with the material that is useful in making sense of institutional logic or just understanding the contents and provenance of particular boxes and collections. We note that the workflow described previously is nonlinear, as reported in other studies (e.g., Mattern et al. (2015)). At any stage, researchers can acquire new information, reorganize their data, refine their hypotheses, or even plan new archive or library visits.

The CENDARI platform has been built to support this workflow with Web-based tools and to augment it with collaboration, sharing, and faceted search through the gathered and shared documents. The main functional requirements can be summarized as taking notes, transcribing, annotating, searching, visualizing, and collaborating.

3.4 Iterative Approach and Support Technologies

The software development in the project was carried out by adopting agile development methods. The software components were developed in short iterative release cycles, and direct user feedback was encouraged and incorporated. The software is released as open source, and the code is hosted on and maintained through GitHub (CENGitHub 2016). Where possible, CENDARI used the existing solutions provided by DARIAH, such as the JIRA (https://www.atlassian.com/software/jira) ticketing system. Researchers who contacted relevant institutions for inclusion of their holdings into the platform used it to support and document the workflow from the first contact through negotiations and agreements for data sharing to the final ingestion. Following the positive experience of historians using the tool, JIRA remained the obvious choice for bug and issue tracking during the software development. By tightly integrating the applications with the DARIAH development platform and establishing close communication between historians and developers, all parties involved in the development process were able to engage in direct and problem-oriented development cycles. Ultimately, the release cycles were reduced to 1 week and included dedicated test and feedback sessions, as well as development focusing on small numbers of issues that were identified and prioritized collaboratively in the group.

One of the decisions that enabled these rapid cycles was the adoption of a DevOps model to manage and develop the CENDARI infrastructure. By applying methods of agile development to system administration and simultaneously combining the development and management of the applications and infrastructure, as discussed, for example, in Kim et al. (2016), the effort and amount of required work from code commit to deployment was dramatically reduced. This was achieved by implementing automated code build and deployment using the Jenkins CI (https://jenkins.io/) platform and Puppet (https://puppet.com/) configuration management on dedicated staging and production servers.

4 THE CENDARI INFRASTRUCTURE

The CENDARI infrastructure combines *integration* of existing components and tools, *extension* of other components and tools, and *tailored development* of the missing pieces. The overall goal was to offer a great user experience to the researchers, part of which was already implemented by existing tools, while avoiding development of an infrastructure from scratch.

Finding a set of tools that can be either developed or combined to form an integrated environment was a challenge. We realized that to address the user requirements, it was necessary to provide multiple tools to support several stages of the researchers' workflow, and the production of research outputs and formats, as no single tool could offer all of the required features. The challenging part was to select and decide upon a limited number of components, technologies, and tools that users could use intuitively and without extensive training.

To implement this infrastructure, we pursued a modular approach. In using existing tools and services, we were able to offer some major features. At the same time, several important parts of the infrastructure (Figure 3) were developed especially for CENDARI.

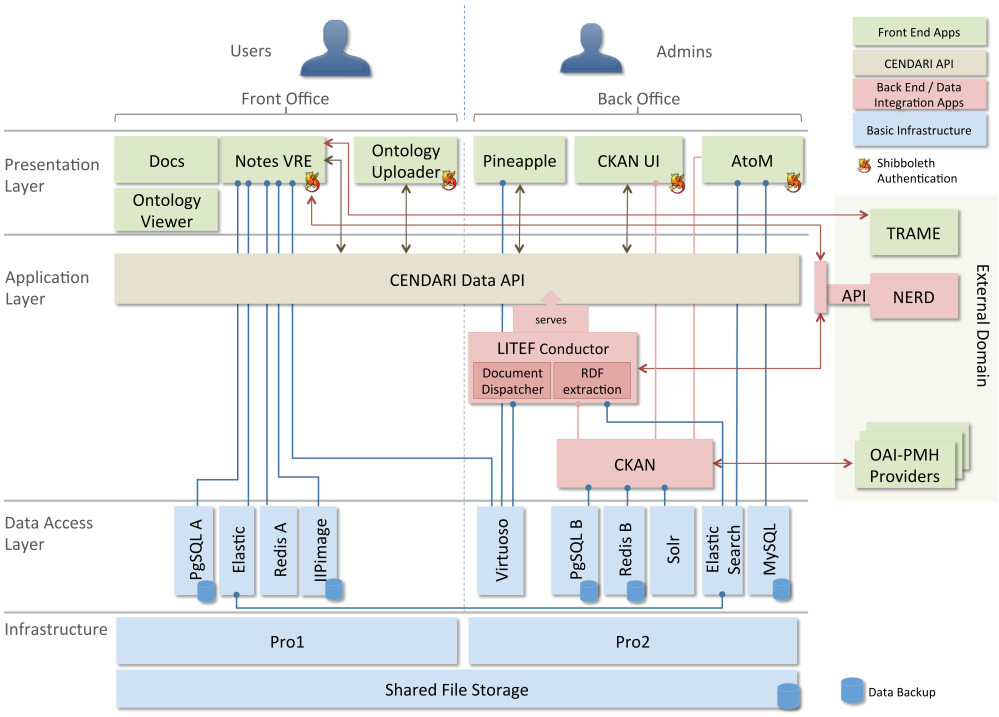

Fig. 3. CENDARI infrastructure model, originally by Stefan Buddenbohm.

4.1 The Backend

In the application layer, the central data exchange and processing is carried out by the Data API, a component implemented within the project (Section 6.2). The data is stored in the repository, based on CKAN, and can be accessed through a Web browser. Within the project, several extensions to CKAN were developed to support the data harvesting and user authentication mechanisms. The development of the Data API and the Litef (Section 6.2.1) focused on providing specific data services required for the project. The NERD service (Lopez 2009) (Section 6.2.3) provides identification and disambiguation of entities from text.

At the data persistence level, we take a polyglot approach: relational databases such as PostgreSQL (http://www.postgresql.org) and MySQL (https://www.mysql.com) for most Web-based tools, Elasticsearch for our faceted search, and the Virtuoso (https://virtuoso.openlinksw.com) triple store to manage generated semantic data and triples created by the NTE and the Litef.

4.2 The Frontend

Several user applications in CENDARI support the historian's workflows: the NTE, the Archival Directory (CENArch 2015) and the *Pineapple Resource Browser* tool (https://resources.cendari.dariah.eu/).

The NTE, described in detail in Section 5, combines access to the faceted search and the repository data with individual research data. The NTE is an adaptation and extension of the EditorsNotes system (http://editorsnotes.org). The Archival Directory, based on the AtoM software, is used for manual creation and curation of archival descriptions. It has a strong transnational focus and includes "hidden" archives and institutions "little known or rarely used by researchers" (CENArch 2015). At present, it offers information about more than 5,000 institutional and archival descriptions curated by CENDARI researchers. Pineapple provides free-text search and faceted browsing through our KB, containing resources and ontologies from both domains (WWI and medieval).

Technically, Pineapple is a SPARQL client, sending predefined parameterized queries to Virtuoso to extract and render information about available resources, related entities such as people, places, events, or organizations, or resources that share same entity mentions from different data sources. These resources are generated by the semantic processing services (Section 6.2.4) or are integrated from the medieval KB TRAME (Trame 2016) that provides the search service for distributed medieval resources. Pineapple offers a Web-based user interface and uses content negotiation (https://www.w3.org/Protocols/rfc2616/rfc2616-sec12.html) to provide a REST-based (non-SPARQL) interface to the KB, delivering JSON formatted data. Advanced and ontology-savvy users can use the *Ontology Viewer* tool or upload their own ontologies via the *Ontology Uploader*.

4.3 Integrated Infrastructure

The main applications were implemented as shown in Figure 3, which provides a high-level overview of the infrastructure design, distinguishing between the internal application and data layers and the user facing presentation layer vertically. The model also shows the distinction between the front office and back office, split over two machines, separating the common user applications from those components used by power users only. This modular approach is exemplified by the role of the Data API as a central component for data exchange between the user facing applications and the repository, including user authorization on the resource level. All user-facing applications communicate through the Data API. Thus, by adopting the API, any underlying repository solution could be used.

All CENDARI services authenticate users through the DARIAH AAI (https://wiki.de.dariah.eu/x/JQurAg), an authentication solution based on Shibboleth (https://shibboleth.net), for which dedicated plugins for AtoM and CKAN were also created. The DARIAH AAI handles the verification of users and the creation of user accounts including support for password management. From an end user's perspective, this is a single-sign-on experience, where CENDARI components are visible as one of the many DARIAH services.

CENDARI was designed from the start as an infrastructure that can ultimately be sustained and reused through DARIAH. To achieve this, we used tools and solutions offered by DARIAH or other parties, then integrated the services into the existing ecosystem where possible, such as the development platform, the AAI, CKAN, and AtoM. Additionally, the technical design of the infrastructure was aligned with the efforts undertaken in parallel to migrate the TextGrid infrastructure into the DARIAH ecosystem.

5 NOTE-TAKING ENVIRONMENT

Most historians collect notes in files using standard text editors. The files are sometimes well sorted in folders, but most of the time the files are difficult to find due to the loose organization of early research projects. The information related to the notes is also scattered in multiple locations. Sometimes historians take pictures at archives or libraries; these pictures end up being stored in whatever location their camera or system stores them. The file names associated with these pictures are also unrelated to the notes or research projects. Even if some historians are well organized, it takes them considerable time and discipline to organize their virtual working space on their own computer. And even with the strongest discipline, it is almost impossible to link the multiple documents together, connecting notes to pictures, scans, PDF files, spreadsheets, or other kinds of documents they use. With all of these problems in mind, and to facilitate the collaboration between researchers, we have designed and implemented the NTE in tight integration with the whole CENDARI infrastructure. The NTE implements the historian's workflow described in Section 3.3.

5.1 Overview

The NTE is designed to manage documents and notes gathered for a project. Typically, a project is a thesis, or a journal article, but more generally it is a container for gathering and linking information, refining it, collaborating, and preparing publications. The final publication or production is not carried out inside the NTE, as there are already many editing environments to perform that task.

The main user interface of the NTE has three main panels coordinated using brushing and linking (Becker and Cleveland 1987): the search and browse panel (Figure 1(A)); a library where the user can manage projects and browse allocated resources (Figure 1(B)); a central space for editing, linking and tagging resources (Figure 1(C)); and a visualization space for showing trends and relationships in the data (Figure 1(D)).

The resources panel. In the resources panel, resources are organized per project into three main folders corresponding roughly to the way historians organize their material on their machines. The notes folder contains files, and each file is a note describing archival material related to a project. The user can select a note, and its content is shown in the central panel. The user can edit the note and tag words to create entities such as event, organization, person, place, and tag. They can add references to documents, which can be letters, newspaper articles, contracts, or any text that acts as evidence for an observation or a statement in a note. Documents can contain named entities, a transcript, and references to other documents and resources, as well as scanned images, which are displayed in the high-resolution Web-based image viewer at the bottom of the central panel.

The central panel. The central panel acts as a viewing space for any type of resource and mimics the function of a working desk of a historian. This is where entity tagging and resolution takes place. The user may not be entirely clear about the true identity of an entity, such as in the case of a city name that exists in different countries. The user has an option to manually resolve it by assigning a unique resource identifier (URI) (e.g., to a unique Wikipedia entry).

The visualization panel. The visualization panel provides useful charts to show an overview of entities, distributions, frequencies, and outliers in the resources of the project, and a map that shows the location of any *place* entity.

In summary, the NTE supports the following features: (1) *editing and annotation* through a rich set of editing, formatting, and tagging options provided by RDFaCE (http://rdface.aksw.org/); (2) *faceted search* for thematic search and access to resources; (3) *visualization*, showing histograms of three entity types (names, places, and events) and a geographical map with support for aggregation (other visualizations could be easily integrated); (4) *automatic entity recognition* in the form of an integrated NERD service in the editor; and (5) *interaction*, in which the visualizations support selection, highlight, and pan and zoom for the map. Brushing and linking is implemented with one flow direction for consistency from left to right. This is to support the workflow in the NTE: select resource, view and update, and then tag and visualize. Note that additionally, referencing, collaboration, and privacy settings are available in the NTE. In terms of privacy settings, notes are private and entities are public by default, but users can change these permissions.

5.2 Technologies

The NTE implements a client-server architecture, with the client side relying on modern Web technologies (HTML5, JavaScript, D3.js) and the server on the Django Web framework (https://djangoproject.com/). Faceted browsing and search functionalities are implemented using Elasticsearch, which unifies the exploration of all resources provided by the project. Django provides a rich ecosystem of extensions that we experimented with to support, for example, Elasticsearch and authentication. As we needed a tight integration to communicate precisely between services, we resorted to implementing our own Django extensions for faceted-search support, indexing and searching with Elasticsearch, access to Semantic Web platforms, and support for very large images through an image server. Although the principles behind all of these services are well known, the implementation is always unexpectedly difficult and time consuming when it comes to interoperability and large datasets.

One example of scalability relates to the faceted search, which allows searching a large set of documents through important types of information or "facets.". A search query is made of two parts: a textual query as search engine support, and a set of facet names and facet values to restrict the search to these particular values. For example, a search can be related to a date range of 1914 to 1915 (the *date* facet with an interval value), and a

person such as "Wilhelm Röntgen" (a *person* facet restricted to one name), plus a query term such as "Columbia." The result of a faceted search is a set of matching documents, showing snippets of text where searched terms occur, and a list of all facets and all facet values appearing in the matching documents (or the 10 most frequent facet values to avoid saturating the user interface). We defined 10 facet types: *application* where the document was created, the document's *creator*, *language* used, name of the research *project* holding the document, and mentions of dates/periods (*date*), organization names (*org*), historical persons (*person*), geocoordinates (*location*), places (*place*), and document identifiers (*ref*) such as ISBN and URL.

Elasticsearch allows defining a structure for searching (called a *mapping*) and provides powerful aggregation functions to support scalability. For example, for each query, we receive the list of matching geographical locations and dates; if we show one point per result, the visualizations are overplotted and the communication between the server and the Web client takes minutes to complete. A typical historical project easily references 10,000 locations and thousands of names. Searching over many projects multiplies the number. Furthermore, CENDARI also provides reference datasets such as DBPedia, which defines millions of locations and dates. Therefore, our faceted search relies on Elasticsearch aggregation mechanisms, returning ranges and aggregates. Locations are returned as geohash identifiers with a count of matches in each geohash area, which enables visualizing results quickly at any zoom level. Dates are also returned as intervals with the number of documents for the specified interval.

The NTE fulfills its role of editor, image viewer, and search interface at scale. It currently serves about 3 million entities and approximately 800,000 documents with a latency around 1 to 10 seconds depending on the number of users and complexity of the search queries.

6 DATA INTEGRATION AND SEMANTIC SERVICES

The primary objectives of the data integration platform (DIP) were to integrate relevant archival and historical content from disparate sources into a curated repository to provide tools for describing and integrating hidden archives and to implement semantic services for inquiring and interlinking of content. The DIP directly or indirectly supports several stages in the historian's research workflow (from Section 3.3): selection of sources (2), planning to visit/visiting archives and libraries (3), knowledge organization and structuring (8), research refinement and annotation (9), and searching through relevant material. It contributes to CENDARI's inquiry environment by offering new ways to discover meaning and perform historical research (CENDARI 2015). It ensures that data from external sources remains available in the exact format and version in which they have been used for the research in the first place, thus contributing to the reproducibility of the research. It preserves the original data and its provenance information and sustains the derivatives from the data processing, transformation, or data modification operations, ensuring data traceability.

To create a representative and rich pool of resources related to the modern and medieval history, the team identified and contacted more than 250 cultural heritage institutions. We encountered a significant diversity among institutions in the level of digitization and digital presence. Some institutions provide excellent digital access, whereas others are still in the process of digitizing their original analog finding aids. It should be noted that neither the digital presence itself nor the existence of digitized material guarantees that the material is publicly available and accessible outside of the institution's own internal system. Furthermore, differences exist among institutions' data provision protocols (when available).

To address these challenges and to access and harvest the content from different institutions, we had to establish a flexible and robust data acquisition workflow, confronting at the same time legal, social, and technical challenges as described in detail in Edmond et al. (2015a). Our process is consistent with the FAIR data principles, designed to make data *findable*, *accessible*, *interoperable*, and *reusable* (Wilkinson et al. 2016). Harvested data is preserved in the original form, enriched during processing (see Section 6.2), and further interlinked based on the enriched and extracted information. Data is retrievable by a CENDARI identifier, along with its data origin

(e.g., the provider institution or the original identifiers). The NTE and Pineapple provide search and browse functionalities for end users, whereas the Data API exposes data in a machine-readable fashion. When data is processed or enriched, DIP makes sure that a full log of applied transformations and final outputs are preserved and FAIR. An exception to some of the principles concerns private research data, as we decided to balance transparency and confidentiality for metadata extracted from researchers' notes. Additionally, the FAIRness of the integrated data at large also depends on the original data sources.

6.1 Incremental Approach to Data Integration

Data in CENDARI originates from archives, libraries, research institutions, researchers, or other types of contributors of original content, including data created in CENDARI. The data has the following characteristics in common: variety of data licenses and usage policies; heterogeneous formats, conventions, or standards used to structure data; multilingualism; and diverse granularity and content quality. In addition, initial prospects suggested that CENDARI would have to accommodate data beyond just the textual and include audio and video materials, thus building an infrastructure with high tolerance for such heterogeneity.

This scenery leads to the term *data soup*, defined as "a hearty mixture of objects, descriptions, sources and XML formats, database exports, PDF files and RDF-formatted data" (Edmond et al. 2015a). From a higher-level perspective, the data soup comprises raw data, KB data, connected data, and CENDARI-produced data, which require different data management approaches (Edmond et al. 2015b). A mapping to a common data model (as applied in most data integration approaches) would not be possible or preferred for several reasons: *lack of a priori knowledge about data* (plenty of standards or custom data formats), and often standard formats were brought in multitude of flavors, sometimes even with contradictory semantics (e.g., "creator" was used both as an author of an archival description or as a person who wrote a letter to another person); *nonexistence of a widely accepted domain data model*, noting that WWI and medievalist groups had different requirements and perspectives on data granularity. The development of a single, widely accepted new schema (or extension of an existing one) takes time and does not guarantee flexibility for future modifications of the schema and the tools. New data acquisitions may impose new changes in the metadata schema, which, apart from modifying the tools, causes further delays to the data integration scenarios. It also increases the risk of incompleteness, as data would be limited only to the structure supported by a single schema, and thus a resource not fitting the schema would have to be omitted.

Even if CENDARI would have established an all-inclusive new metadata schema, it would still have not guaranteed that it will serve researchers' needs, ensure their transnational and interdisciplinary engagement, and provide an inquiry-savvy environment. Such a schema would either be highly generic and comprehensive, thus defeating the purpose of having a schema, or it would be too specific, thus failing to fulfill the needs of both current and future researcher groups.

Consequently, it was necessary to adapt a lean model to the data integration, avoiding thorough domain modeling until a better understanding about user requirements and scenarios develops. The resulting system should enable integration of data processing components in a pay-as-you-go fashion (Hedeler et al. 2013), deliver early results and corrective feedback on the quality and consistency of data, and perform refinement and data enrichment in an incremental rather than a prescribed way.

For this purpose, we adopted an approach combining two conceptual frameworks: *dataspace* (Franklin et al. 2005) and *blackboard* (Hayes-Roth 1985). This allowed us to separate the data integration process from the data interpretation and developments of domain-specific application models and tools (Edmond et al. 2015a, 2015b). The dataspace promotes coexistence of data from heterogeneous sources and a data unification agnostic to the domain specifics. Such a system contributes to creating and inferring new knowledge, and benefits from the "dirt in data" (Yoakum-Stover 2010) by preserving the information in its original form, enabling serendipitous discoveries in data. The blackboard architecture is often illustrated with the metaphor of a group of specialists

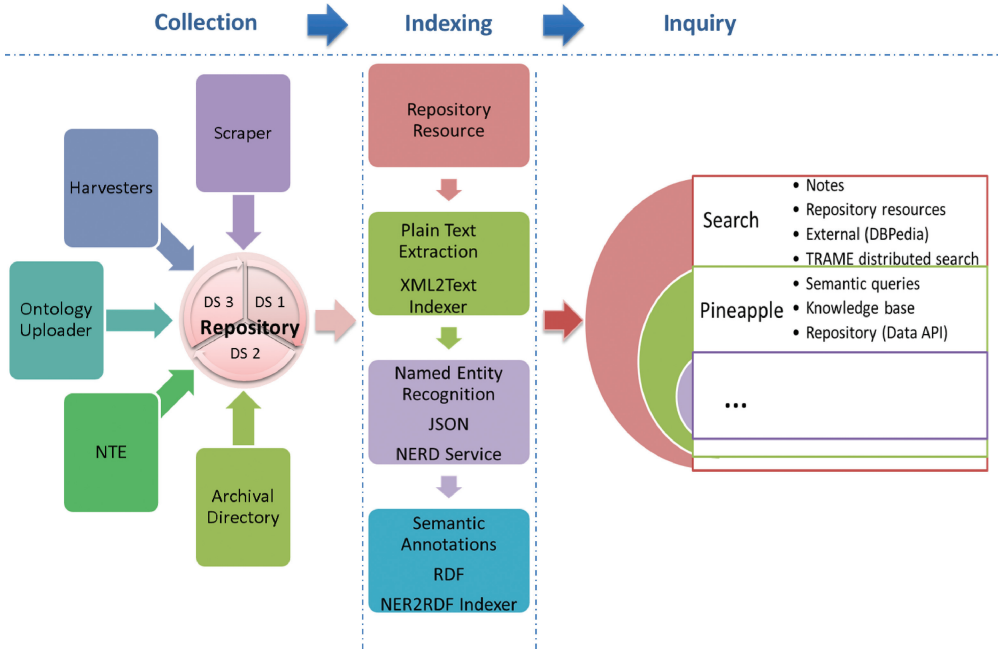

Fig. 4. Data integration workflows within the CENDARI infrastructure (Edmond et al. 2015a).

working on a problem (Hayes-Roth 1985), sharing their solutions to the "blackboard" and reusing them to develop further solutions until the problem is solved. We applied a slightly modified blackboard model, a specialist being a data processing service (agent) triggered by a data insertion or update; the processing service produces an output that may be used by other services and components; and additionally, it specifies the certainty score of the output, which is used to filter out the less optimal results.

6.2 Data Integration and Processing Components

CENDARI workflows are grouped around three major processes: collection, indexing, and enquiry (Figure 4) (Edmond et al. 2015a). The DIP is essential for the data collection and indexing and integrates services to transform and process data for searching and semantic inquiries by the end users (CENTools 2016).

The *repository* is a central storage for harvested data and CENDARI-created content. It assigns basic provenance information to the data and ensures that data is versioned. In addition, it keeps the overall authorization information at a single place. The repository implements the dataspace model and has no understanding of the data semantics.

The *data acquisition components* support the collection process by harvesting and ingesting content into the repository. They vary from dedicated harvesters for content providers APIs, to a Scraper service, which extracts structured data directly from the Web pages of selected providers, in absence of other data interfaces. Additionally, a special data synchronization component was developed to transfer the archival directory data into the repository.

The *KB* is the central node for CENDARI's domain knowledge, based on the Virtuoso triple store. Whereas repository data represents file-based resources, the KB is where historical knowledge, formalized through ontologies, vocabularies, or user annotations, is persisted in a schema free but still structured form. The KB is populated from several strands: acquisition of existing structured domain ontologies, acquisition of

knowledge from collected nonstructured/non-RDF resources through semantic data processing, or via the NTE user annotations. The KB supports the dataspace model, implements an RDF named graph for each resource, and provides authorization features during the research refinement, annotation, and knowledge organization processes. The KB is accessed through the NTE and Pineapple.

The *Data API* (https://docs.cendari.dariah.eu/developer/litef-conductor/docs/rest.html) is a REST interface that provides machine-readable access to all data in the repository accordingly to the user permissions. It is agnostic toward semantically rich descriptions and is primarily aware of the dataspace, provenance, authorization, and data format. The Data API additionally provides unified access to all derivatives of the original data in the repository, generated through the processing services.

Data processing components enlist a document dispatching service and several internal or external services for resource indexing, acting as a modified blackboard system. These are the Litef, TikaIndexer, NERD services, semantic processing services, and VirtuosoFeeder and ElasticFeeder services. The following sections provide an overview of these services.

6.2.1 Litef Conductor. Litef implements a dispatching mechanism for invocation of integrated data processing services, either developed by third parties or by CENDARI. It separates the (general or domain-specific) data processing from the internal organization of data, authentication, and access permissions, and avoids deep internal dependencies between participating data processing services. Litef reacts on the addition or update of a resource in the repository and passes it to all interested data processing services. The results of the processing and their log are stored in the file system and available via the Data API in read-only mode.

For each data processing service, Litef implements a small plugin, an indexer, which informs Litef about the types of resources it is interested in and the type of result it produces. This is a simple concept but is expressive enough to define even complex processing flows. The plugin-based indexing architecture ensures that the system can be extended with new processing services, either to support additional formats or to perform other specific data processing in the future.

6.2.2 Plain-Text Indexing, Metadata Extraction, Indexing for Search. Common processing steps for most of the resources are plain-text extraction and metadata extraction. These have been integrated in Litef as a standalone Java library named *TikaExtensions* (https://github.com/CENDARI/tika-extensions), based on the Apache Tika toolkit (https://tika.apache.org/). The library implements dedicated parsers for most frequent and preferred formats.[1] For other formats, default parsers were used, providing less precise extraction. Litef transforms the parsed output from the library and generates several resources: a *plain-text* file containing extracted textual selection, a *key-value pairs* file serializing the output from the library, and a *search index document* in JSON format, in accordance with the defined search facets (see Section 5.2). Where possible, a link to the original location of the resource is provided. The ElasticFeeder service then recognizes the newly generated index document and sends it to the ElasticSearch service, integrated by the NTE, enabling search across all resources, independent from the tool where the resource was originally created.

This approach allows us to separate the metadata extraction logic from Litef internals and to iteratively improve it in a pay-as-you-go fashion, as our knowledge about data developed. Furthermore, it allows us to reuse a wide variety of already available Tika-based metadata extraction plugins, to customize them or integrate new parsers. Note that the library can be easily reused outside of the CENDARI context.

6.2.3 NERD Services. The participatory design workshops showed that historians are mostly interested in identifying places, persons, events, dates, and institutions in archival material, annotating them as entities and linking them to the resources where these terms originally appeared. To support the entity identification over a large corpus of data, the NERD service was used for automatic entity extraction from preprocessed plain-text

[1]EAD/XML, EAG/XML, EDM RDF, METS/XML, MODS/XML, OAI-PMH records, TEI/XML.

content. Two NERD services were developed in the project, one for the English language (Lopez 2009) and another for multiple languages (Bulgarian, German, Greek, English, Spanish, Finnish, French, Italian, Ripuarisch Platt, Latin, Dutch, Serbian (Cyrillic), Swedish) (Meyer 2016). Both services expose very similar REST APIs and provide JSON-formatted outputs of the recognized entities and the confidence of the result. Although their entity recognition methods vary, they both use Wikipedia for entity disambiguation.

6.2.4 Semantic Data Extraction and Processing. For each resource in the repository, Litef creates a semantic representation as a named document graph with extracted metadata added as properties of the resource within that graph. Depending on the processing results of the NERD services, semantic representations will be created for entity resources of type person, place, event, and organization, following the CENDARI ontology structures. For the most common formats, such as EAD and EAG, more accurate semantic mapping is performed. All outputs of the semantic processing services are persisted in the KB.

6.3 The Development of the CENDARI KB

Several ontologies (see conceptualizations in Gruber (1993)) were used within the CENDARI Knowledge Organization Framework developments. These vary from metadata schema for describing archival institutions, through controlled vocabularies and gazetteers, to domain ontologies structuring knowledge about both supported research domains.

We created extensive ontology development guidelines (CENOnt 2014), focusing on the reuse of an existing ontology element set and suitable instances for each domain. In a joint workshop, researchers from both domains identified similar types of concepts and entities to be represented, broadly fitting within the Europeana data model (EDM) classes (http://pro.europeana.eu/page/edm-documentation): Agent, Place, Timespan, Event, and Concept. Domain differences could be accommodated through EDM extensions, allowing a finer level of granularity while enabling unification at a courser level of granularity and fostering the data interoperability. For example, to coincide with the 100th anniversary of the start of WWI, many research projects published WWI data. However the format and the quality of this data varied considerably: Excel spreadsheets, data scraped from Web pages, and as a badly formed single large RDF file. We implemented custom solutions to transform relevant existing vocabularies into an appropriate EDM extension. Transformed ontologies were aligned to provide better integrated coverage of the domain than was provided by any single ontology on its own. Due to the nature of data, we used a terminological approach to the alignment (for other approaches, see Shvaiko and Euzenat (2013)), more specifically, a character-based similarity measure and the I-SUB technique (Stoilos et al. 2005). The concepts in the ontologies also made use of MACS (Clavel-Merrin 2004) to facilitate multilingual access to the resources for English, German, and French.

Transformed ontologies form a large part of the KB, along with the data automatically generated by the DIP. Data can be browsed and searched through Pineapple, providing a unified view over transformed ontologies; the semantic data extracted from heterogeneous archival resources; and medieval manuscript ontologies, including visual navigation through the text and the structure of the medieval manuscripts. As an example, for an entry about "Somme," Pineapple will deliver entries from DBpedia, WWI Linked Open Data (http://www.ldf.fi/dataset/ww1lod/), and 1914-1918 Encyclopedia (https://docs.cendari.dariah.eu/user/pineapple.html) ontologies (https://repository.cendari.dariah.eu/organization/cendari-ontologies). By navigating to one of them (e.g., the "Battle of Somme" event), more information about the event and potentially associated archival resources is displayed. For the latter, more details are available, such as the extracted text from the resource raw data; generated mentions of organizations, places, periods, and persons; or other semantically related archival resources.

Transformed domain ontologies were published in the repository with the *Ontology Uploader*, which creates additional provenance metadata, based on the OAI-ORE resource maps model (http://www.openarchives.org/ore/1.0/datamodel), expressing the relationships between the original and transformed data.

Element set ontologies developed or adapted within the project are published in raw format on GitHub (CENGitHub 2016) and available for visual exploration through the WebVowl tool (http://vowl.visualdataweb.org/webvowl.html). A smaller portion of the KB was created by researchers, through the NTE. They were primarily focused on identifying and tagging entities within notes and uploaded documents, and resolving them against DBPedia (see Section 5). This knowledge was captured according to the schema.org general-purpose ontology built into the original software used for the NTE. Provisioning of a richer semantic editor or annotator tool to support user-friendly ontology developments by researchers, along with relations between entities, notes, and any other archival resources, proved to be very challenging. Further developments required a strong balance between the flexibility of the tool and the simplicity of the user interface, which was deemed to be beyond the scope of the CENDARI project.

7 DISCUSSION

The goal of the CENDARI project was to support historical researchers in their daily work. This goal is relatively original in the DH landscape and required a lot of experiments and a complicated infrastructure. Although the goal has been reached technically, we only knew at the end of the project what essential components were needed and the capabilities and digital literacy of historians in general. We try to summarize here the positive outcomes of the project and some of the pitfalls.

It was very clear at the end of the project that historians benefit immensely from supporting tools such as the one offered by CENDARI. Data management at the historian level has become a technical nightmare without proper support. As discussed in Section 3, CENDARI benefited from the feedback offered by historians to express their workflows and needs; to our knowledge, they were never explicitly stated before in such a constructive way. Even if our workflow does not support all of the research steps performed by historians, it supports a large portion of them.

The infrastructure to support faceted search and semantic enrichment is very complex. Is it worth the effort when large companies such as Google and Microsoft are investing in search engines? We believe that historians need more specific support than what search engines currently offer. It may sound banal, but not all data is digital; rather, the opposite is true, with the largest part of historical data neither digitized nor available via metadata. Modern search tools are not meant for these goals and should be augmented by more focused tools. Our ingestion and enrichment tools are complex because they need to deal with multiple types of data and extract useful information out of them to be usable by historians. In addition to the automatic extraction of entities from existing documents and data, it is now clear that a large portion of archive and library data will never get digitized, as acknowledged by all of the institutions with which the CENDARI project interacted. CENDARI has decided to use the data gathered by researchers as a historical resource; we believe that this decision is essential to better support historical research in general. We believe that the "political" decision to keep the historians' notes private by default but to publish the tagged entities publicly is an effective way to spread and share historical knowledge with little risk of disclosing historians' work.

We realized that there is a tension between the capabilities that technology can provide and the digital training of historians to understand and use these capabilities. Our user interface supports some powerful functions simply, but there are limitations. For example, allowing manual or automatic tagging of notes and transcriptions was perceived as very useful and important, but we only support seven types of entities because the user interface would become very complicated if we needed to support more. Allowing more complex enrichment is possible using the RDFace editor that we provide, but it has seldom been used because of the complexity of manually tagging text through ontologies.

We gathered knowledge about how researchers work and some typical workflows. We documented all of these and started a co-evolution between our tools and the historians' capabilities and needs, but more iterations became necessary: our agile development processes, as outlined in Section 3.4, allowed us to perform short

iterations, allowing code deployments to production in mere hours before evaluation workshops. But conversely, this added additional complexity at every step. However, some level of simplification is always needed to reach a new target audience, such as historian researchers. Early integration of the infrastructural components is essential to ensure timely feedback and reduce friction later, but in a setup of parallel and distributed development efforts with several teams on individual development iteration schedules, efficient and clear communication among all participants is a crucial factor to align the work and create a common and collaborative development effort.

By mixing agile methods and long-term planning, CENDARI built a reproducible infrastructure that has since been taken over by DARIAH in an effort to ensure its sustainability and availability for future use by historians and scholars from other disciplines alike.

8 CONCLUSION

Through the functional requirements identified during the participatory design workshops, a firm basis could be laid to support historians in performing their specific workflow. The tools and services provided by the CENDARI research infrastructure favor ordering, organization, search, taking notes, annotations, and transcribing, yet also the analysis of networks and time-space relationships as well as sharing of resources. Researchers are thus supported in drawing together resources and facts about persons, organizations, and structures in the time frame under consideration. These can be pulled together to make patterns visible that cannot be easily taken into focus by classical historiographical methods.

An interesting result of the CENDARI project was the formulation of requirements by historians that support not just their specific research workflow but rather the research process as a whole. The visualizations and the built-in collaboration functionalities of the CENDARI infrastructure—such as the sharing of resources, the establishment of collaborative projects, or the possibility of collaborative writing of articles—seem at first glance secondary to the research process but enhance the analysis of search results and the community of historians in general. This can be seen as the "pedagogical" offer of the infrastructure. Although historians are generally trained to work all by themselves, the infrastructure offers a range of possibilities for collaborative information collection and writing. It thus lays the basis for a truly collaborative and transnational historiography.

Furthermore, the examples resulting from the collaboration of information engineers, historians, and archivists are very promising beyond the achievement of the CENDARI project. The development of ontologies and the possibility of their collaborative enlargement by users can be regarded as a potentially fruitful domain for interaction between the disciplines involved. Another example is the enrichment of metadata by researchers in international standard formats and their handover to the cultural heritage institutions that established and provided the metadata. Quite obviously, an important part of historians' work in archives consists of a deeper description of archival material than the one provided by archivists and librarians, who aim at a much more formal level of description. Provided the interoperability of the data through the infrastructure, enriched metadata shared by users and cultural heritage institutions can be described as a win-win situation for all sides involved. To achieve a cross-domain level of interoperability of data and services, however, syntactic conformance and semantic data per se are not sufficient. Enabling tools for researchers to structure their knowledge and map it across different domains calls for joint efforts in the domain modeling, technical implementation across research infrastructures, training and communication with researchers, and strong research community participation. Could a medieval archaeologist working, for example, with ARIADNE (http://www.ariadne-infrastructure.eu/) benefit from CENDARI? Our answer is positive, but we are aware that additional service-level integration or semantic data–level and ontology developments alignment would be needed for a flawless user experience.

The infrastructure built by the CENDARI project does not support several steps of classical hermeneutic interpretation, which is typical for historians. It can be questioned whether there will ever be tools to support humanists in the specific practices and competencies that mark this profession—the observation and

interpretation of ambivalence and polysemy, of ambiguities and contradiction, and the differentiated analysis of cultural artifacts. The broad range of tools, services, and resources offered by the CENDARI infrastructure underlines the fact that not every need formulated by historians can be satisfied, and a mission creep with respect to requirements must be avoided.

ACKNOWLEDGMENTS

We are thankful to all CENDARI data providers who contributed their content and made it available for research. We would also like to express our sincere gratitude to all CENDARI partners for their great contributions to the development and setup of the infrastructure. The fusion of researchers, archivists, librarians, and IT experts made the CENDARI project a unique learning experience for all of us.

REFERENCES

Michel Beaudouin-Lafon and Wendy Mackay. 2002. Prototyping development and tools. In *Human Computer Interaction Handbook*, J. A. Jacko and A. Sears (Eds.). Lawrence Erlbaum Associates, Hillsdale, NJ, 1006–1031. http://www.isrc.umbc.edu/HCIHandbook/.

Richard A. Becker and William S. Cleveland. 1987. Brushing scatterplots. *Technometrics* 29, 2, 127–142.

Tobias Blanke and Conny Kristel. 2013. Integrating holocaust research. *International Journal of Humanities and Arts Computing* 7, 1–2, 41–57.

Nadia Boukhelifa, Emmanouil Giannisakis, Evanthia Dimara, Wesley Willett, and Jean-Daniel Fekete. 2015. Supporting historical research through user-centered visual analytics. In *Proceedings of the EuroVis Workshop on Visual Analytics (EuroVA'15)*.

CENArch. 2015. CENDARI Archival Directory. Retrieved March 7, 2018, from https://archives.cendari.dariah.eu/.

CENDARI. 2015. Collaborative European Digital Archive Infrastructure. Retrieved March 7, 2018, from http://www.cendari.eu.

CENGitHub. 2016. Cendari Development Repository on GitHub. Retrieved March 7, 2018, from https://github.com/cendari/.

CENOnt. 2014. Deliverable 6.3: Guidelines for Ontology Building. Retrieved March 7, 2018, from http://www.cendari.eu/sites/default/files/CENDARI%20_D6.3%20Guidelines%20for%20Ontology%20Building.pdf.

CENTools. 2016. Deliverable D7.4: Final releases of toolkits. Retrieved March 7, 2018, from http://www.cendari.eu/sites/default/files/CENDARI_D7.4%20-%20Final%20releases%20of%20toolkits.pdf.

Genevieve Clavel-Merrin. 2004. MACS (multilingual access to subjects): A virtual authority file across languages. *Cataloging and Classification Quarterly* 39, 1–2, 323–330.

Jennifer Edmond, Jakub Beneš, Nataša Bulatović, Milica Knežević, Jörg Lehmann, Francesca Morselli, and Andrei Zamoiski. 2015a. *The CENDARI White Book of Archives*. Technical Report. CENDARI. http://hdl.handle.net/2262/75683

Jennifer Edmond, Natasa Bulatovic, and Alexander O'Connor. 2015b. The taste of "data soup" and the creation of a pipeline for transnational historical research. *Journal of the Japanese Association for Digital Humanities* 1, 1, 107–122.

Michael Franklin, Alon Halevy, and David Maier. 2005. From databases to dataspaces: A new abstraction for information management. *ACM SIGMOD Record* 34, 4, 27–33.

Thomas R. Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5, 2, 199–220.

Barbara Hayes-Roth. 1985. A blackboard architecture for control. *Artificial Intelligence* 26, 3, 251–321.

Cornelia Hedeler, Alvaro A. A. Fernandes, Khalid Belhajjame, Lu Mao, Chenjuan Guo, Norman W. Paton, and Suzanne M. Embury. 2013. A functional model for dataspace management systems. In *Advanced Query Processing: Volume 1: Issues and Trends*. Springer, Berlin, Germany, 305–341.

Georg G. Iggers. 2005. *Historiography in the Twentieth Century: From Scientific Objectivity to the PostModern Challenge*. Wesleyan University Press, Middletown, CT.

Gene Kim, Patrick Debois, John Willis, and Jez Humble. 2016. *The DevOps Handbook: How to Create World-Class Agility, Reliability, and Security in Technology Organizations*. IT Revolution Press.

Patrice Lopez. 2009. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries*. Springer, Berlin, Germany, 473–474.

Eleanor Mattern, Wei Jeng, Daqing He, Liz Lyon, and Aaron Brenner. 2015. Using participatory design and visual narrative inquiry to investigate researchers? Data challenges and recommendations for library research data services. *Program: Electronic Library and Information Systems* 49, 4, 408–423.

Alexander Meyer. 2016. mner—Multilingual Named Entity Recognition and Resolution. Retrieved March 7, 2018, from http://136.243.145.239/nerd/.

Michael J. Muller. 2003. Participatory design: The third space in HCI. In *The Human-Computer Interaction Handbook*, J. A. Jacko and A. Sears (Eds.). Lawrence Erlbaum Associates, Hillsdale, NJ, 1051–1068. http://dl.acm.org/citation.cfm?id=772072.772138

P. Shvaiko and J. Euzenat. 2013. Ontology matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering* 25, 1, 158–176.

G. Stoilos, G. Stamou, and S. Kollias. 2005. A string metric for ontology alignment. In *Semantic Web—ISWC 2005*. Lecture Notes in Computer Science, Vol. 3729. Springer, 624–637.

Trame. 2016. Home Page. Retrieved March 7, 2018, from http://trame.fefonlus.it/.

Amanda Visconti. 2016. Home Page. Retrieved March 7, 2018, from http://www.infiniteulysses.com/.

Claire Warwick. 2012. Studying users in digital humanities. In *Digital Humanities in Practice*, C. Warwick, M. M. Terras, and J. Nyhan (Eds.). Facet Publishing in association with UCL Centre for Digital Humanities, London, 1–22.

Bridgette Wessels, Keira Borrill, Louise Sorensen, Jamie McLaughlin, and Michael Pidd. 2015. *Understanding Design for the Digital Humanities*. Studies in the Digital Humanities. Sheffield: HRI Online Publications. http://www.hrionline.ac.uk/openbook/chapter/understanding-design-for-the-digital-humanities.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3, 160018. DOI: http://dx.doi.org/10.1038/sdata.2016.18

Suzanne Yoakum-Stover. 2010. Keynote Address "Data and Dirt." Available at http://www.information-management.com/resource-center/?id=10019338

Received April 2016; revised May 2017; accepted May 2017