CENDARI Virtual Research Environment & Named Entity Recognition techniques

Patrice Lopez, Alexander Meyer, Laurent Romary

CENDARI Virtual Research Environment & Named Entity Recognition techniques

Patrice Lopez **Alexander Meyer** **Laurent Romary**
Inria (Institut national de recherche en informatique et en automatique) & HU Berlin
forename.surname@inria.fr

with support from the Inria Aviz team (Jean-Daniel Fekete et al.)

The CENDARI project

Overview

CENDARI (**Collaborative European Digital Archive Infrastructure**) is a research infrastructure project aimed at integrating digital archives and resources for research on medieval and modern European history.
The project brings together information and computer scientists with historians and existing historical research infrastructures (archives, libraries, other digital projects) to improve conditions for digital historical scholarship. CENDARI has engaged in extensive networking with the archives and libraries of Europe, especially those in Eastern Europe.
CENDARI is a 4-year, European-Commission-funded project led by Trinity College Dublin, in partnership with 14 institutions across 8 countries.

Case study areas

- Medieval culture
- World War I

CENDARI has carried out multiple **participatory design workshops** in order for historians to articulate their needs and wishes regarding a digital research environment, and for the computer scientists to understand those needs. Two major outcomes of those workshops are discussed on this poster:

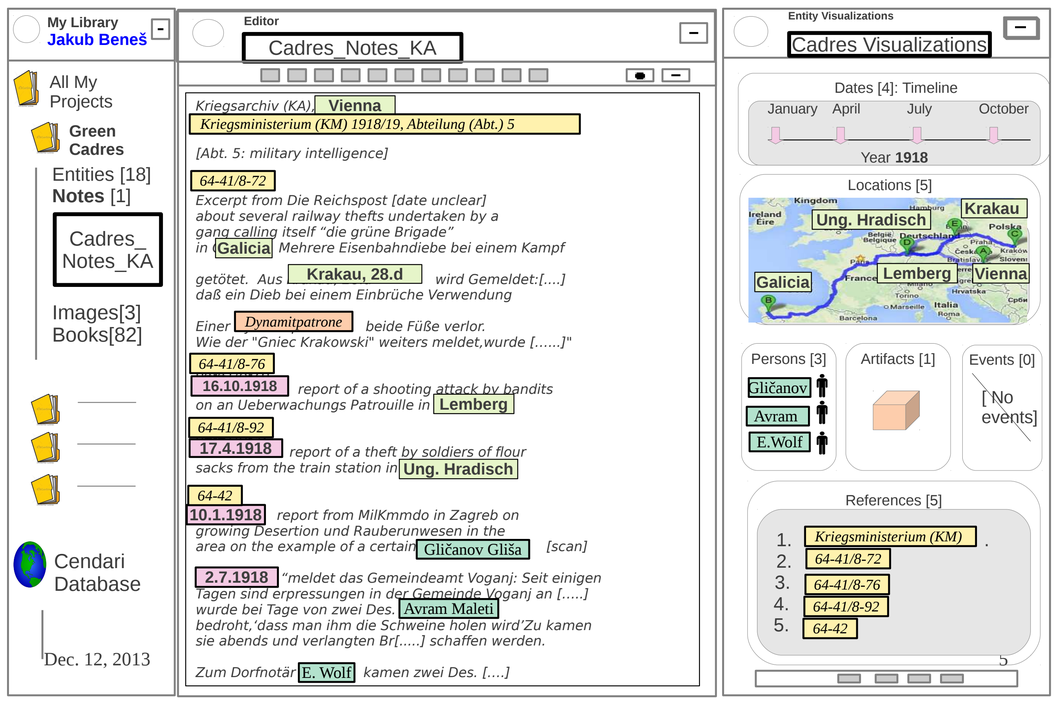
VRE Virtual Research Environment

NER Named Entity Recognition & Resolution techniques
(called from within the VRE)

Dealing with entities

Recognizing and resolving **Named Entities** has been mentioned by historians as being one of the most important features that would leverage the CENDARI VRE and distinguish it from other software currently used in the field.
Named Entities are **persons**, **organizations**, **places**, **dates**, **events**, **artifacts**.
The VRE will allow for

- manual tagging of entities
- **automatic** tagging of entities (with possible manual correction)

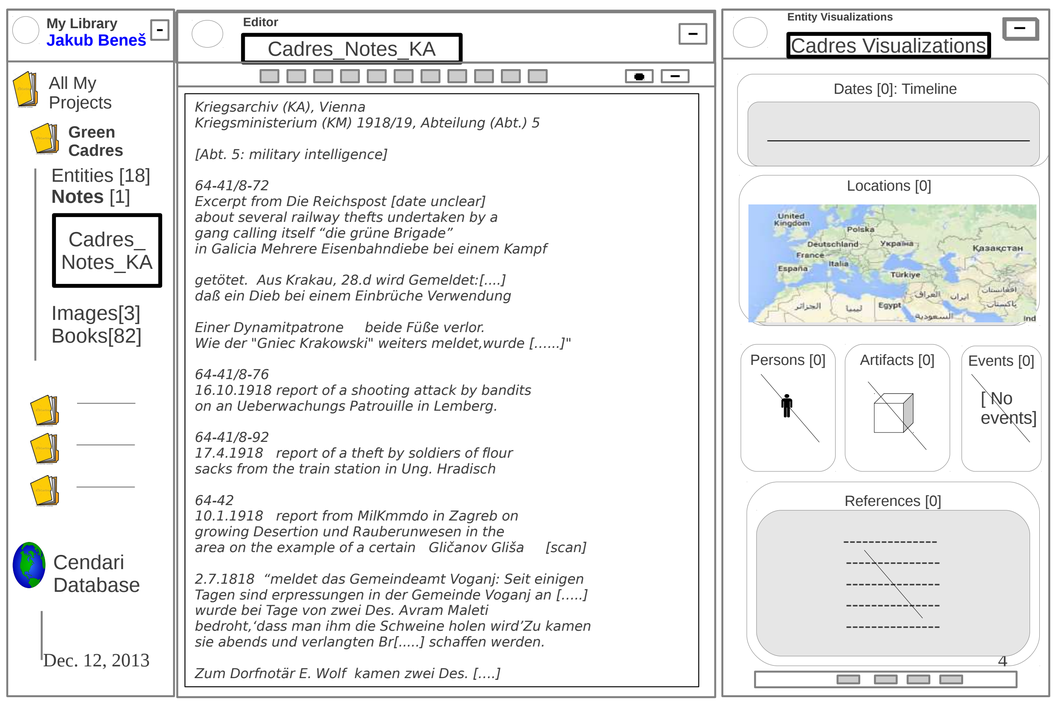


Named Entity highlighting in the notes & visualizations directly beside!

VRE prototype

Historians' requirements and wishes

- two data spaces: *personal* and *project-wide*
- collecting notes (taken in e. g. archives)
- uploading files (e. g. scans of documents)
- recognition of Named Entities in notes
- visualization of found entities & documents → foster exploration and analysis
- collaboration and sharing of notes & documents (if wanted)
- enrichment of a common repository of historical information



VRE mockups by Jean-Daniel Fekete

3 major components:

- **LEFT: Storage** user data space / CENDARI shared data space
- **MIDDLE: Note-taking environment** creation, modification & use
- **RIGHT: Entity visualization** overview, exploration & use

✔ **Simplicity:** minimalistic design for ease of learning
✔ **Gentle learning:** typing in an editor is the only pre-required user knowledge
✔ **Unification:** all CENDARI services in one platform

Named Entity Recognition

Problems in the historical domain

Statistical approaches are state-of-the-art in NER. They are accurate, provide high coverage and are portable when applied to new domains. However, the customization of such algorithms towards the historical domain raises several specific challenges:

- **Lack of training data** and reference/evaluation corpora
- **Lack of knowledge resources** (gazetteers, terminological databases). Existing gazetteers and terminological databases are only partially helpful for the historical researcher. They are more relevant for contemporary history.
- **Multilinguality of sources** and heterogeneous writing systems in use (for the World War I domain, especially Eastern European languages)
- **Digitalization at an early stage**: the documents to be processed are poorly integrated and normalized.

The gap between unstructured and semi-structured data on the one hand and semantic representation on the other is thus significantly larger than for fields such as biotechnology or chemistry where NER currently is most advanced.

Solutions

For **contemporary English**, we use statistical NER based on Conditional Random Fields (CRF), allowing for very accurate and fine-grained resolution (e. g. not only choosing entity type "person", but "military person").

For **other languages**, we are currently developing an original approach based on intensive exploitation of existing generalist knowledge bases:

1. A huge lexicon is assembled from Freebase and Wikipedia, containing entity names and translations into various languages.
⇓
2. Text in a target language is searched for entities from the lexicon. (The lexicon is large enough so that some will always be found.)
⇓
3. For unambiguous matches, machine learning is applied using language-independent features to find entities not in the lexicon.
⇓
4. Resolution of entities against the lexicon is done using measures of semantic relatedness.

Proof-of-concept demo for 1. & 2.: `http://dev1.cendari.saclay.inria.fr/bulgarian/`

Poster made with LaTeX and *fancytikzposter* by Elena Botoeva: `http://www.inf.unibz.it/~ebotoeva/fancytikzposter.html`