Workflows, Digital Data Management and Curation in the RETOPEA Project

Ilenia Eleonor Laudito
Leibniz Institute for European History (IEG)
Mainz, Germany
laudito@ieg-mainz.de

Abstract

English. The main aim of the RETOPEA project is to give insight into political and religious peace-making history offering a prism for interpreting contemporary social issues related to religious diversity and tolerance. The project emphasises a reflective learning from history, where teenagers with different cultural, national and ethnic backgrounds will actively interpret historical and contemporary information about past religious conflicts and the present representation of similar conflicts, in contrast to the classical schooling methods where students are merely passive receptors. To achieve this purpose the students will create a series of short films that will be published on the project's website among the background informations developed by the historical researchers. The project's aim demands a multilingual dataset to guarantee a proper understanding of the produced research data by the students residing in eight different European countries. Due to this requirement, not only all data but also its metadata necessitated a proper translation from English in seven different European languages. This paper describes the workflow and the procedures used in this on-going project and depicts the possibilities and the necessity of multilingualism and automatic translations, as well as the technical issues encountered concerning these topics.

Italiano. Il principale obiettivo del progetto RETOPEA è quello di fornire informazioni storiche sulla pace politica e religiosa, offrendo un prisma per l'interpretazione di attuali questioni sociali legate alla diversità religiosa e alla tolleranza. Il progetto pone l'accento su un apprendimento riflessivo della storia in contrasto ai classici metodi scolastici, in cui gli studenti sono meri recettori passivi. A tal fine gli adolescenti partecepanti al progetto, residenti in otto diversi paesi europei e dunque aventi diversa identità culturale, nazionale ed etnica, realizzaranno una serie di cortometraggi, rappresentando e confrontando attivamente conflitti religiosi passati e presenti. Sia i cortometraggi che il materiale elaborato dal gruppo di ricerca storica saranno pubblicati sul sito web del progetto. Al fine di garantire una corretta comprensione dei dati di ricerca, è necessario un dataset multilinguale. Perció tutti i dati di ricerca con i relativi metadati saranno tradotti dall'inglese nelle sette diverse lingue europee del progetto. Questo articolo descrive l'organizzazione del flusso di lavoro, le metodologie e le procedure utilizzate nel progetto e illustra la necessità di dati multilinguali e di traduzioni automatiche in progetti di umanistica digitale, soffermandosi sulle ulteriori possibilità di sviluppo di tali funzioni e sulle problematiche techniche incontrate nel corso del progetto.

1 Description and aim of the RETOPEA project

Funded by the European Commission under the program Horizon 2020, RETOPEA (REligious TOleration and PEAce) aims at creating a modern understanding of religious conflicts and peace-making history among youngsters and students throughout Europe. The intention is to teach in a comprehensible and appealing way, complex aspects of the past and present society. The project's target group are students between 12 and 18 years old, attending schools as well as non-academic institutions in European countries partnered with the project (which are Spain, UK, France, Belgium, Germany, Finland, Estonia and Macedonia).

Characteristic to this project is its mixture of a historical corpus of peace treaties and agreements – spanning from settlements prior to the anno domini to the most recent Charter of Fundamental Rights of the European Union – and contemporary political discourses, popular culture among teenagers, new spiritual initiatives and heritage. The materials selected and processed by the historical research groups (called "clippings"[1]) will be disclosed on the RETOPEA official website and will serve as primary resources and background information for the creation of short documentary films about the different aspects of tolerance and religious coexistence

[1] A clipping is a piece of information with a length of ca. 200 – 500 words, about a specific subject, possibly containing different media types and formats.

by the students participating to the project. These short films (called "docutubes") will be published on the project's online platform and can be used in the future for further teachings.

It is essential to the project's aim and purpose that all research data is correctly translated in the seven languages corresponding to the above-mentioned partnered European countries. This fundamental requirement assures that all students, independently from their knowledge of the topics presented and their level of understanding of the English language, are able to comprehend solidly the informations provided by the researchers.

2 Technical environment, workflow and data management

The Data Management Plan guarantees the storage sustainability and the long-term preservation of all relevant research data produced and processed by the historical research groups. The collected data will be publicly available via a Virtual Research Environment (VRE). Additionally, the designed workflow establishes a proper coordination between the historical research groups and the technical requirements of the project.

2.1 Technical enviroment

The technical specifications for the storage, preservation and visual presentation of the collected data consists of three main components: a collective access database (with an API to link material from other databases and platforms), a digital repository (TENEO) and a publishing tool (Omeka) that also serves as the project's official website.

Crucial for the publication platform's selection was a user-friendly environment for students, teachers and researchers. Omeka is a web publication platform for sharing digital collections and creating media-rich online exhibits (Omeka S User Manual, 2019). This tool is mainly used by universities, archives, museums and galleries and fits the project's specific demands, due to the heterogeneity of its data.

2.2 Workflow

The workflow divides the clipping's production process in four main stages: creation of the source clippings in English, automatic translation of the clippings, review of the automatic translations and upload by the research data manager into the VRE.

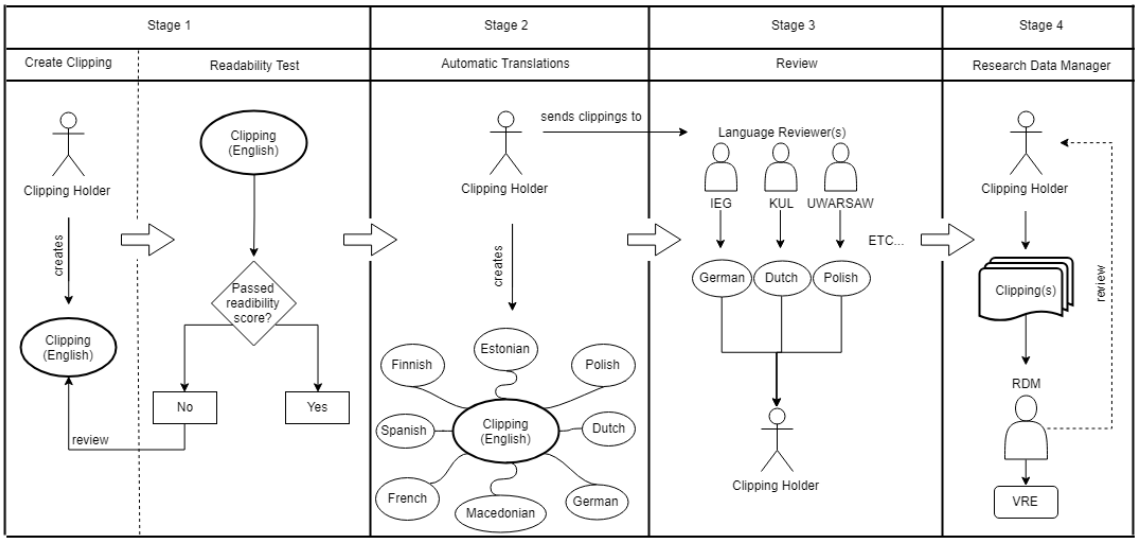

Figure 1: The four stages of the workflow.

All clippings require to be written in English to facilitate the translation of its content afterwards. Additionally, the clippings need to pass a readability test – which is based on the English language – to match the student's reading and grade level. The selected material differs widely in subject matter, field and case study. According to the analysis made on the clippings reading ease and grade level, RETOPEA safeguards an appropriate level of understanding of the complex topics selected by the researchers. The pedagogical importance of this step underlines once again the project's priority concerning the didactical value of the data produced.

The second stage concerns the automatic translation of the clippings, whereas the third stage concentrates on the review of the automatically translated content of the clippings. As explained, it is fundamental that all clippings are grammatically, linguistically and thematically comprehendible by youngsters. The current translation tools available are fallible, thus the need for a reviewing process by native language speakers. Nonetheless, the possibility to accelerate the workflow and lighten the workload of the translation process through automation is a major advantage. At the present state, the tools used for the automation were tested only on a small number of clippings, since the research groups must work on its production first. From the various web translators tested, the following tools will be used in the project: DeepL for Dutch, German, Spanish, French and Polish, Google Translate for Macedonian and Finnish and Tilde for Estonian. The most arduous languages to translate result to be Finnish and Macedonian, not only for the complexity of the language *per se*, but mostly for the lack of tools available.

In the final step the research data manager will upload the complete dataset in the VRE and make small adjustments concerning the clippings' visual representation based on the resources and formats used (text, video and/or audio), in concordance with the clipping's creators.

Albeit this workflow is efficient and functional to RETOPEA's work organisation, it ought to be adapted to fit other requisites and necessities, if used in different projects. There are mainly two things that should be taken into consideration: first, the volume of the research data produced; second, the financial aspect of a reviewing process should not be underestimated and adequately evaluated.

RETOPEA has a relatively small dataset of short text snippets (approximately 400 clippings) which can be uncomplicatedly handled and humanly reviewed. Due to this factor, the costs and extent of a reviewing process are limited. If a project has a broader dataset, then the human involvement and financial aspects should be carefully scrutinised, especially in the case of minor European languages. For example, as for the Finnish language review, it may be more convenient in terms of financial resources and time management to directly translate the data from the source language without going through an automatic translation.

This notwithstanding, one major advantage of this workflow is that it can be easily managed, exploited and followed by researchers, independently from their computer skills and know-how.

2.3 Metadata structure

Every clipping necessarily includes a written part and may contain different media types in various formats beyond its written content (e.g. additional textual resources, images, video and/or audio material, YouTube or other URLs, etc…). To facilitate both the researchers' and the research data manager's work, the researchers will fill the clipping's metadata in two simple spreadsheets, containing the tags selected by the research data manager in the column header. The spreadsheets will be pasted and exported to CSV format and uploaded in the Omeka environment. Given the dataset's heterogeneity in terms of composition of resources, the development of a project-wide metadata scheme to normalise the metadata records was mandatory.

The metadata standards and controlled vocabulary used in the project are the *Dublin Core Metadata Standard* (DC), the *Bibliographic Ontology* (BIBO), *GeoNames* (GN) and the *Canadian Writing Research Collaboratory Ontology* (CWRC). Omeka provides an automapping module that maps the metadata terms of the spreadsheet's header to the imported vocabularies (Omeka S User Manual, 2019).[2] The module also allows to associate a media source (e.g. HTML, URL, YouTube link, etc…) to a selected metadata tag. In RETOPEA's case, the clippings' contents are ingested as HTML-code through the Bibliographic Ontology Term <Content>. The tag will be hidden afterwards and will not be displayed as metadata, yet the content will be visible as media attachment.

[2] For example, "dcterms:title" is automatically mapped to the Dublin Core <Title> property.

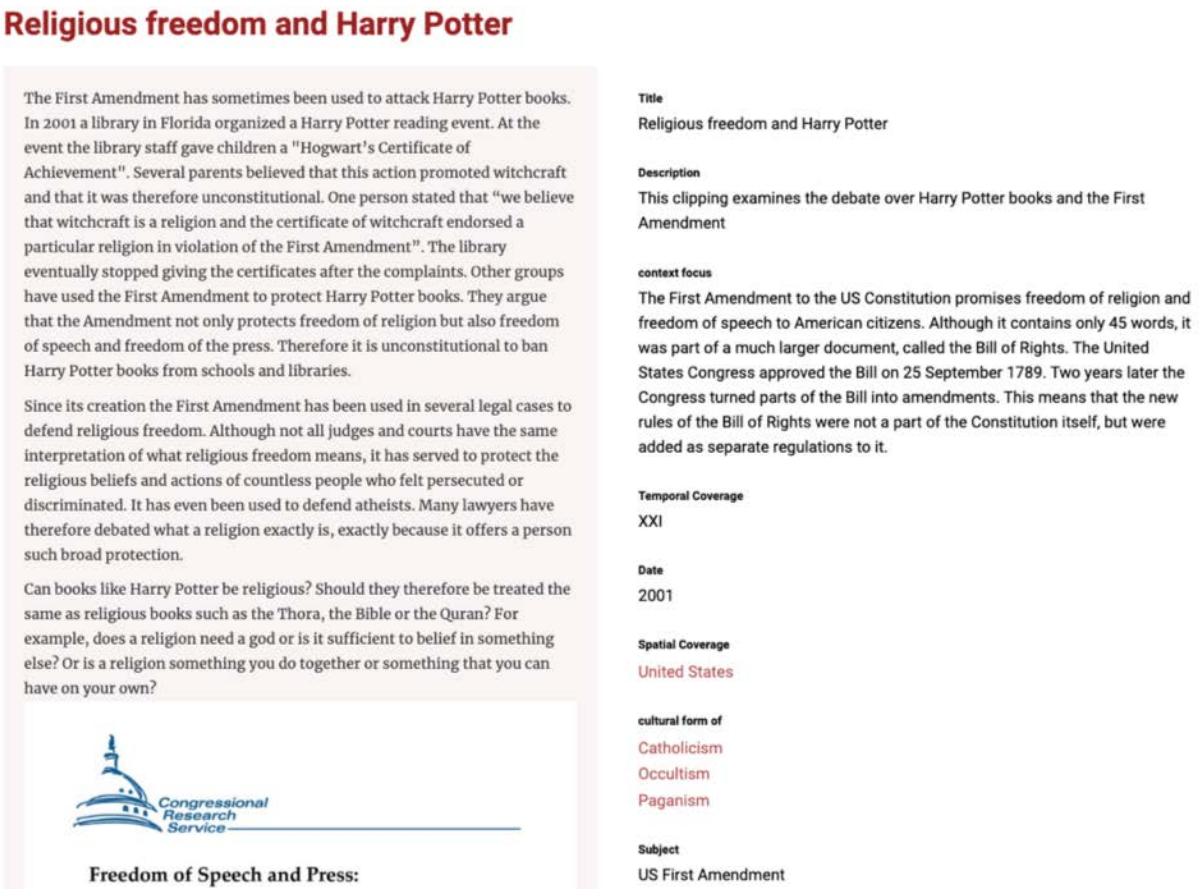

Figure 2: Draft of a clipping as displayed on the RETOPEA website.

Further, Omeka allows to aggregate the ingested data in user-made collections. The twelve collections used in this project aim at grouping the clippings into abstract thematic classifications of project-relevant keynotes. The collections used in RETOPEA represent generic topical focuses (e.g. "Religious practice", "Gender and Sexuality", "Propaganda and stereotyping", etc…), to which clippings can belong independently of their subject. Subjects differ from the topical focuses, for the latter have a broader thematic range and may apply to an indefinite number of clippings, whereas the intent of the subject's list is to bundle a relative small number of clippings into strictly defined subjects (e.g. "Peace of Augsburg", "Edict of Nantes", "YouTube channels", "Political speeches", etc…). The same clipping can appear in more than one collection, depending on how many relations the researcher associated to the clipping. This type of arrangement constructs an intricate entanglement between clippings in order to create vast links and relations between clippings that do not share the same subject matter. These relations and clusters belong to and are part of the metadata description, creating both a vertical and a horizontal hierarchical structure. The main purpose of this structure is to drive the website users and the teenagers using the provided clippings to produce their docutubes to discover as many clippings as possible, independently of the clipping's affiliation or subject matter.

Besides the clipping's title, description, contextual focus and content, the implemented BabelNet API will automatically extract and translate all other metadata tags and keywords through an HTTP interface that returns JSON (Navigli and Pozzetto, 2012). BabelNet not only functions as an online translator, but also recognises synonyms, word sense and (multilingual) semantic relatedness, shaping the possibility to generate linked data and semantic networks.

The described metadata structure was designed specifically for RETOPEA and determined by the intensive

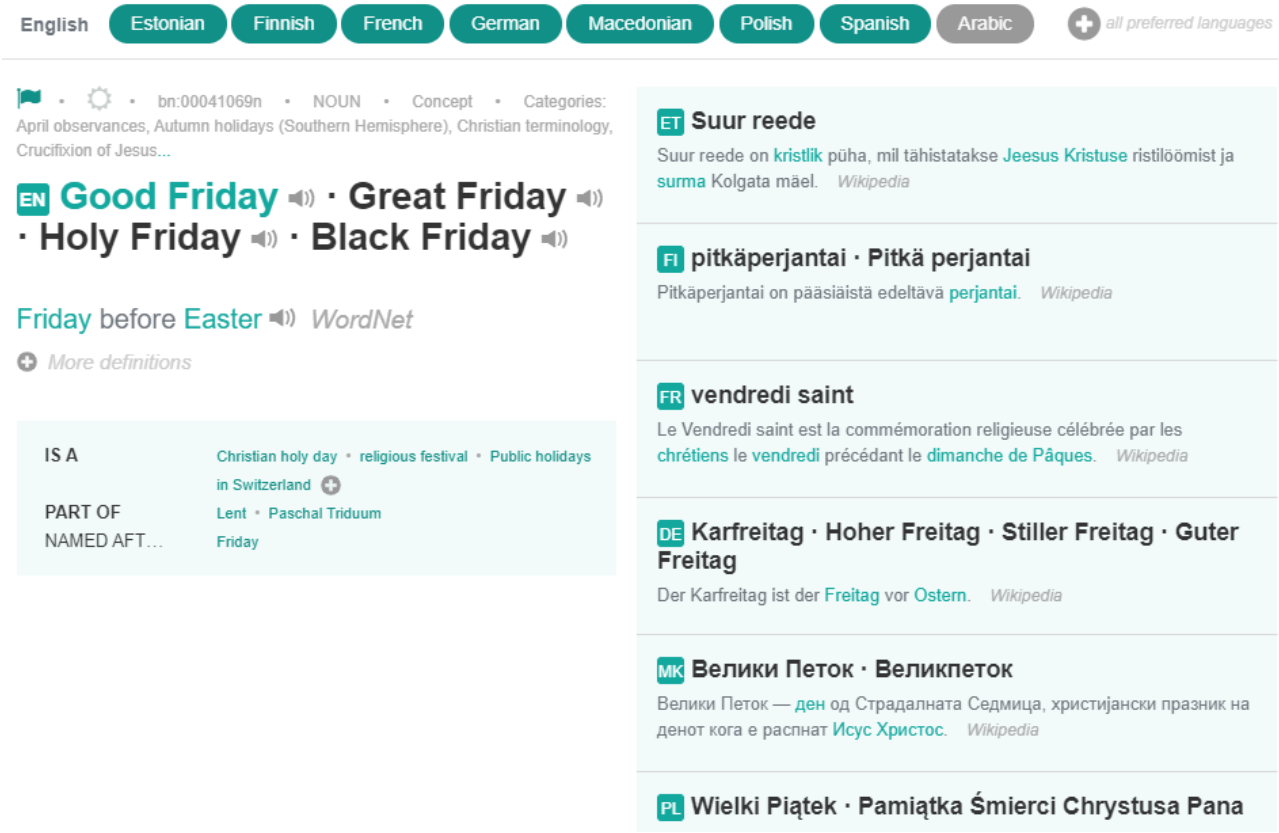

Figure 3: BabelNet translation of "Good Friday".

collaboration, confrontation and discussion with the two historical research groups. Due to the project's distinctive didactical purpose and sundry data, it contains peculiar arrangements that are not always feasible or desirable in other DH-projects. This notwithstanding, this structure could be readily adopted and accordingly modified to fit other requirements.

Considering RETOPEA's didactical and pedagogical purpose, the project's resources will be disclosed under the Creative Commons Licenses (i.e. CC BY-NC-SA). Most of the external materials used in the project can be likewise used and remixed by third parties. Additionally, future tasks will concern the implementation of external databases, like the IEGs "Maps" and "European History Online" (EGO) Databases, and the "On site, in time" project.

Further developments in RETOPEA will give a more precise evaluation about the translation tools used and the metadata structure. Moreover, the controlled vocabularies used leave open the possibility to connect, organise, retrieve and interlink the project's resources in Linked Open Data.

3 Innovation possibilities and DH importance

The methodology and the workflow described in this paper aims at giving a suggestion on how humanistic projects can organise and arrange the digital data produced and the immense possibilities that Natural Language Processing tools and approach may offer.

In the last years the availability and growing amount of data extremely increased, also through the thriving of Digital Humanities related projects. The need for automatically translated documents, data and metadata will increase as more DH-projects arise worldwide. This need does not only apply to major and minor European languages, but also to Arabic and Asiatic languages as well as dialects.

Acknowledgements

The research was supported by the Leibniz Institute for European History (IEG) and the Religious Toleration and Peace (RETOPEA) research project. This paper is based upon work supported and funded by the European Commission under the funding program Horizon 2020, Grant CULT-COOP-05-2017. Special thanks go to Marco Büchler, who provided insight and expertise and collaborated to the development of the workflow. Further gratitude goes to Bram De Ridder, who wrote the clipping shown in "Figure 2", for granting and permitting the publication of his draft example.

References

Navigli Roberto and Ponzetto Simone Paolo. 2012. *Multilingual WSD with Just a Few Lines of Code: the BabelNet API*. Proc. of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012), Jeju, Korea, July 9-11, pp. 67-72. http://wwwusers.di.uniroma1.it/~navigli/pubs/ACL_2012_Navigli_Ponzetto.pdf

"Omeka S User Manual". *Omeka S*, Corporation for Digital Scholarship, Roy Rosenzweig Center for History and New Media, George Mason University. Accessed 15.09.2019. https://omeka.org/s/docs/user-manual/