Ref. Ares(2020)4026405 - 30/07/2020

Horizon 2020 Grant nº 770309
Religious Toleration and Peace
Mainz, 30.07.2020

Report on tutorials on data integration

Contributor(s): Ilenia E. Laudito

Institution: IEG – Leibniz Institut für Europäische Geschichte

Contact: laudito@ieg-mainz.de

Related project task:	T 4.3
Type of document:	Report
Document reference:	Deliverable 4.11
Version number:	1.0
Status:	Final
Last revision date:	30 July 2020
Verification date & verifying subject:	30 July 2020, Christophe Schellekens
License:	CC BY
Other, if applicable:	/

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under Grant Agreement n° 770309.

The clippings' ingestion in the VRE can be divided into the following steps:

1. Preparation of the research data and its metadata
2. Harmonisation of the metadata sheets
3. Data ingestion and manual edits

Previously, an appropriate folder structure was created in Seafile, where all different stages of the clippings' review process are maintained.

1. Preparation of the research data and its metadata

The clippings' production, translation and metadata attribution were arranged through the development of a workflow that also served as a tutorial for the researchers. This workflow is resumed in 3 different documents, all available in Seafile.[1] The workflow has been divided into four steps:

1. Creation of the source clipping
2. Metadata
3. Translations
4. Upload in the VRE

Creation of the source clipping

The source clipping had to be written in English and undergo a readability test to ensure the appropriate understanding by the project's target group.[2] Every clipping was saved as a single word file containing a title, a description of the clipping (if needed) and the content (including text, images and/or links). The file names were standardised to the following format:

Last name of the researcher – a1 *name/title of the clipping*.docx

where the **letter** (e.g. "a") represents clippings on the same subject and the **number** the amount of clippings on that specific subject. This format facilitated the clippings' identification during the review, translation and ingestion processes, especially for the Language Reviewers and the RDM.

The source clippings were uploaded in Seafile and reviewed by the PIs. Subsequently, all clippings were reviewed by our English-speaking partner (Open University). The Language Reviewer (or the delegated

[1] In the "WP4" folder: "Explanation Workflow.pdf", "Tags Explained.pdf" and "examples.pdf".
[2] A more detailed explanation about the readability test can be found in deliverable 4.1 (Data Management Plan), pp. 6-7.

person) downloaded the files and reviewed them, making the corrections visible for the researchers in order to implement the changes. This step assured consistency and accuracy of all written texts.

The Language Reviewer is responsible for the grammatical correctness of the clipping. However, the Clipping Holder is responsible for every content-related material. The LR (or delegated person) re-uploaded the reviewed clippings in Seafile with the same file name. The RDM controlled this process, updating the researchers per email.

In the meantime, the researchers were asked to fill the metadata sheets created by the RETOPEA's RDM, leaving the metadata tags <content> and <description> blank and filling these tags only after the English LR reviewed their clippings.

Metadata

To avoid congestion between the first clipping review and the ingestion of the source clipping in the VRE, the first part of this step took place while the researchers were waiting for feedback from the English Language Reviewer. After receiving a notification from the RDM confirming the grammatical correctness of the clipping, the reviewed and final version of the source clipping and the according metadata sheets (both with the same file name) were uploaded in Seafile.[3]

Translations

Every textual part (i.e. the clipping's content, context, description and questions) has been translated in seven languages: German, French, Spanish, Polish, Dutch, Macedonian and Estonian. The other metadata tags will be automatically translated via the *Babelnet* API.

To accelerate the clipping's translations from the reviewed English source clipping to the above-mentioned languages, the following automated translation tools have been used:

- *DeepL* for Dutch, German, Spanish, French and Polish.
- *Google Translate* for Macedonian.
- *Tilde* for Estonian.

For each language, a separate Word-document has been created, keeping the original English text inside to facilitate the review for the Language Reviewers. These files were also named following a standard (in consonance with the associated language, conforming to the Iso639-2 language code list):

[3] The metadata scheme and the vocabularies used are thoroughly described in deliverable 4.8 (Harmonized metadata scheme).

Last name of the researcher – a1_lang *(English) name/title of the clipping*.docx

The automatically translated clippings were then uploaded in Seafile, where a folder for each language was created. The RDM informed the respective Language Reviewers and controlled the translation and language review process.[4]

Upload in the VRE

Once the first review and the English review process was completed and the metadata sheets were filled and uploaded, the metadata harmonisation began.

2. Harmonisation of the metadata sheets

The metadata templates were examined and inspected by the RDM, transforming the un-normalised metadata sheets into harmonised metadata. This included checking spelling and punctuation mistakes, to avoid formatting issues when ingesting the csv files, verifying links and webpages, and correcting double categorisations, tags and redundancies. Questions, mistakes and other issues were discussed with the researcher.

After receiving confirmation from the researcher, the harmonised sheets were re-sorted and prepared for the first upload in Omeka. During the first upload, only the following tags were included: Alternative Title, Title, Content, Description, Context focus, Questions, Temporal Coverage, Date, Spatial Coverage, Subject, Source, Bibliographic Citation, Transcript of, Translation of, Audience, Creator, Contributor.

The excel/numbers sheets were then exported into a csv file, conforming Omeka's prescriptions. The csv files were uploaded in Omeka and the <context> and <content> tags transformed into HTML code, so that Omeka classifies it as "media" (*in contra* to "item").

3. Data ingestion and manual edits

The item sets were manually added, whereas the <map> tag was added afterwards with all other media files. Subsequently, the researchers were asked to verify the correctness of their clippings, both in terms of language and visualisation.

[4] For further clarifications about this stage consult deliverable 4.1 (Data Management Plan), pp. 7-8.

Future modifications

The online clippings have been reviewed by the project team members and PI's. All modifications of the clippings will be implemented in the next months. Moreover, the following matters will be modified in the next revision of the online clippings:

- <date>: will be reviewed and harmonised to a common format, in order to avoid difficulties in understanding the exact date, since every country has its own way to display them (as MM/DD/YY or DD/MM/YY). Therefore, we opted for a "DD. Month YYYY" format (e.g. 01. July 2020).
- <temporal coverage>: it was pointed out that the roman number format of this tag could provide troubles in the comprehension of the temporal coverage. Therefore, a new format will be used here ("3rd Century" for "III").

After these implementations, the metadata sheets of the media used in the clippings will be uploaded and available, as well as the translated version of all clippings.