Weave Project KU Leuven / UNIGE – 'Places of Translation: A comparative study of the emergence of local translation policies in Belgium and Switzerland (1830/1848-1918)'

Data Management Plan

1. Data collection and documentation

1.1. What data will you collect, observe, generate or reuse?

This project will collect and generate data on the translation policies of six Belgian and Swiss cities (Brussels, Antwerp, Liège, Bern, Geneva and Biel/Bienne) in the 19th century. The project team consists of researchers from KU Leuven and the University of Geneva, who will be responsible for the part of the study relating to their respective countries. The data will be compiled and derived from printed sources in public archives. Some of these sources are digitized and available online, while the others are in paper form and can only be consulted on-site. They consist of institutional texts issued by local authorities, which are documents *about* translation (e.g. minutes of city councils reporting discussions on how translation services should be organized) or *resulting from* translation (e.g. legal or administrative bilingual, or 'parallel', texts).

The project will produce two types of data. The first is a selection of relevant primary sources available from the archives. This will consist of a catalogue of references to archival sources (*a*) and a documentary database (*b*) bringing together some of these sources, either taken from the archives in digital form or digitized by the research team itself. The second is data that will result from a study of the sources included in the database, which will be carried out by the research team using digital tools for qualitative content analysis (*c*) and bilingual corpus analysis (*d*).

The files will be in CSV format for *a*, *c* and *d*, and in OCR'd PDF format for *b*. At the time of project submission, it can be estimated that the volume of data will not exceed 50 GB.

1.2. How will the data be collected, observed or generated?

The data will be collected and generated through several means: automated keyword searches and manual searches of the archives, cataloguing (*a*) and compilation (*b*), computer-aided qualitative content analysis with software such as QDA Miner or NVivo (*c*), and corpus-based analysis of aligned bitexts with concordancers such as ParaConc or NoSketch Engine (*d*).

For each dataset, the files will be organized according to the general structure of the project and the research team, with subfolders by country and by city:

- Catalogue of references
  - Belgium
    - Brussels
    - Antwerp
    - Liège
  - Switzerland
    - Bern
    - Geneva
    - Biel/Bienne
- Documentary database
  - Belgium

…
Qualitative content analysis
Belgium
…
Bilingual corpus analysis
Belgium
…

File naming will follow a convention such as: CABr_translation_quality_v003_20210325.csv (CABr = 'content analysis Brussels').

1.3. What documentation and metadata will you provide with the data?

README.txt files will be created for the project data as a whole and for each dataset (*a*–*d*). They will contain the following information: title of the project/dataset, DOI (at the final stage, i.e. data sharing), name and contribution of the researcher(s), dates of data collection and processing, methods and tools used, description of the tree structure and file naming principles, list of abbreviations and conventions, and access conditions (licence).

Developing a metadata model to record the primary sources and create the database (*a*–*b*) and defining coding grids and principles to carry out the computer-aided analyses (*c*–*d*) is an integral part of the research project. This can be used as a basis for completing the documentation to be provided with the data (e.g. 'methods and tools' items in the README.txt files). Moreover, some of the publications planned under the project, which will be open access, have a methodological focus. They will facilitate the possible reuse of the data by other researchers.

2. Ethics, legal and security issues

2.1. How will ethical issues be addressed and handled?

In principle, the data that will be collected is not sensitive. The primary sources that will be used are available from public archives and relate to people who were born before 1900 (and thus were either adults in 1918 or had died prior to that date). In practice, this rules out the risk of unauthorized disclosure of personal or other sensitive data.

2.2. How will data access and security be managed?

Since there are no ethical issues (see 2.1), the only risks are the loss and corruption of data. During the research process, the data will be stored in private access or in shared access by team members, protected by a password, and backed up regularly, under the responsibility of the two PIs (see 3.1).

2.3. How will you handle copyright and Intellectual Property Rights issues?

The internal structure of the datasets (see 1.2) makes it possible to clearly distinguish the data that have been generated, and are therefore owned, by each of the two universities according to their regulations.

With regard to the database (*b*), it should be noted that, as a general rule, the reproduction of printed documents from public archives dating from the period under study is permitted by law, provided that specific referencing principles are respected.

The licences that will be used for long-term archiving (see 4.1), in accordance with the open science policy, are CC0 for the database (reproducing documents that are already in the public domain) and CC BY for the other datasets, in order to achieve maximum interoperability.

3. Data storage and preservation

3.1. How will your data be stored and backed-up during the research?

During the research process, the data will be stored separately by country and city (see 1.2), under the responsibility of the two PIs. The data produced by each university will be stored in triplicate: one copy on the academic NAS (Network Attached Storage) managed by the local IT department, which provides automated daily backups and allows researchers to revert to earlier versions of the files over two weeks; one on a personal computer and one on a portable storage device (external hard drive), both updated weekly by the researcher involved. Working documents (e.g. metadata models for the catalogue and database, coding grids and examples for the computer-aided analyses) will be shared in the online storage space (e.g. OneDrive Entreprise). The expected volume of data (see 1.1) is compatible with the storage capacity of these various devices.

3.2. What is your data preservation plan?

The data collected or generated will be preserved for five years after the completion of the project and shared at the latest at the time of the relevant publications. The project design provides for a progressive selection of sources in two phases of analysis (macro-analysis and selection of topics in work package 2; digital analysis of sources on selected topics in work package 3). Only the selected sources will be included in the documentary database for the second phase. Among the raw data extracted during this second phase, only those actually used in the analyses and publications will be shared in the long term through permanent archiving, while working files that can be considered as drafts will be discarded. The files will eventually be archived in PDF/A (*b*) and CSV (*a*, *c*, *d*) format.

4. Data sharing and reuse

4.1. How and where will the data be shared?

The data will be shared in line with the open science policy of the two universities and funding agencies (in particular, the *FAIR Data Principles*) without breaking up the data pool after the completion of the project. The data selected to be shared in the long term (see 3.2), together with the metadata files (see 1.3), will be archived without time limit on Yareta, the long-term repository of the University of Geneva. The licences used will be CC0 for the database and CC BY for the other datasets (see 2.3).

4.2. Are there any necessary limitations to protect sensitive data?

In principle, there will be no sensitive data (see 2.1).

4.3. All digital repositories I will choose are conform to the FAIR Data Principles.

X

4.4. I will choose digital repositories maintained by a non-profit organisation.

X

VD & RM, 25.03.2021