Open Science and Linked Data to Create a FAIR Corpus of Intra-Belgian Book Translations 1970-2020

Sven Lieber, KBR (Royal Library of Belgium)
Ann Van Camp, KBR (Royal Library of Belgium)

Until now, intra-Belgian literary translation flows have never been studied at a large and systematic scale, neither quantitatively nor qualitatively, for the past fifty years. To fill this gap, KULeuven, UCLouvain and the Royal Library of Belgium (KBR) launched the BELTRANS project, founded by the Belgian Science Policy Office (BELSPO): an interdisciplinary research project on intra-Belgian literary translations from 1970 to 2020. One of the objectives is the creation of a FAIR Linked Data corpus which not only supports our research but can be reused by the community. In the following we present the specific challenges of the BELTRANS project, identify the gap of available open data for existing translation studies, and present our openly available data processing workflow to create a FAIR corpus.

Specifically, the BELTRANS project aims to provide an as complete picture as possible of what kind of books from Belgian authors, illustrators or scenarists have been translated from Dutch to French or French to Dutch from 1970 to 2020 in Belgium or abroad, anywhere in the world. The project investigates among others (1) book translations of different literary genres, i.e. novels/comics/poetry/youth literature, but also literary non-fiction and life writing in different formats such as printed, e-book or audiobook; (2) a mapping of an unexplored period of cross-cultural actors, translators and networks in Belgian translation history; (3) linking literary book translations to reciprocal cultural perception between the Flemish and French Communities, to regional and national identity and to political developments. This enables analyses on size and structure of translation flows, gender (im)balance, or the diachronic evolutions of intra-Belgian translation flows.

However, the creation of a FAIR data corpus is not straightforward considering these challenges, because different types of heterogeneous data need to be processed. KBR has already gathered a vast collection of contemporary Belgian publications through legal deposit, available in CSV or MARCXML format. However, there are factors preventing us from using KBR’s catalogue as a sole data source. Belgium is a multilingual country and Belgian authors are often published abroad, e.g. in the Netherlands, France, Switzerland or Canada, and therefore are not always aware of the legal deposit (or do not comply). Additionally, BELTRANS-relevant information such as the source language of a translation, the link from a translation to the original work, gender or nationality of contributors were not always curated in the catalogue in the past 50 years. Furthermore, linking literary book translations to reciprocal cultural perception requires non bibliographic data. A plethora of other data sources exist, from high-quality and structured information of other national libraries (BnF[1], KB[2], etc) or the ISNI International Agency[3] to smaller less structured databases specialised in translations (Index Translationum[4], Vertalingendatabase[5], DLBT[6], etc) or even lists in Excel from public institutions that offered translation grants. Besides these data sources, curated by single institutions, also crowd-sourced data sources such as Wikidata[7] exist.

Good data management is the key conduit leading to knowledge discovery and innovation (Wilkinson, 2016), therefore BELTRANS aims to publish the underlying research data as well. Existing translation studies such as (Heilbron, 2015) or (Voogel, 2018) already offer diachronic statistics covering several decades. However, only aggregated statistics are provided, but for further research it would have been interesting to analyse the specific datasets that were taken into account. Translation studies often involve large-scale analysis, both geographically and over time (Roig-Sanz, 2021). This makes it necessary to use computational methods for data analysis. Recently, the term Big Translation History (BTH) was coined and defined by Diana Roig-Sanz & Laura Fólica as “translation history that can be analysed computationally” (Roig-Sanz, 2021). Within their presented case study the authors emphasise that the data cleaning process was time consuming but fundamental to get a curated database. This is possibly applicable in most other digital humanities projects as well. Unfortunately, such time consuming work, especially in multi-year and collaborative projects, is liable to go undocumented and thus stays unrecognised (Edmond, 2020). For BELTRANS we find it necessary that this kind of data processing is recognisable as well. Software has become an essential part of scientific research and thus it is desirable to apply the FAIR principles also to software (Lamprecht, 2020).

In this short paper, we will present our publicly available semi-automatic workflow of cleaning, enriching and integrating different types of heterogeneous data. The workflow is ongoing work but its components are publicly available in a GitHub repository under the MIT licence[8]. The workflow itself is not novel as it concerns standard data processing combined with Linked Data techniques. However, to the best of our knowledge this is the first openly available repository to share code with

[1] https://www.bnf.fr/fr
[2] https://www.kb.nl/
[3] https://isni.org/
[4] http://portal.unesco.org/culture/en/ev.php-URL_ID=7810&URL_DO=DO_TOPIC&URL_SECTION=201.html
[5] https://letterenfonds.secure.force.com/vertalingendatabase/zoeken
[6] https://dlbt.univie.ac.at/
[7] https://www.wikidata.org/
[8] https://github.com/kbrbe/beltrans-data-integration

respect to Big Translation History. We explain how we use the issue and milestone features of GitHub to plan versions of the data corpus. Regarding the workflow, we explain how different preprocessing components developed in Python support us to filter large RDF data dumps of several dozen Gigabytes performantly in a streaming fashion with limited resources. Furthermore, we present how we exploit the advantages of Linked Data already during the data integration process, instead of merely for the last step of data publication after the data integration. To this end, we use the rml.io (Dimou, 2014) framework to map heterogeneous data sources to RDF in a declarative fashion. These integrated data enable us to apply BELTRANS-relevant filters such as the Belgian nationality of the author, illustrator or scenarist using a declarative SPARQL query. Finally, we discuss short-comings of the workflow as well as challenges and lessons learned. For example the risk of moving the problem of an opaque but straightforward data integration process to a Linked Data blackbox, only accessible by experts familiar with RDF technologies and computer science methods.

References:

Roig-Sanz, Diana, and Laura Fólica. "Big translation history: Data science applied to translated literature in the Spanish-speaking world, 1898–1945." *Translation Spaces* 10.2 (2021): 231-259.

Edmond, Jennifer, and Francesca Morselli. "Sustainability of digital humanities projects as a publication and documentation challenge." *Journal of Documentation* (2020).

Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., & Van de Walle, R. (2014, January). RML: a generic language for integrated RDF mappings of heterogeneous data. In Ldow.

Heilbron, J. & Van Es, N. (2015), 'In de wereldrepubliek der letteren'. In: Bevers, T., Colenbrander, B., Heilbron, J. & Wilterdink, N., *Nederlandse kunst in de wereld*. Nijmegen: Vantilt, 20-56.

Lamprecht, A. L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., ... & Capella-Gutierrez, S. (2020). Towards FAIR principles for research software. *Data Science*, 3(1), 37-59.

Voogel, M. (2018). *Bon ton of boring? De ontwikkeling van het Frans in onderwijs en uitgeverij in Nederland.* Amsterdam: AUP*.*

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9.