Using the Semantic Web to Improve Knowledge of Translations
Presentation

Karen Smith-Yoshimura
OCLC, USA
smithyok@oclc.org

Keywords: metadata; linked data; translations; multilingualism

Abstract

More than half of the almost 400 million bibliographic records in WorldCat are for languages other than English. Most of the monographs described were published only once. But a few million represent the core of our shared culture—works that have been translated into multiple languages, and sometimes translated multiple times into the same language. We learn about other cultures, and other cultures learn about ours, through these translations. As the world's largest bibliographic database, WorldCat is positioned to provide the translation history of works, using the W3C bib extension translationOfWork to communicate the relationship of each translation to the original work.

In our multilingual data enhancements project, our goal was to improve the descriptions of the most frequently published works, as they are the ones most likely to be translated and searched by users. In a database of MARC records, machine processes cannot support browsing or searching of works and their translations. Critical entities such as the title of the original work and the names of the translators are not always expressed in a machine-understandable form—and sometimes the information is missing altogether. Since a manual cleanup is not scalable, we explored the possibility of enriching MARC records with Linked Data from a third-party source, Wikidata. By integrating information from both WorldCat and Wikidata, we may be in a better position to present information about frequently-translated works in the preferred language and script of the user.

MARC records include data elements that can explicitly state that the record represents a translation, the language of the original and any intermediate translations, the title of the original work, and the translator(s) responsible for the translation. As long as *some* records accurately record this information, we can assert the correct relationships for the records that lack the information. Unfortunately, only a subset of all the relevant translations in WorldCat include such rich information. Many books written in non-Latin characters (such as Cyrillic, Greek, Russian) are often represented in WorldCat by the romanization only. This also makes the search by native-speakers unnatural and difficult. Using WorldCat records alone could not identify all the translations and their translators.

We enhanced the data retrieved from WorldCat with data retrieved from Wikidata by retrieving the Wikidata entries for a few works and its labels in multiple languages, even those written in non-Latin scripts. With the title/author match in a different language other than the original one, we can infer with high confidence a translation of the original work, even if the MARC record does not indicate it is a translation. The Wikidata entry often includes the non-Latin script for languages represented in WorldCat only in romanization. For example, we could use the label *Είναι και Χρόνος* from Wikidata rather than *Einei kai chronos* from WorldCat for the Greek translation of Heidegger's *Sein und Zeit.*

Data enrichment could be mutual. For example, Wikidata entries focus on the original title and do not describe all the translations represented in WorldCat; few Wikidata entries include translators, crucial to differentiate translations in the same language. Leveraging the strengths of each resource through linked data offers us the ability to present users an enriched view of our shared culture through translations.