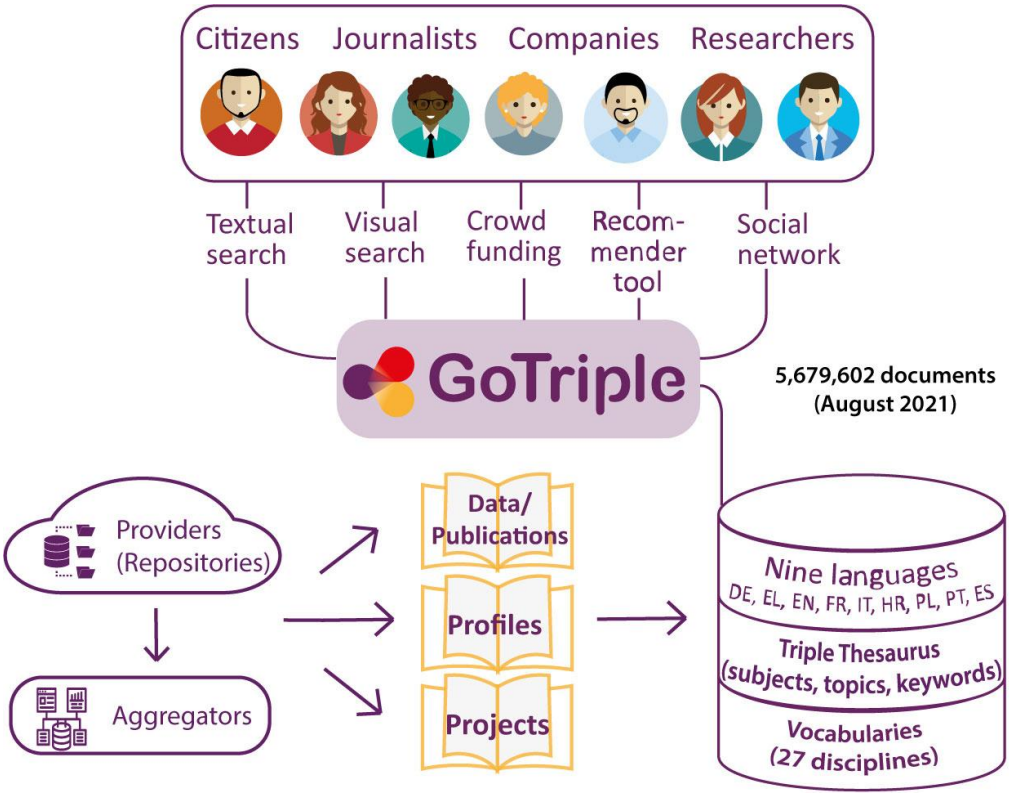
SNSF MARK16 second international conference – 2-3 June 2022, online; org. Claire Clivaz and Mina Monier (DH+, SIB, Lausanne) and Dan Batovici (KU Leuven, BE).

Emilie Blotière (TGIR Huma-Num), Haris Georgadis (EKT), Ondrej Matsuka (Lexical Computing) and Judith Schulte (MWS)

GoTriple, a multicultural and a multilingual discovery service in Social Sciences and Humanities

At the heart of the TRIPLE project launched in October 2019 is the development of the GoTriple platform, an innovative multilingual and multicultural discovery solution for the social sciences and humanities. It will be one of the dedicated services of OPERAS, the Research Infrastructure supporting open scholarly communication in the social sciences and humanities (SSH) in the European Research Area.



GoTriple provides a single access point for users (researchers, institutions such as universities and libraries, but also enterprises and the media):

- to discover and reuse open scholarly SSH resources in nine European languages (Croatian, English, French, German, Greek, Italian, Polish, Portuguese, Spanish), i.e. research data and publications, which are currently scattered across local repositories;
- to find and connect with other researchers and projects across disciplinary and language boundaries;
- to make use of innovative tools to support research (e.g. visualisation, annotation, trust building/social network and recommender system);
- to discover new ways of funding research (e.g. crowdfunding).

In this presentation we will focus on multilingualism which is one of the main aspects of the platform. English is the common language within research to establish a common pattern. Nevertheless the widespread use of English in international conferences and scientific

publications has a tendency to marginalize native language works and to make these works more difficult to access by the majority of researchers. From the perspective of open science, GoTriple should promote closer connections to each language.

A. Controlled vocabulary

One of the technical tasks to achieve this goal was to develop a vocabulary of subjects in SSH to be used by the annotation service of the GoTriple.
In order for the annotation mechanism to be effective for publications of all the nine languages that will be supported, the vocabulary must contain a sufficient number of concepts and, at the same time, these concepts must have labels in as many of these languages as possible. The first action was to gather existing vocabularies and thesauri in SSH. All partners contributed to a list of available vocabularies. After evaluation, the consortium decided that the best candidate was to use a subset of the Library of Congress Subject Heading[1] (LCSH). LCSH is one of the most popular and authoritative vocabularies, it contains a very large number of concepts from which the most common SSH concepts were selected.

The second action involved creating the Triple vocabulary based on LCSH. The vocabulary was originally created by selecting a subset of the LCSH that is related to SSH. The methodology used for selecting the SSH-related concepts was based on identifying 14 basic concepts from the Frascati taxonomy under SSH, then mapping these to 37 broad terms of LCSH and, finally, extracting these LCSH concepts along with their children, using the Linked Data API of the Library of Congress. We ended up with a hierarchical vocabulary of 2565 concepts. The vocabulary is currently hosted in a platform for managing and publishing LOD Vocabularies developed by EKT, Semantics.gr.

The third action was to increase the multilingualism of the TRIPLE vocabulary, since, initially, it had only English Labels, like the original LCSH. At first, existing links of LCSH to other vocabularies were imported, from which labels in our target languages were extracted and added to the vocabulary. Particularly for wiki data links, we extracted and added all the different labels in our target languages provided by wiki data. We also processed mappings from the National Library of Florence and Rameau thesaurus to LCSH, adding more labels in French and Italian, respectively. After all these attempts to increase the multilingualism of the Triple vocabulary, we achieved the following coverage per language from 10% (Croatian) up to 60% (French), with an average of 20%.

The next challenge was to further increase the language coverage. Initially, it was planned that for the missing labels, the partners would manually add them (translating the English labels) directly in the semantics.gr platform. However, we concluded that manual translation is a labour-intensive process. We estimated that it would require on average 3.23 PM (Person Month) for each language. To overcome this problem, it was suggested that we could use a translation service to generate the missing labels by automatic translation from the English labels. IBL PAN developed a script that, using the Google Translation service, generated all missing labels. Based on these results, we created 9 spreadsheets, one per

[1] https://id.loc.gov/authorities/subjects.html

target language, that included translations in that language that need validation.Currently, partners work on these spreadsheets (each language is assigned to a different partner) in order to validate - and occasionally curate - the automatically generated translations. Thus all validated labels have been batch-imported in the TRIPLE vocabulary.

B. Machine learning

Machine learning aims to create a SSH corpus to train the search engine of the platform. Technically speaking, Machine learning is an artificial intelligence algorithm allowing computers to learn a model by using a large quantity of data provided by humans. 100 documents per langage and per discipline have been gathered to be classified and categorized. It is part of the semantic enrichment process

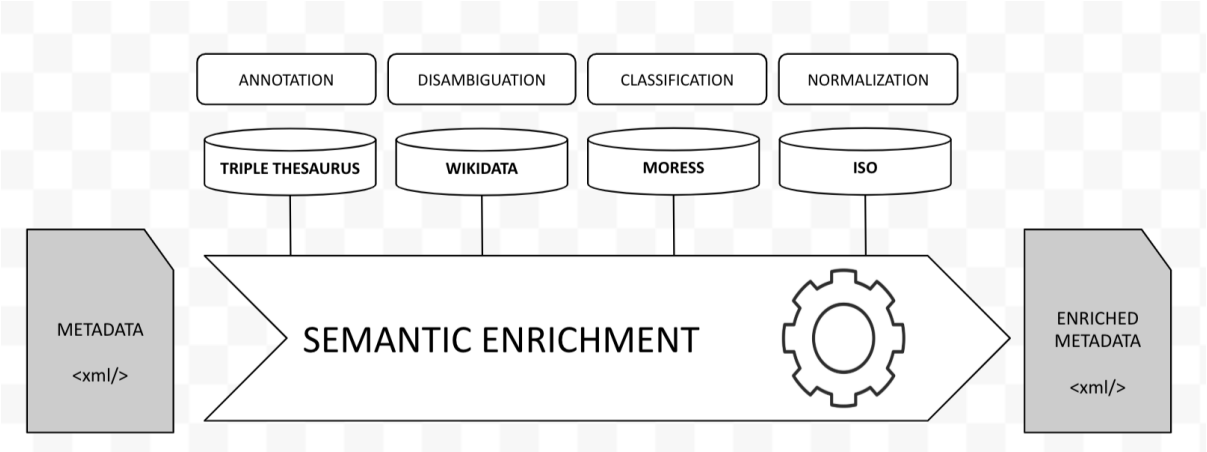


Fig.1 Semantic enrichment process

Each new document/record is ingested in the database and will be automatically classified in a specific SSH category in the nine languages. An algorithm was set up, according to the MORESS categories, which is composed of 27 disciplines in Social Sciences and Humanities.

The task has been divided in two parts :

(1) an automatic collection of texts and metadata (title+keywords+abstract) through the catalog of the Directory of Open Access Journals (DOAJ[2]), according to the MORESS categories, and

(2) a manual collection to complete the corpus, i.e. provide additional material for disciplines and languages underrepresented in DOAJ. The training data comprising more than 240,000 papers in nine languages were validated by the partners for their correspondence to MORESS categories.

The TRIPLE vocabulary is hosted in semantics.gr, a platform created by EKT, the partner in charge of this task, for creating, curating, linking and publishing vocabularies, thesauri and authority files as LOD[3]. The platform was extended with additional functionality to meet the

[2] https://doaj.org/
[3] Linked Open Data

needs for the TRIPLE project. Partners have been given accounts in order to be able to enrich and curate the GoTriple vocabulary in semantics.gr. The TRIPLE vocabulary is currently private but it is planned to be published as LOD, with the TRIPLE Consortium being the creator, after the curation phase is finished.

For the sustainability of the platform, the consortium also set up mechanisms for enriching the GoTriple vocabulary. (1) new concepts (and their descendants) from LCSH and missing will be added, via possible synergies with other digital initiatives in SSH the language coverage will be improved to existing concepts. One of the difficulties is to ensure a satisfactory level of coverage for all languages and in particular to preserve the less used and less represented languages.

In order to increase the multilingualism of the vocabulary, the TRIPLE team also involved an automatic translation service to produce the missing labels in our nine target languages. The automatically generated translations are currently being validated and curated by partners.