DOI: 10.5281/zenodo.6008274

Triple

Transforming Research through Innovative
Practices for Linked Interdisciplinary Exploration

[December 2021]

Advancing Open Scholarship

D2.3 – Machine learning

Version 1.0 – Draft
PUBLIC

H2020-INFRAEOSC-2019
Grant Agreement 863420

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863420

Disclaimer- "The content of this publication is the sole responsibility of the TRIPLE consortium and can in no way be taken to reflect the views of the European Commission. The European Commission is not responsible for any use that may be made of the information it contains."

Triple

Deliverable Name

Project Acronym:	**TRIPLE**
Project Name:	**Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration**
Grant Agreement No:	**863420**
Start Date:	**1/10/2019**
End Date:	**31/03/2023**
Contributing WP	**WP2**
WP Leader:	**IBL-PAN**
Deliverable identifier	**D2.3**
Contractual Delivery Date: 09/2021	**Actual Delivery Date: 12/2021**
Nature: Report	**Version: 1.0 Draft**
Dissemination level	**PU/PR**

Revision History

Version	Created/Modifier	Comments
0.0	Ondřej Matuška, Lexical Computing; Laurent Capelli, CNRS (HN)	writing of the content
1	Paula Forbes, Abertay University; Luca de Santis, Net7; Maciej Maryl, IBL-PAN, Marta Błaszczyńska; IBL-PAN	review

Table of Contents

Acronyms

DOAJ	Directory of Open Access Journals
SSH	Social Sciences and Humanities
API	Application Programming Interface
SSOAR	Social Science Open Access Repository
HALSHS	French Open Archive

Publishable Summary

This deliverable describes the work implemented in the task 2.3 Machine learning. This task provided support for machine learning training on classification of SSH documents according to MORESS classification of disciplines[1]. Native speaker researchers were asked to provide at least 100 documents per language and per discipline to train the system via machine learning and have been involved in the training phase of the system. The task also established how the machine learning results will be later integrated in the TRIPLE environment through the work of WP4 (Building of the platform).

Due to the postponement of the use of the classification model in the processing chain, the original submission date (September 30) has been moved to the end of December 2021. This draft version will be then prepared before the end of the Second reporting period and will be updated once the classification model has been completed.

[1] https://project.gotriple.eu/wp-content/uploads/2020/12/MORESS-categories-for-GOTRIPLE-2020_12.pdf

1| Creation of the TRIPLE Training Database

The TRIPLE training database is a collection of social sciences and humanities scientific papers in the nine languages supported by TRIPLE (French, English, Spanish, Portuguese, Italian, German, Polish, Croatian, and Greek) divided into 27 disciplines according to the classification prepared in the EC-funded MORESS[2] project into which all items in the platform will be divided. The database serves as input data for machine learning to train a classification model for the automatic classification software integrated into the GoTriple platform. The classification software will use this model to assign each new addition (article, thesis, dataset, software and other content) into one of the MORESS categories.

1.1 MORESS categories

The reason behind using the MORESS categories is that they were recently proposed by the scientific community which conducted a mapping exercise of SSH outputs within the framework of an EC-funded project. The outcome was a categorisation closely aligned with the goals of TRIPLE.

The training database required at least 100 papers in each category of each language and 30 additional papers per category and language which served for the validation of the classification model. The creation of the database involved a two-step procedure. First, the resources with already categorised papers were identified. Secondly, for those categories where an insufficient number of papers was identified automatically, a manual selection and classification was performed. The final database is a product of the combination of both of these approaches.

While every effort was made to find a sufficient number of papers per each category in all languages, it turned out that the *Biological anthropology* category was far too specific and relevant papers were very scarce or could not be found in some languages at all. The reason is that in some languages this category was part of another category, e.g. biology or anthropology, while, in some other languages, papers were not available in a sufficient quantity: in fact, although journals in this domain exist, they are mostly published in English. The category was, therefore, dismissed for the purpose of TRIPLE and the final classification model only recognizes 26 MORESS categories.

[2] MORESS (Mapping of Research in European Social Sciences and Humanities - https://cordis.europa.eu/project/id/HPSE-CT-2002-60060) was a EU-funded project that ran from 2003 to 2005 and was coordinated by the European University Association (EUA). These categories, which proposed a classification for SSH disciplines, were one of its outcomes.

1.2 DOAJ and SSOAR

The Directory of Open Access Journals (DOAJ[3]) was used as the starting point and the primary source of the training data. It contains a very large number of papers in many languages and many more categories than required for the purposes of TRIPLE which makes it an ideal resource.

The disadvantage of the database is that it does not store the actual papers but only provides links to locations where the papers can be found. Frequently, the link is not the direct download link, but only the URL to the website that hosts the paper. To facilitate the collection of such data, a specialised crawling tool was developed. The purpose of the tool was to follow the link provided by DOAJ and download the pages of the destination website until the required paper was found. This involved downloading numerous pages from the destination, in rare cases even thousands, before the required paper could be retrieved. This process, which was already very time consuming, was hindered further by the fact that different destination sites use an enormous variety of web technologies and the download tool repeatedly had to cope with content which it was not designed to handle. This required the tool to be continuously adjusted, updated, and reconfigured before the download process could continue. The downloaded content was then converted to plain text and added to the training database.

In the course of the collection, it turned out that the DOAJ language classification was not reliable. The problems included papers classified as the wrong language or classified according to the main body of the text while the abstract was in a different language and similar inconsistencies. The LangID language identification tool (https://github.com/saffsd/langid.py) had to be used to re-process the papers and assign the correct language to the papers. This resulted in removing papers in non-TRIPLE languages and the reduction of the database.

In the case of German, the SSOAR[4] database was added to the training data using a similar procedure.

1.3 Manual collection

While the DOAJ database offered an abundance of papers in certain languages and categories, it turned out that papers in other languages and/or categories were scarce or not available at all. This generated the need for additional training data in all languages. In the case of English, French and Spanish, the existing ISIDORE[5] database was used to complete the required quantities. For these languages, we used documents from the HALSHS[6] open archive developed by the CCSD[7]. When depositing in this database, users must classify each document in one or more MORESS disciplines. In the case of Portuguese, Italian, German, Polish, Croatian and Greek, TRIPLE project partners, with the relevant expertise, were asked to identify suitable

[3] https://doaj.org

[4] https://www.gesis.org/ssoar/home

[5] https://isidore.science

[6] https://halshs.archives-ouvertes.fr

[7] https://ccsd.cnrs.fr

sources manually and collect the required data to meet the minimum target size for each category.

The manually collected data were then added to the automatically collected data to complete the required quantities.

1.4 Statistics of data submitted for the machine learning procedure

The table below shows the numbers of papers per subject and language collected from DOAJ and SSOAR databases and also the additional manually collected data. It also includes the *Biological anthropology* category, which was removed later. French, Portuguese and English were later complemented with data from ISIDORE to meet the planned quantities.

MORESS category	Croatian	Spanish	German	Polish	Italian	English	Portuguese	French	Greek
TOTAL	**38,459**	**134,838**	**29,890**	**4,708**	**9,387**	**623,394**	**124,435**	**12,178**	**6,169**
Biological anthropology	0	0	31	130	19	164	131	0	8
Social Anthropology and ethnology	4,491	3,830	375	155	550	8,867	1,139	181	104
Archaeology and Prehistory	782	1,897	311	137	158	5,991	560	109	118
Architecture, space management	529	2,599	331	139	642	8,313	926	29	70
Art and art history	3,248	2,194	133	136	305	10,676	2,422	189	182
Classical studies	289	246	130	134	131	691	468	29	61
Demography	1,988	1,972	1,182	135	132	9,175	934	3	63
Law	793	8,046	659	133	539	16,862	4,227	107	266
Economies and finances	328	11,274	1,009	458	226	80,513	7,595	185	102
Education	1,793	20,383	2,024	401	355	61,397	22,672	417	524
Environmental studies	4,088	3,928	596	139	131	85,692	4,987	751	104
Gender studies	2,679	1,049	908	135	132	1,480	669	23	129
Geography	136	4,135	810	135	754	46,188	5,817	1,225	113
Management	1,474	3,074	1,256	134	169	11,691	1,962	35	47
History, Philosophy and Sociology of Sciences	2,034	7,704	1,251	210	1,059	25,442	8,225	1,850	145
History	1,508	5,564	1,639	213	297	12,328	3,224	12	217
Communication sciences	476	11,018	1,266	144	1,454	43,438	15,273	2,818	111
Linguistics	479	5,283	498	132	708	17,990	5,844	2,224	334
Literature	1,523	658	833	132	132	17,386	1,734	308	211
Cultural heritage and museology	1,386	128	134	133	131	1,493	680	0	135
Musicology and performing arts	261	391	150	134	130	2,329	470	33	106

Philosophy	1,801	4,198	160	190	399	17,432	3,274	527	164
Psychology	1,093	6,284	1,399	141	207	34,342	10,100	32	1,396
Religions	3,593	728	135	180	145	19,144	2,205	62	121
Political science	1,381	6,649	4,339	137	177	29,015	4,085	212	641
Sociology	173	21,562	6,908	322	174	52,751	14,749	817	657
Methods and statistics	133	44	1,423	139	131	2,604	63	0	40

Table 1. Numbers of papers collected per language et per discipline to train the machine

2| Classification model

Automatic classifying is a method of categorising a document using advanced methods based on statistics and language analysis.

Classifier uses artificial intelligence and machine learning to automatically detect the characteristics of documents. Each category (also referred here as tag or label) is associated with a unique signature that is subsequently used to select which tags to apply to new documents. Ongoing quality control provides a feedback loop that adjusts incorrect tagging, increasing precision over time.

2.1 The Antidot software

In the ISIDORE search assistant, Huma-Num has used a classifier tool which has been developed by Antidot[8] for the last ten years. Antidot is a software vendor that leverages 20 years of advanced research in semantic search and content enrichment.

To test the feasibility of automatic classification on 9 languages and 26 categories, we used this robust but expensive tool in the first part of the TRIPLE project. This classifier can apply any number of tags to each document and it is language agnostic. It employs cutting-edge machine learning algorithms, guaranteeing a precision of the outcome, even when only a limited reference corpus is available. Due to the proprietary nature of the tool we are not able to provide a detailed methodology.

The *afs_classifier_train*[9], used here, allows us to generate a classifier database model. It takes XML documents as input, containing text and one or several tags. It outputs a classifier database allowing us to tag new "free" documents. The process is similar to that of the "spam filter" mechanism in a mail client.

[8] https://antidot.net

[9] https://doc.antidot.net/r/AFS-Filters-Description/Linguistic-Filters/afs_classifier_train

2.2 Training phase

In the learning phase, Classifier creates its semantic signature repository based on a subset of already tagged documents. Each XML document, built as indicated in the previous chapter, contains the three types of information needed to train the classifier:

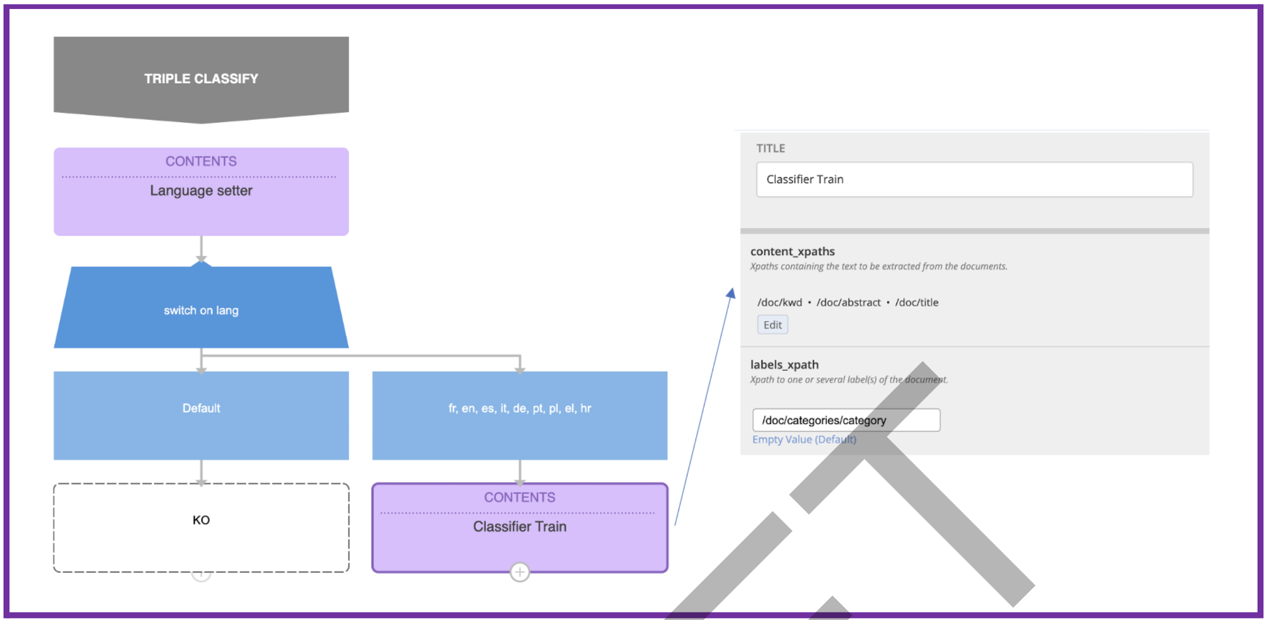
- text (title, abstract and keywords)
- language of the whole text
- one or more disciplines, which represent the categories, and follow the model presented below

```
<?xml version="1.0" encoding="UTF-8"?>
<doc>
  <lang></lang>
  <title></title>
  <abstract></abstract>
  <categories>
    <category></category>
  </categories>
  <kwd></kwd>
</doc>
```

Fig.1 TRIPLE Classification model

The pipeline used to build the classifier database in the Antidot framework is shown in Figure 2 below.

First, the language of the document is set according to the one written in the XPath /doc/lang (language setter). Then, only documents in the nine TRIPLE languages are kept (switch on lang). Finally, the Antidot classifier train filter is used on the document texts (XPath /doc/[title or abstract or kwd]) for each /doc/categories/category tags.



TITLE

Classifier Train

content_xpaths

Xpaths containing the text to be extracted from the documents.

/doc/kwd • /doc/abstract • /doc/title

Edit

labels_xpath

Xpath to one or several label(s) of the document.

/doc/categories/category

Empty Value (Default)

Figure 2: Pipeline of the classifier database

Documents are processed at a speed of about 400 items per second and the consolidation of the training database takes less than 1 minute. The system outputs the quality (F-score[10]) of the global database at 62% while for each language.

Language	de	en	es	fr	it	pl	pt	el	hr
Quality (F1) %	71	60	57	73	77	75	63	57	40

Table 2. Quality of the global database per language

2.3 Classifier validation

Based on the training database, we have developed a tool whose purpose is to classify in one of the MORESS categories a text based on its title and its summary in a given language. The URL of this application is :

https://rd.isidore.science/ondemand/en/classify.html

The title and abstract are mandatory to ask the system to provide a MORESS category or categories that fit the given document best. These two metadata can be requested from CrossRef by simply entering the DOI. The template to use is the one from the TRIPLE training. A probability for each discipline is also given.

[10] https://en.wikipedia.org/wiki/F1_score

3| Next steps

This exploratory work, conducted in WP2 (Data acquisition), will feed, into the future, work in WP4 dedicated to the creation of a classifier, having the same characteristics and purposes of the one presented herein but based on open source technologies.

We will certainly test the possibilities of Keras API[11] as a candidate. This framework allows for creating, training and using neural networks.

The architecture we envisage for this future enrichment filter in the pipeline, will be in the form of a web service, accessible through an API. This will make it possible to offer, elsewhere than in the TRIPLE pipeline, the possibility of automatically classifying SSH documents in the MORESS categories.

[11] https://keras.io