DARIAH-EU

Use of vocabularies for metadata curation and quality assessment in Social Sciences and Humanities

Triple

How to use vocabularies to enrich GoTriple ?

Discover
Connect
Collaborate

Triple

Julien Homo **(Foxcub)** | March, 27th 2023 | Topical vocabularies used in the SSH

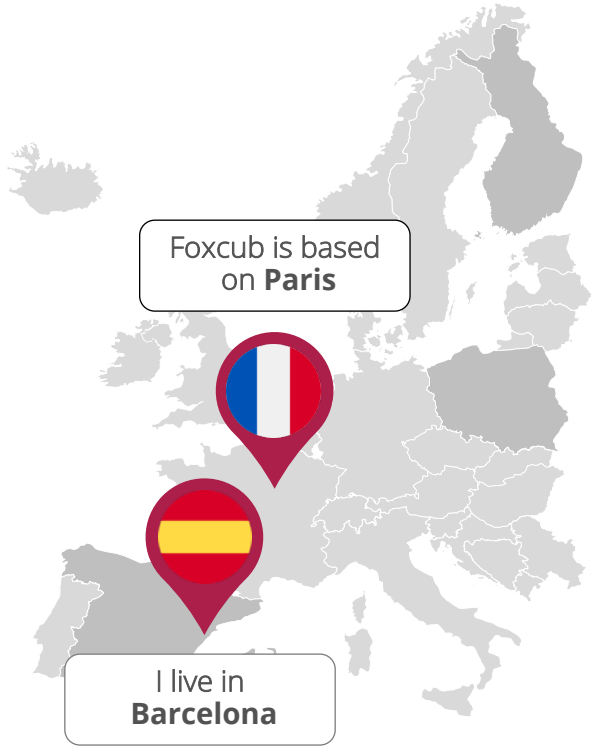
I am your speaker for this session

JULIEN HOMO

Senior Data Architect
AI Expert

foxcub.
12 years of experience
in Data

Specialized in
Public sector



Discover
Connect
Collaborate

Triple

How to use vocabularies to enrich GoTriple ?

1. **Enriching GoTriple data thanks to vocabularies ?**
2. The TRIPLE Vocabulary and MORESS
3. The annotation and classification GoTriple services based on vocabularies
4. Conclusion & future work
5. Questions

Discover
Connect
Collaborate

Triple

Definition of a vocabulary in GoTriple

A collection of **semantic resources (or terms)** usually specific to a **subject domain** and created following a **particular methodology**. The semantic resources of a vocabulary are frequently organised hierarchically and may be interlinked forming a whole network of terms.

Thesauri

Vocabularies of terms that are usually domain-specific, organised hierarchically in one or more levels, and containing synonyms, antonyms, etc.

Getty AAT...

Classifications and classification schemes

Presents terms in thematic categories. Classification schemes often contain serial alphanumeric values

DDC, LCC, MORESS...

Authority files

List of standardised terms for person names, corporate bodies or concepts commonly used within a domain

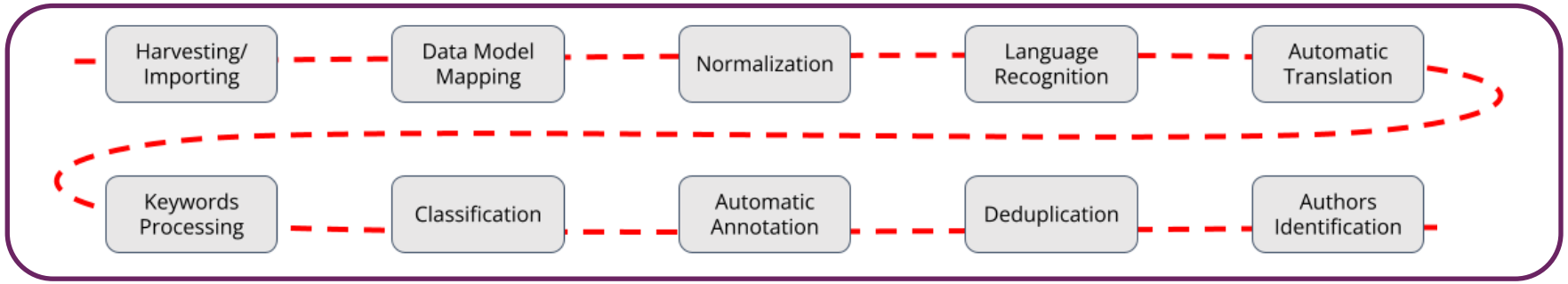
VIAF, LCSH, MeSH...

Discover
Connect
Collaborate

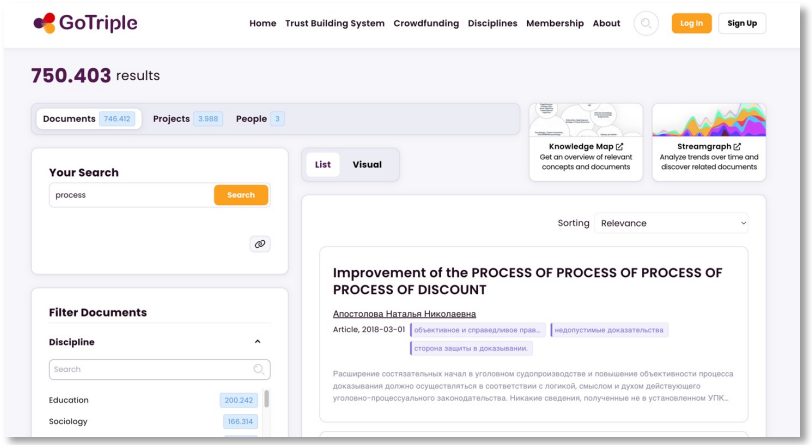
Triple

The GoTriple platform

GoTriple pipeline (SCRE)



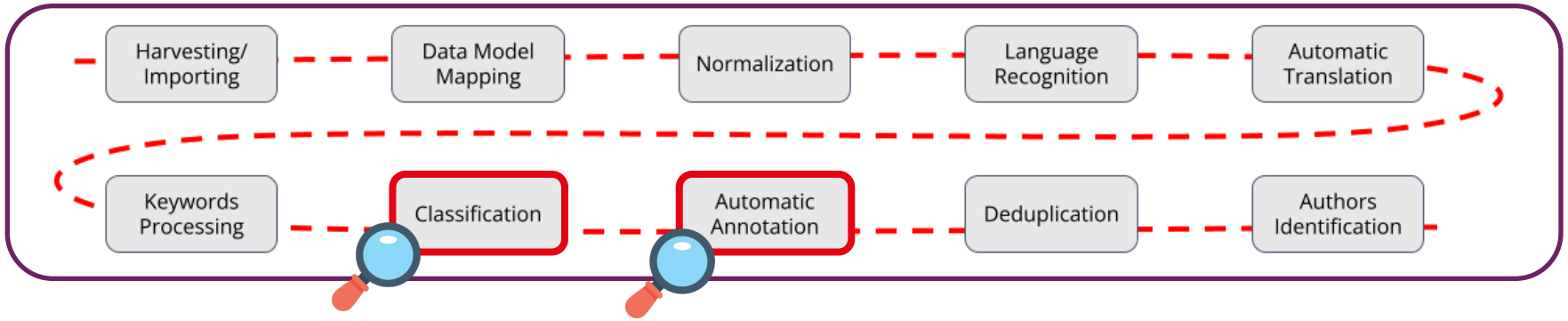
What are the vocabularies involved and where are they (in the GoTriple pipeline) ?



Discover
Connect
Collaborate

Triple

Classification and annotation: 2 vocabularies



Classification scheme

MORESS Vocabulary

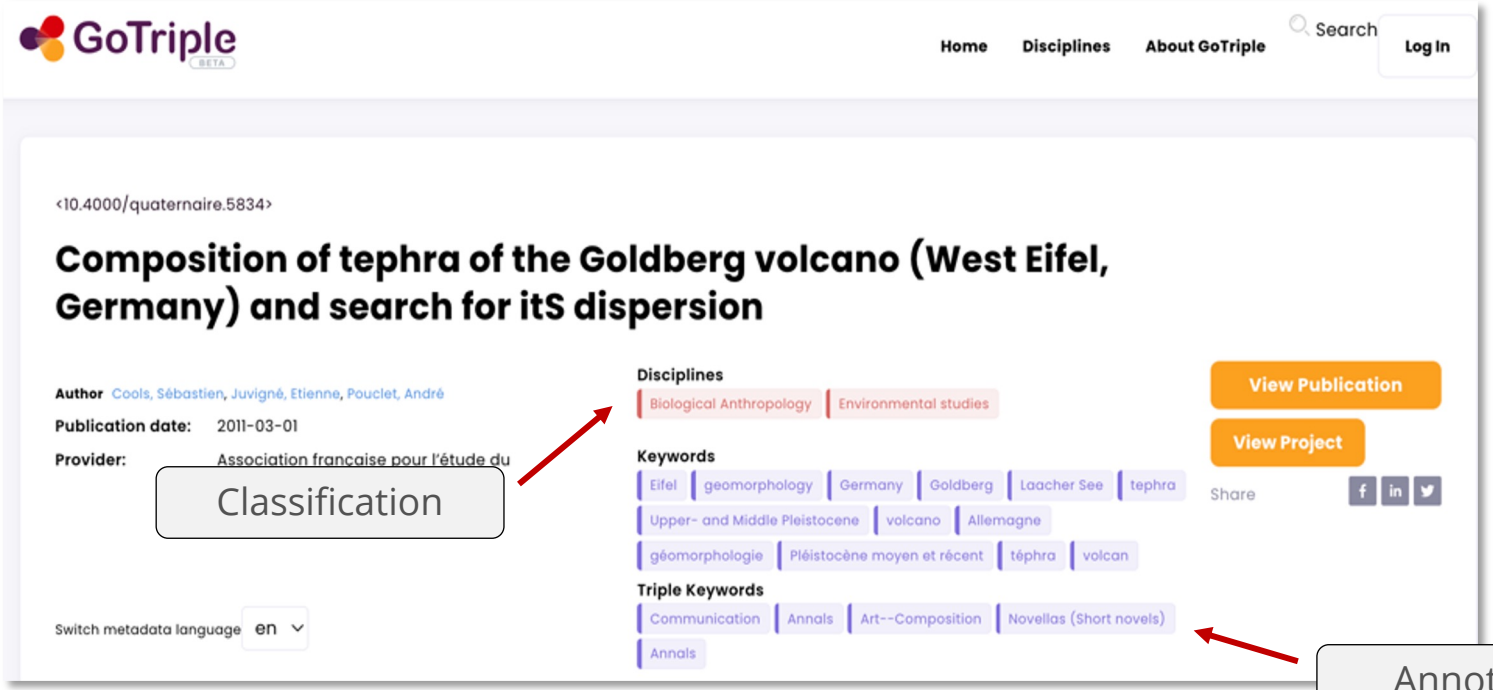
Classification

Thesauri

NEW

TRIPLE Vocabulary

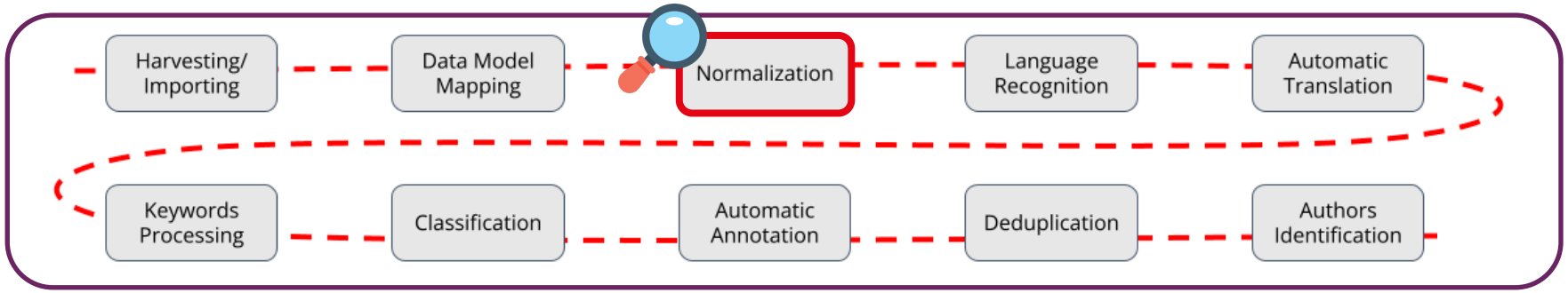
Annotation



Discover Connect Collaborate

Triple

Other vocabularies involved in the GoTriple pipeline



Classification scheme

TRIPLE Language Vocabulary

Normalization

- Croatian (hr)
- Catalan (ca)
- English (en)
- French (fr)
- German (de)
- Greek (el)

Classification scheme

COAR Resource Types Vocabulary

Normalization

TRIPLE Document Type	Codification	COAR Resource Type
Article	typ_article	Journal Article: https://vocabularies.coar-repositories.org/resource_types/c_6501/
Bibliography	typ_bibliography	Bibliography: https://vocabularies.coar-repositories.org/resource_types/c_86bc/
Blog post	typ_blog-post	Blog post: https://vocabularies.coar-repositories.org/resource_types/c_6947/

Classification scheme

TRIPLE Licences and Access Rights Vocabulary

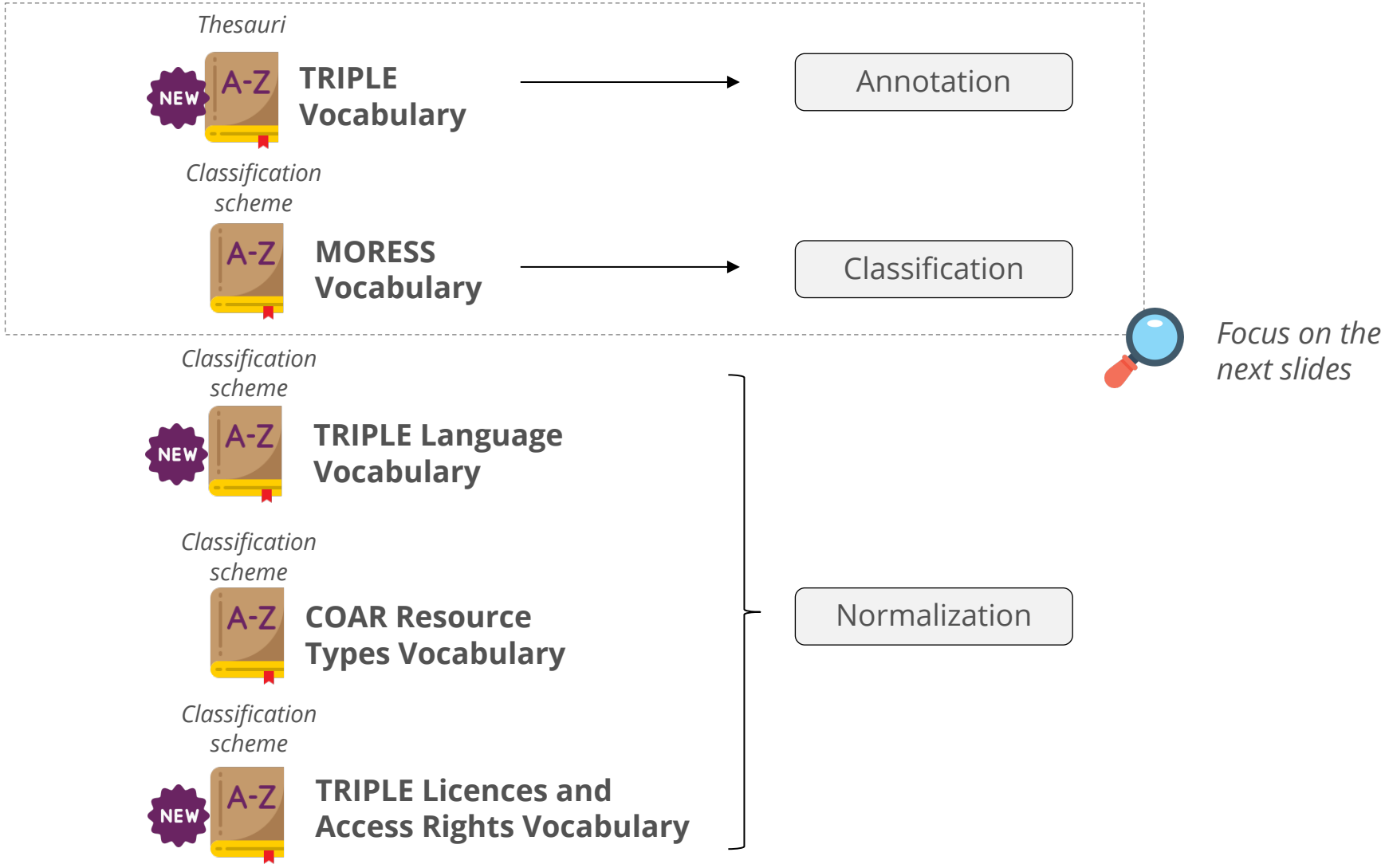
Normalization

TRIPLE License	Codification
CAIRN	lic_cairn
Creative Commons Various spelling and acronyms, e.g. CC0, CCBY,..., and full Creative Commons URLs as well..	lic_creative-commons
Open source Various spelling and licence names, e.g. Apache, GPL, BSD, MIT	lic_open-source

Discover Connect Collaborate

Triple

There is a strong link between the vocabulary and the service or application that consumes it



Discover
Connect
Collaborate

Triple

How to use vocabularies to enrich GoTriple ?

1. Enriching GoTriple data thanks to vocabularies ?
2. **The TRIPLE Vocabulary and MORESS**
3. The annotation and classification GoTriple services based on vocabularies
4. Conclusion & future work
5. Questions

Discover
Connect
Collaborate

Triple

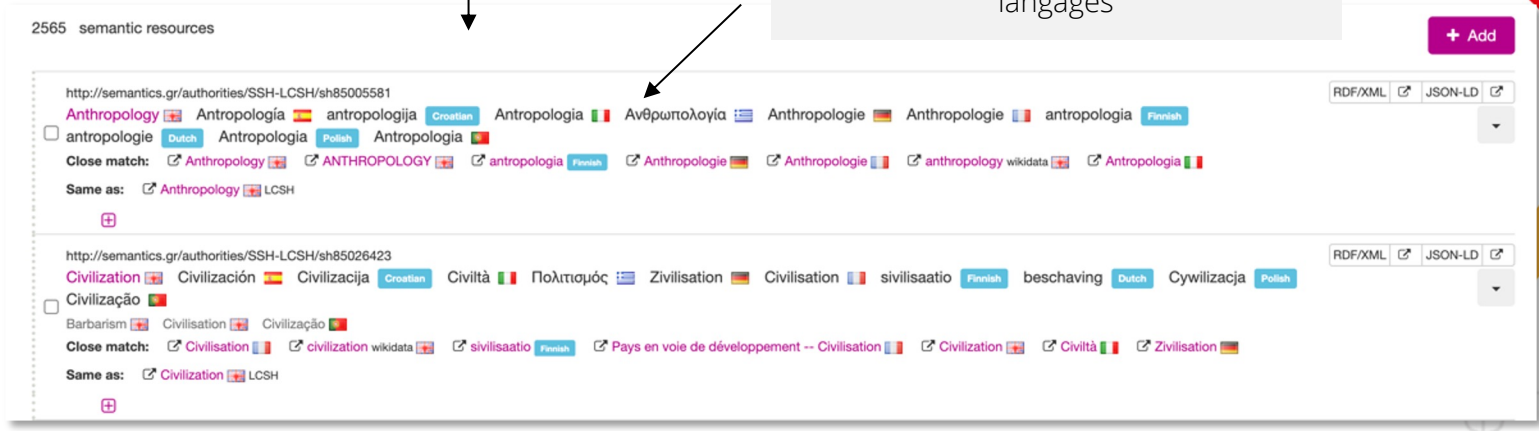
The TRIPLE Vocabulary: overview (1/2)

Semantics.gr — Platform for publishing semantic resources as Linked Open Data
Vocabularies - Thesauri - Authority files

The Triple Vocabulary is published through the Semantics.gr platform

More than **3300 concepts** are currently available in the vocabulary

The labels are currently translated in 10 langages

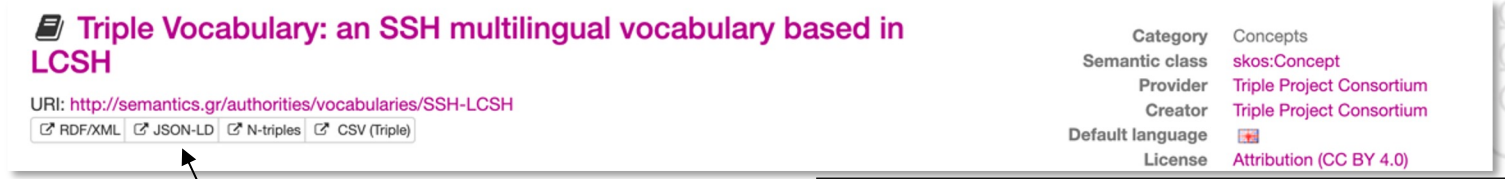


Thesauri

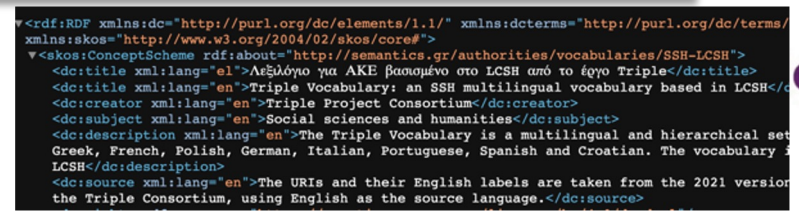
TRIPLE Vocabulary

The Triple Vocabulary is a **multilingual and hierarchical set of SSH-related concepts**. It is a subset of LCSH (Library of Congress Subject Headings) that cover popular SSH aspects and is enhanced with labels in Greek, French, Polish, German, Italian, Portuguese, Spanish, Croatian, Slovenian).

The vocabulary is used for **the automatic annotation of the publications hosted in the GoTriple platform**



The vocabulary can be extracted in 4 different formats :
RDF/SKOS, JSON-LD, N-triples and CSV

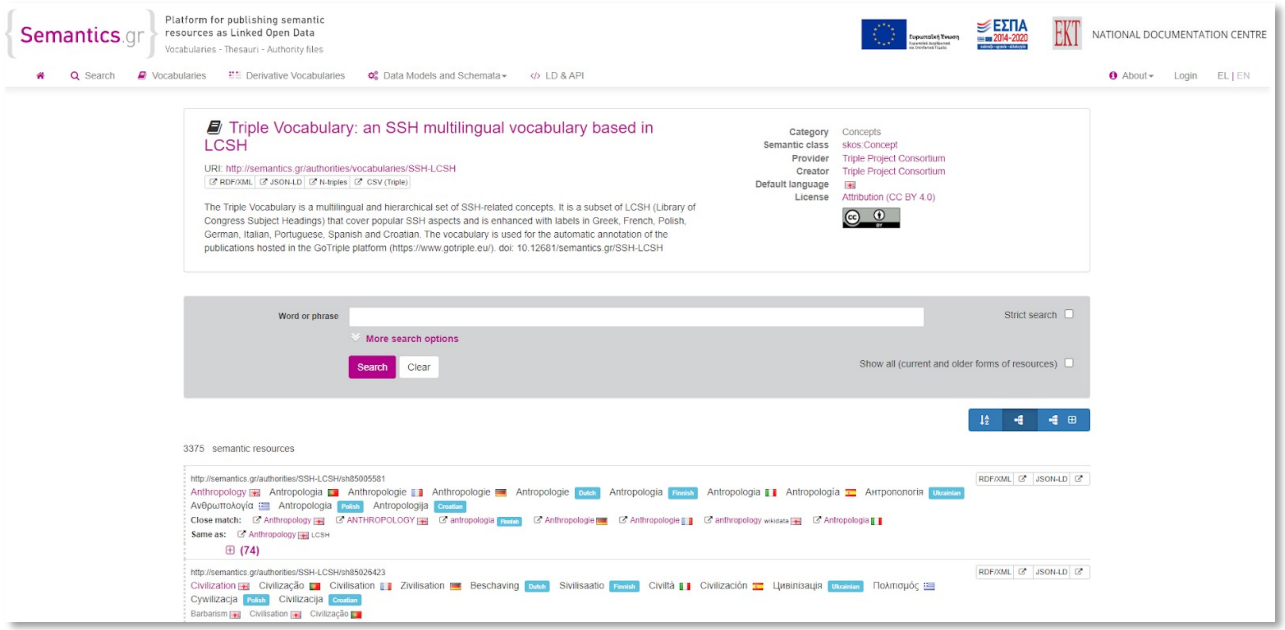


The TRIPLE Vocabulary: overview (2/2)

The data model of the GoTriple Vocabulary is **SKOS**, a standard for data modeling vocabularies and thesauri. SKOS is an RDF Data Model. It defines classes and properties to describe concepts and their relationship. All concepts or the GoTriple Vocabulary are instances of the skos:Concept class.

A **dedicated schema** was created in Semantics.gr based on SKOS to which only the GoTriple Vocabulary conforms. This way, new properties can be easily in the future, even custom ones, without affecting other SKOS vocabularies in Semantics.gr.

http://semantics.gr/authorities/vocabularies/SSH-LCSH



Class	skos:Concept			
Properties	**qualified name**	**description**	**type**	**multivalued**
	skos:prefLabel	preferable labels (one per language)	Literal	TRUE (one per language)
	skos:altLabel	alternative labels	Literal	TRUE
	skos:browder	URI Ref to broader concept(s) from the same vocabulary. The property defines the hierarchy. Its symmetric property is skos:narrower.	URI Ref	TRUE
	skos:narrower	URI Ref to narrower concept(s) from the same vocabulary. The property defines the hierarchy. Its symmetric property is skos:browder.	URI Ref	TRUE
	skos:closeMatch	URI Refs to similar resources (with similar meaning) of external vocabularies	URI Ref	TRUE
	skos:exactMatch	URI Refs to exact resources (with exact meaning) of external vocabularies	URI Ref	TRUE

The MORESS Vocabulary: overview

Classification scheme

MORESS Vocabulary

The MORESS vocabulary (Mapping of Research in European Social Sciences and humanities) is mapping to support the development of the European Research Area through improving information on pan-European research in social sciences and humanities. This initiative is part of an EC-funded main program “FP5-HUMAN POTENTIAL - Programme for research, technological development and demonstration on "Improving the human research potential and the socio-economic knowledge base" (1998-2002)”

The MORESS Vocabulary **contains 27 categories** which are used to classify the GoTriples document with their associated **“discipline”.**

https://cordis.europa.eu/project/id/HPSE-CT-2002-60060/fr

MORESS IDs	Labels (FR, EN, ES)	DBPedia / Calenda mapping	HAL mapping
anthro-bio	Biological anthropology	http://dbpedia.org/resource/Biological_anthropology	http://aurehal.archives-ouvertes.fr/subject/shs.anthro-bio
anthro-se	Social Anthropology and ethnology	http://calenda.org/categories.rdf#categorie213 http://dbpedia.org/resource/Social_anthropology	http://aurehal.archives-ouvertes.fr/subject/shs.anthro-se
archeo	Archaeology and Prehistory	http://calenda.org/categories.rdf#categorie293 http://dbpedia.org/resource/Archaeology	http://aurehal.archives-ouvertes.fr/subject/shs.archeo
archi	Architecture, space management	http://dbpedia.org/resource/Architecture	http://aurehal.archives-ouvertes.fr/subject/shs.archi
art	Art and art history	http://calenda.org/categories.rdf#categorie278 http://dbpedia.org/resource/Art_history	http://aurehal.archives-ouvertes.fr/subject/shs.art
class	Classical studies	http://dbpedia.org/resource/Classics	http://aurehal.archives-ouvertes.fr/subject/shs.class
demo	Demography	http://calenda.org/categories.rdf#categorie203 http://dbpedia.org/resource/Demography	http://aurehal.archives-ouvertes.fr/subject/shs.demo
droit	Law	http://calenda.org/categories.rdf#categorie251 http://dbpedia.org/resource/Law	http://aurehal.archives-ouvertes.fr/subject/shs.droit
eco	Economies and finances	http://calenda.org/categories.rdf#categorie236 http://dbpedia.org/resource/Economy	http://aurehal.archives-ouvertes.fr/subject/shs.eco
edu	Education Educación	http://calenda.org/categories.rdf#categorie283 http://dbpedia.org/resource/Education	http://aurehal.archives-ouvertes.fr/subject/shs.edu
envir	Environmental studies	http://dbpedia.org/resource/Environmental_studies	http://aurehal.archives-ouvertes.fr/subject/shs.envir
genre	Gender studies	http://calenda.org/categories.rdf#categorie205 http://dbpedia.org/resource/Gender_studies	http://aurehal.archives-ouvertes.fr/subject/shs.genre
geo	Geography	http://calenda.org/categories.rdf#categorie218 http://dbpedia.org/resource/Geography	http://aurehal.archives-ouvertes.fr/subject/shs.geo
manag	Management	http://calenda.org/categories.rdf#categorie239 http://dbpedia.org/resource/Management	http://aurehal.archives-ouvertes.fr/subject/shs.gestion
hisphilso	History, Philosophy and Sociology of Sciences	http://dbpedia.org/resource/History_of_science http://dbpedia.org/resource/Philosophy_of_science http://dbpedia.org/resource/Sociology_of_scientific_knowledge	http://aurehal.archives-ouvertes.fr/subject/shs.hisphilso
hist	History	http://calenda.org/categories.rdf#categorie228 http://dbpedia.org/resource/History	http://aurehal.archives-ouvertes.fr/subject/shs.hist
info	Communication sciences	http://calenda.org/categories.rdf#categorie271 http://dbpedia.org/resource/Communication_sciences	http://aurehal.archives-ouvertes.fr/subject/shs.info
lang	Linguistics	http://calenda.org/categories.rdf#categorie268 http://dbpedia.org/resource/Linguistics	http://aurehal.archives-ouvertes.fr/subject/shs.langue
litt	Literature	http://calenda.org/categories.rdf#categorie269 http://dbpedia.org/resource/Literature	http://aurehal.archives-ouvertes.fr/subject/shs.litt
museo	Cultural heritage and museology	http://dbpedia.org/resource/Museology	http://aurehal.archives-ouvertes.fr/subject/shs.museo
musiq	Musicology and performing arts	http://dbpedia.org/resource/Musicology http://dbpedia.org/resource/Performing_arts	http://aurehal.archives-ouvertes.fr/subject/shs.musiq
phil	Philosophy	http://calenda.org/categories.rdf#categorie261 http://dbpedia.org/resource/Philosophy	http://aurehal.archives-ouvertes.fr/subject/shs.phil
psy	Psychology	http://calenda.org/categories.rdf#categorie266 http://dbpedia.org/resource/Psychology	http://aurehal.archives-ouvertes.fr/subject/shs.psy
relig	Religions	http://dbpedia.org/resource/Religion	http://aurehal.archives-ouvertes.fr/subject/shs.relig
scipo	Political science	http://calenda.org/categories.rdf#categorie242 http://dbpedia.org/resource/Political_science	http://aurehal.archives-ouvertes.fr/subject/shs.scipo
socio	Sociology	http://calenda.org/categories.rdf#categorie201 http://dbpedia.org/resource/Sociology	http://aurehal.archives-ouvertes.fr/subject/shs.socio
stat	Methods and statistics	http://dbpedia.org/resource/Statistics	http://aurehal.archives-ouvertes.fr/subject/shs.stat

How to use vocabularies to enrich GoTriple ?

1. Enriching GoTriple data thanks to vocabularies ?
2. The TRIPLE Vocabulary and MORESS
3. **The annotation and classification GoTriple services based on vocabularies**
4. Conclusion & future work
5. Questions

Discover
Connect
Collaborate

Triple

TRIPLE Vocabulary and annotations: build

Annotation

GoTriple uses a **semantic annotation service that**, after being deployed and linked with the **TRIPLE Vocabulary**, tags all TRIPLE keywords in documents in GoTriple.

Semantic rules and applied normalizations:

- Case sensitive or insensitive

Example : "Geography" = "geography"

- Accent sensitive or insensitive

Example : "géographie" = "geographie"

- Inflection sensitive or insensitive

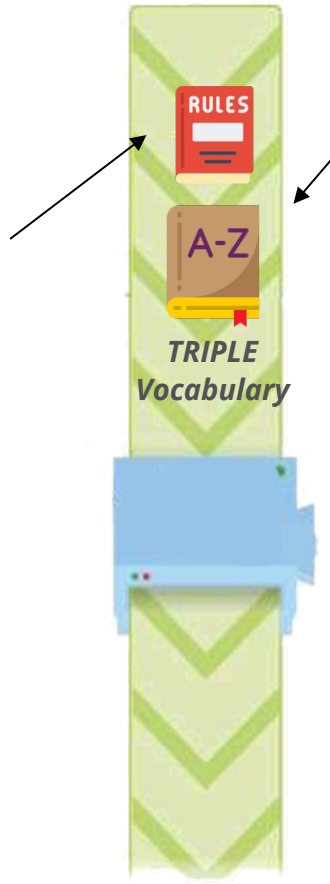
Example : "géographies" = "géographie"

- *Stop words remiving*

Example: " The Geography" => try to find " Geography"

- *Inflections*

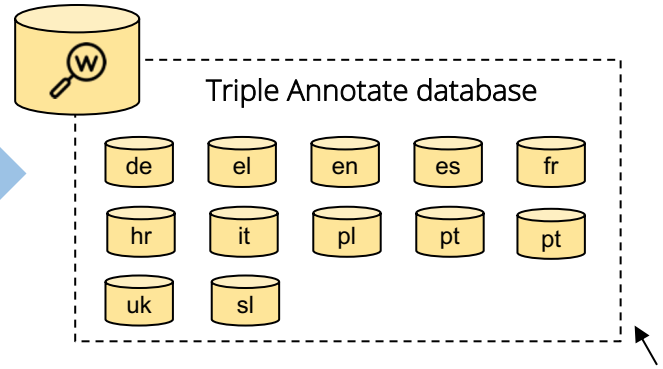
Example: "History Geography" => try to find " Geography History"



Index Job

TRIPLE Vocabulary
SKOS vocabulary to be index

All the SKOS labels (prefLabel, altLabel), closeMatch and sameAs from this vocabulary have to be indexed



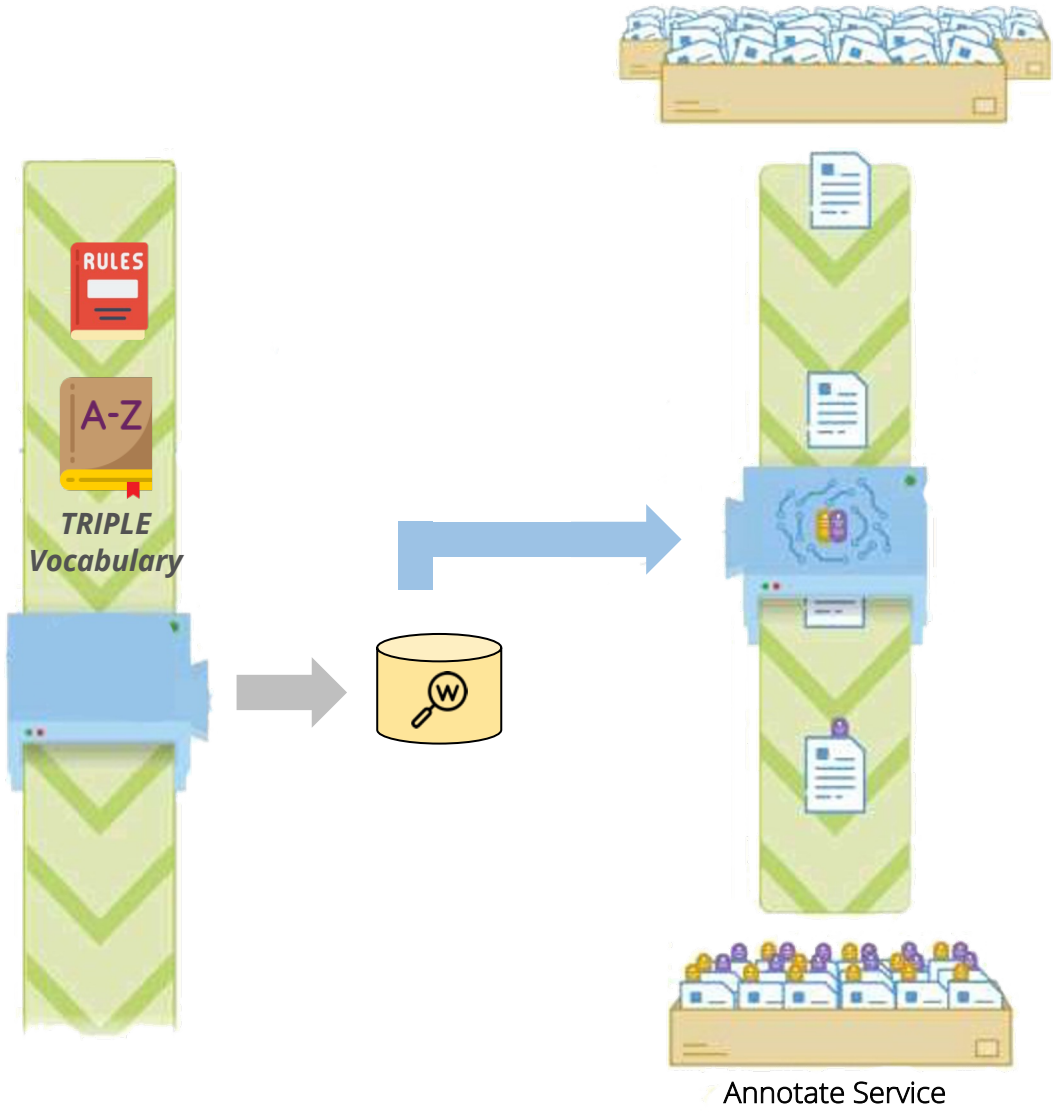
Triple Annotate database : the database need to be indexed in the 11 different langages in order to allow annotation of new documents in its associated langage. When a localized document is annotated, only the corresponding localized annotate database will be used.

Discover Connect Collaborate

Triple

TRIPLE Vocabulary and annotations: serving

Annotation



Documents to annotate

These documents are sent to a **specific API** with **2 inputs** :

- Metadata : lang
- Plain text (could be title, description, abstract or full text)

The **Annotate Service** is a script which are able to request the Triple indexed database from a set of metadata and the full text of the document. This script is the link between the annotation database and the API.

The Annotate Service will **analyze input text to detect if labels** (and extended labsl based on semantic rules) **from the Triple Vocabulary are present in the plain text**. This detection is based on **exact matches + semantic rules (case/accents/flexions sensitivity)**

Connect Collaborate

Triple

TRIPLE Vocabulary and annotations: challenging points

Annotation

The TRIPLE Vocabulary is built on more than 3,300 concepts. Therefore, the Annotate Service can return many results. We only need to filter the most relevant TRIPLE vocabulary tags between the different matched labels.

Strategy #1
Filter TRIPLE vocabulary tags that are “blacklisted”

Example :
- Concepts with labels that are too generic (Example: “value”, “values”...)
- Concepts with labels built on many stopwords (Example: “the New”, ...)
- Concepts with labels based on a combination of the 2 last assumptions (Example: “Other (Philosophy)”)

Strategy #2
Filter TRIPLE Vocabulary tags that are not “closed together”

Complete example

Concepts with their top concepts found in a document:
- Concepts: ['Civilization']
- Contracts: ['Law']
- Democracy: ['Civilization']
- Drawing: ['Civilization']
- Fiction: ['Philology']
- Kings and rulers: ['Civilization']
- Plots (Drama, novel, etc.): ['Philology']
- Political science: ['Civilization']
- Politics, Practical: ['Civilization']
- Population: ['Civilization']
- Value: ['Civilization', 'Anthropology’]

Filter 1 blacklist:
Blacklisted concepts:
- Value: ['Civilization', 'Anthropology']
- Values: ['Civilization’]

Filter 2 “closed together”:
Top concepts score:
- ['Civilization: 10', 'Anthropology: 2', 'Law: 1', 'Philology: 2’]

We keep group of same-top-level keywords with size >= 3 :
- Final keywords = Biopolitics, Democracy, Political science, "Politics, Practical", Concepts, Drawing, Population

MORESS Vocabulary and classify: build

Classification

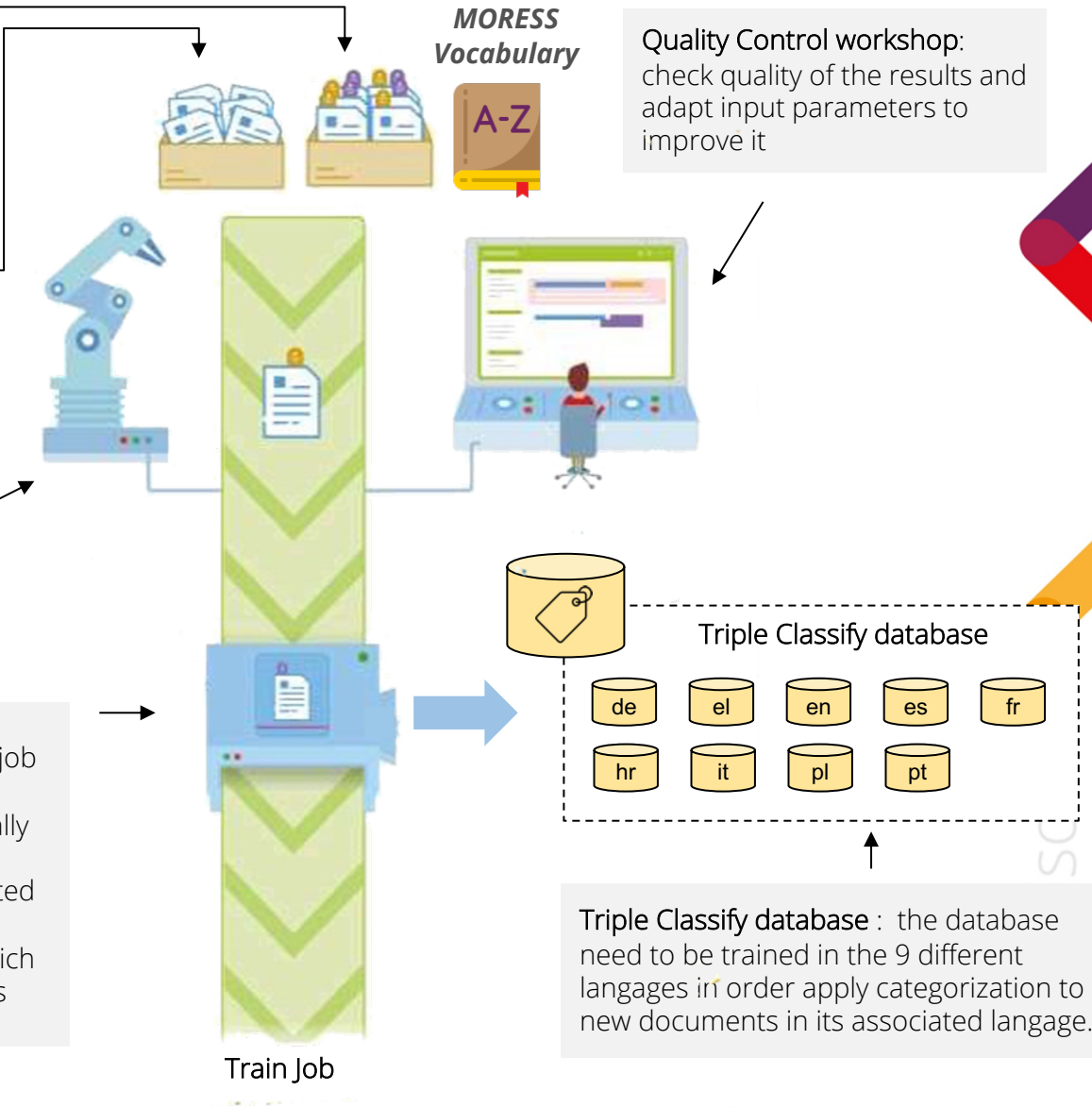
GoTriple uses a **semantic classify service** for classification, one or more “discipline”, that is the **27 MORESS categories** that have been selected in the TRIPLE project as representatives of the SSH domain

Triple Trained database + MORESS categories : inputs of the job

Triple Test database : subset of the Triple Trained database

Pre-processing:
XML to CSV
Merge
Lemmatization:
Cleaning (removing spaces, stopwords, numbers, emails, ...)
UTF8 normalization
N-gram (1,2)
Vectorization ...

Machine Learning model : this job uses artificial intelligence and machine learning to automatically detect the characteristics of documents. Each tag is associated with a unique signature that is subsequently used to select which tags to apply to new documents

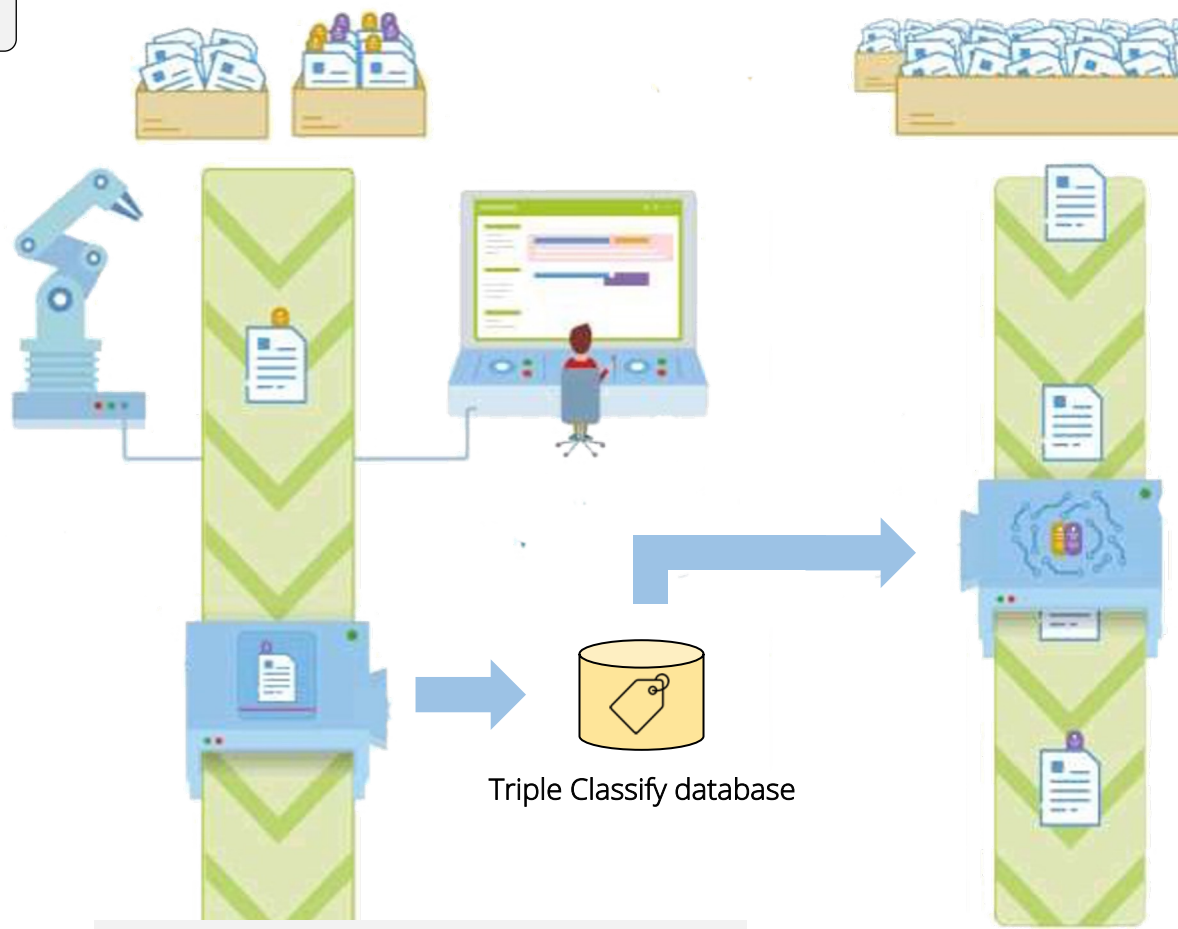


Quality Control workshop: check quality of the results and adapt input parameters to improve it

Triple Classify database : the database need to be trained in the 9 different langages in order apply categorization to new documents in its associated langage.

MORESS Vocabulary and classify: serving

Classification



Triple Classify database

For each document, a score is calculated **for each of the 27 MORESS categories.**

Classify Service

MORESS Vocabulary

New documents to tag

These documents have to be sent with **3 inputs** :

- Metadata : lang
- Theashold : float between 0 and 1
- Plain text (could be title, description, abstract or full text)

The **Classify Service** is a script which are able to request the Triple trained database from a set of metadata and the full text of the document. This script is the link between the prediction model based on MORESS and the API.

The prediction model will analyze its semantic proximity to the different MORESS categories and produce a score.

Discover Connect Collaborate

Triple

MORESS Vocabulary and classify: challenging points

Classification

1. The Classify Service can return all predictions in every MORESS category. We need to filter only the most relevant categories between the different scores of predictions.

Strategy #1
Filter categories with a score higher than a predefined threshold to maximize the quality rather than the number of categories, even if it means having more documents without categories. The goal is to maximize a user's trust so that they do not lose overall trust for the entire feature.

Strategy #2
Filter categories that have no more than 2 categories, as this is considered irrelevant in the context of the documentary corpus from SSH.

Some categories are overrepresented or underrepresented in the train model.

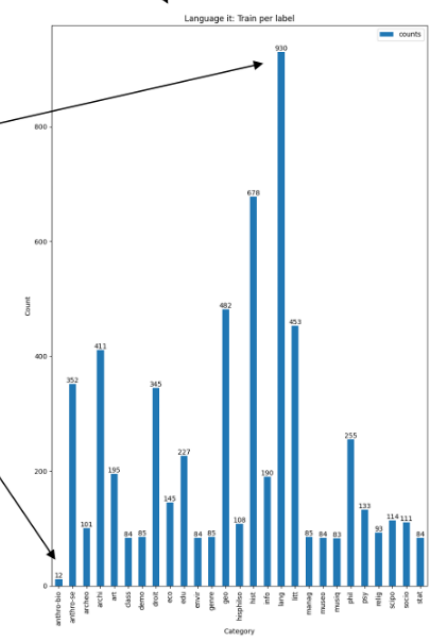
Strategy
In the process of evaluation of the train model (the Quality Control Workshop) the MORESS Distribution measure allows us to understand which categories are overrepresented or underrepresented in the train model.

Results with the italien model

« 930 documents were used to train the lang category »

« Only 12 documents were used to train the anthro-bio category »

MORESS Categories



Discover Connect Collaborate

Triple

How to use vocabularies to enrich GoTriple ?

1. Enriching GoTriple data thanks to vocabularies ?
2. The TRIPLE Vocabulary and MORESS
3. The annotation and classification GoTriple services based on vocabularies
4. **Conclusion & future work**
5. Questions

Discover
Connect
Collaborate

Triple

Conclusions and future work

Annotation

TRIPLE Vocabulary

- A balance of all input parameters has been found limiting noise in favour of more relevant tags
- The application of the blacklist filter composed of 100 entries allowed the removal of more than 30% of erroneous annotations
- The impact of the application of the “closed together” filter seems to have a positive impact on the overall quality of the results, but remains difficult to assess.

- Use a Machine Learning-based algorithm
- Cross the results of the Annotate Service with the results of the Classify Service

Classification

MORESS Vocabulary

- Globally, the models (1 model for each language) are efficient: on average, the F1-Score of the 11 x 27 categories > 50%
- The (relative) perception in terms of user satisfaction is rated "positive" to "very positive". Overall, users have confidence in the discipline displayed to them. This feeling can be partly explained by the use of the confidence threshold and the limit of the number of results to 2.
- The results are unequal between the models (ex: HR has better results than FR)
- The results are unequal between categories (ex: anthro-bio VS lang)

- Review the content of the training base and the training corpus to complete and enrich it (missing fields, representativeness of categories, etc.). This is important work but could considerably improve the MORESS distribution and quality of the results observed.
- Review the confidence threshold and the max limit (=2)

Discover Connect Collaborate

Triple

Triple

Julien Homo – julien.homo@foxcub.fr

foxcub.

TRIPLE will be a dedicated service of the OPERAS RI

OPERAS
open access in the european research area through scholarly communication

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 863420

Collaborate

ple