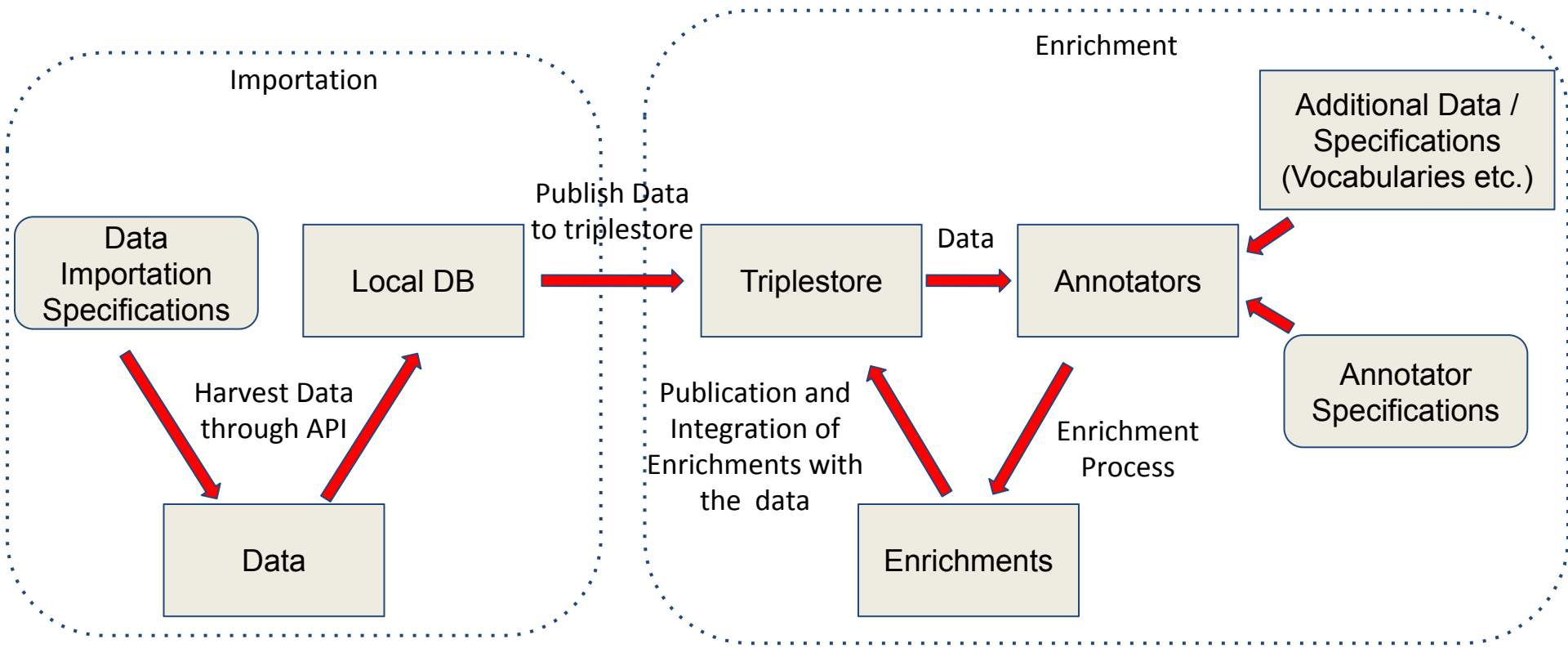
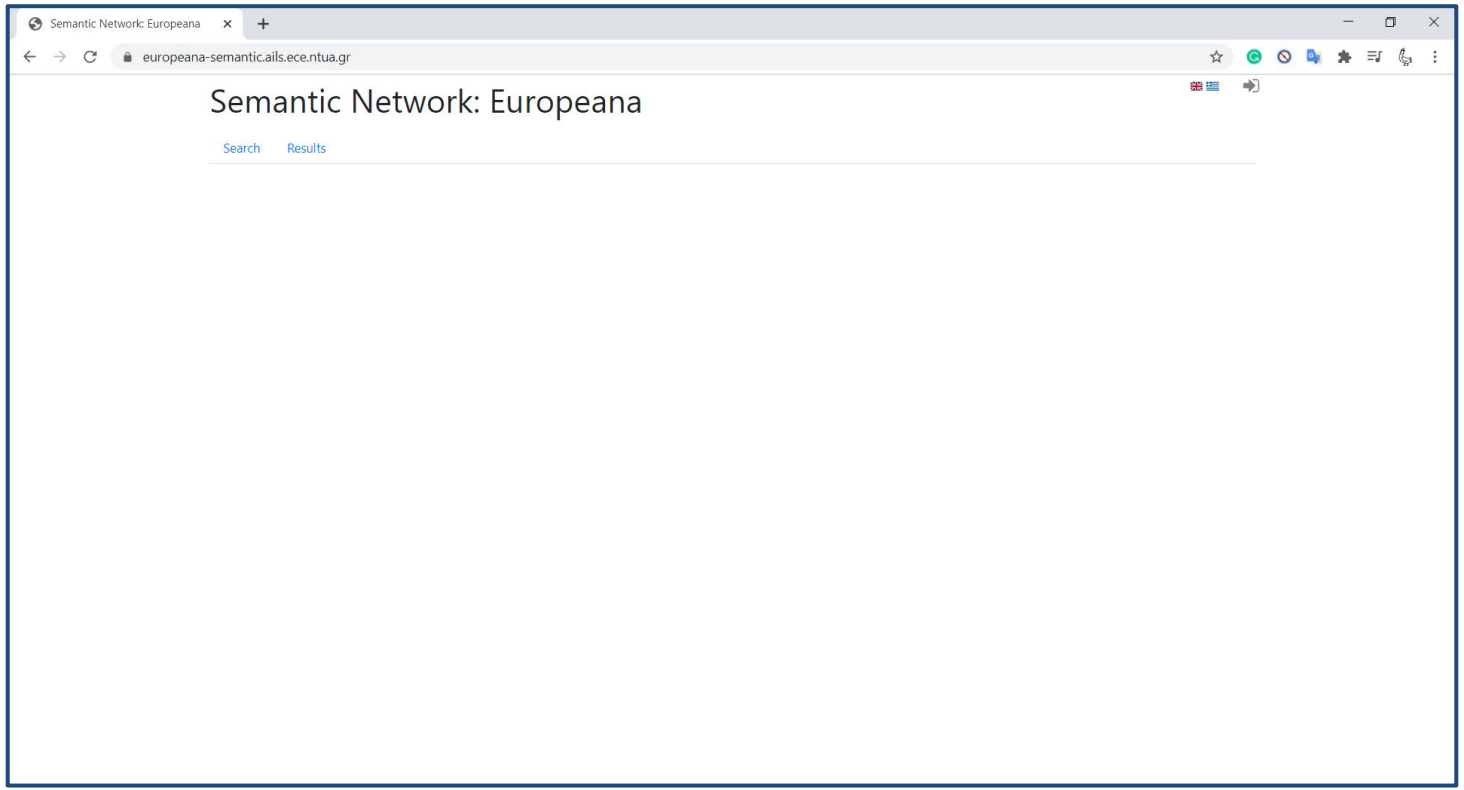
Auto Enrichment Tool

Orfeas Menis - Mastromichalakis

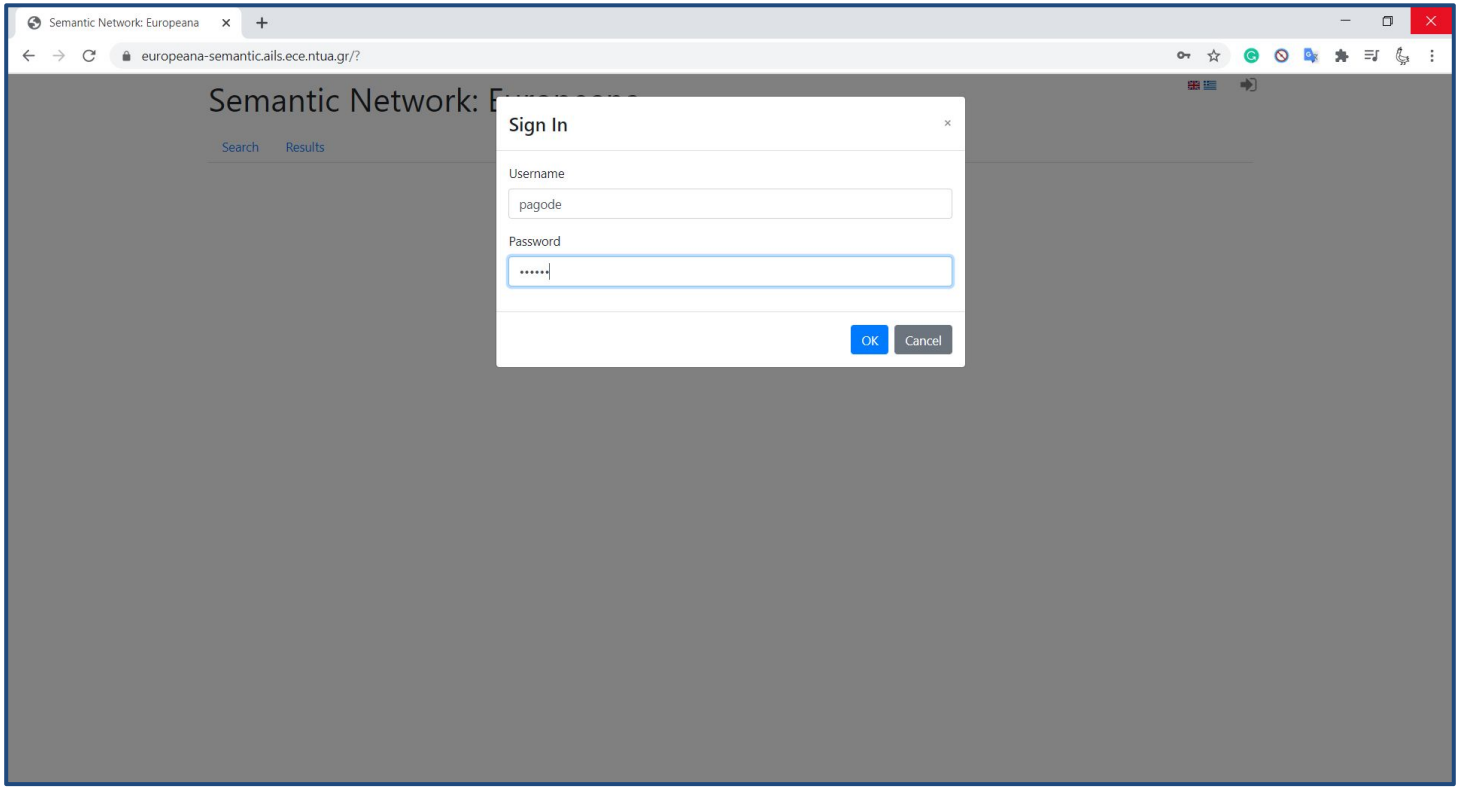
Workflow Overview



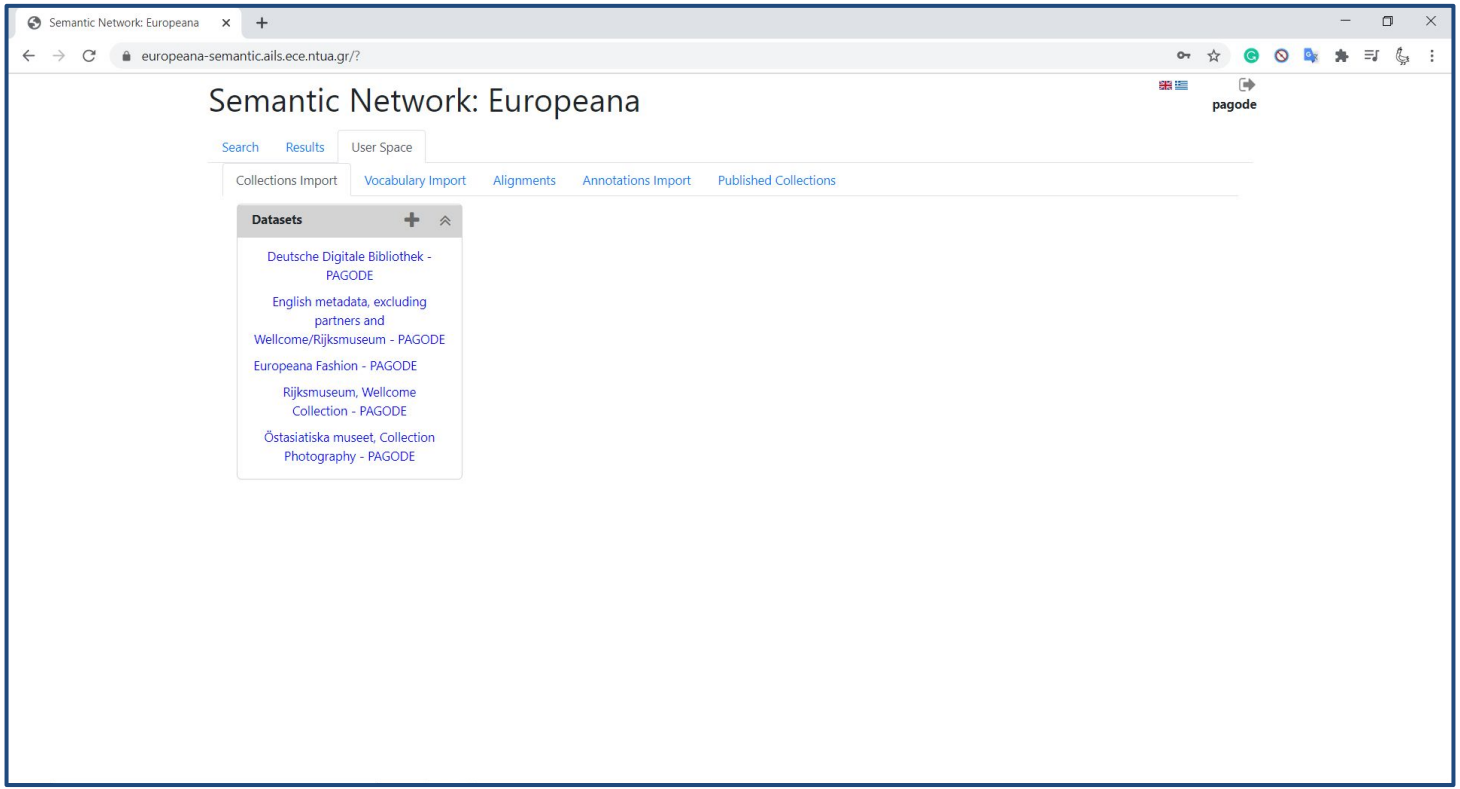
Main Page - Visitor



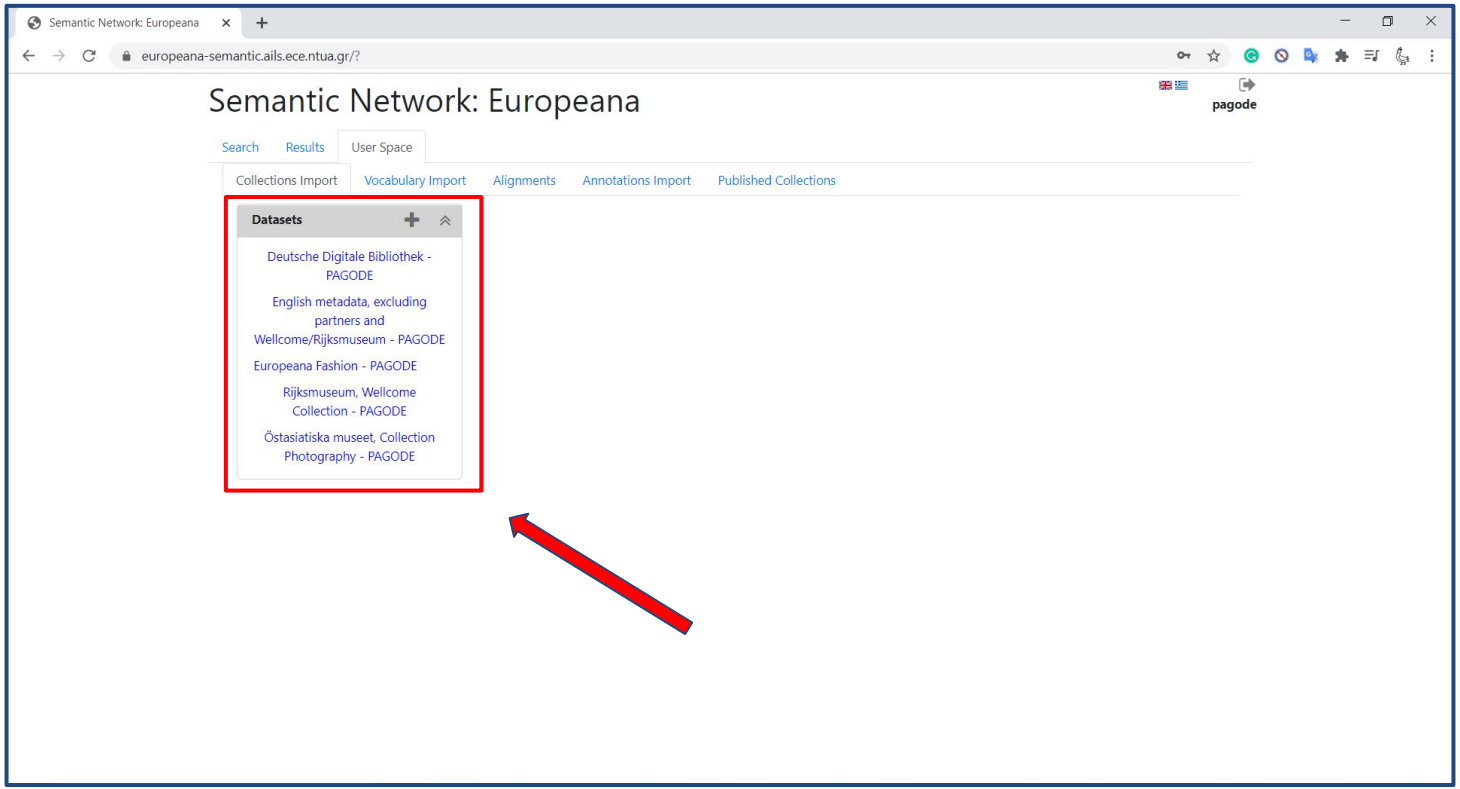
Login Page


Semantic Network: Europeana
europeana-semantic.ails.ece.ntua.gr/?
Semantic Network:
Search
Results
Sign In
Username
pagode
Password
OK
Cancel

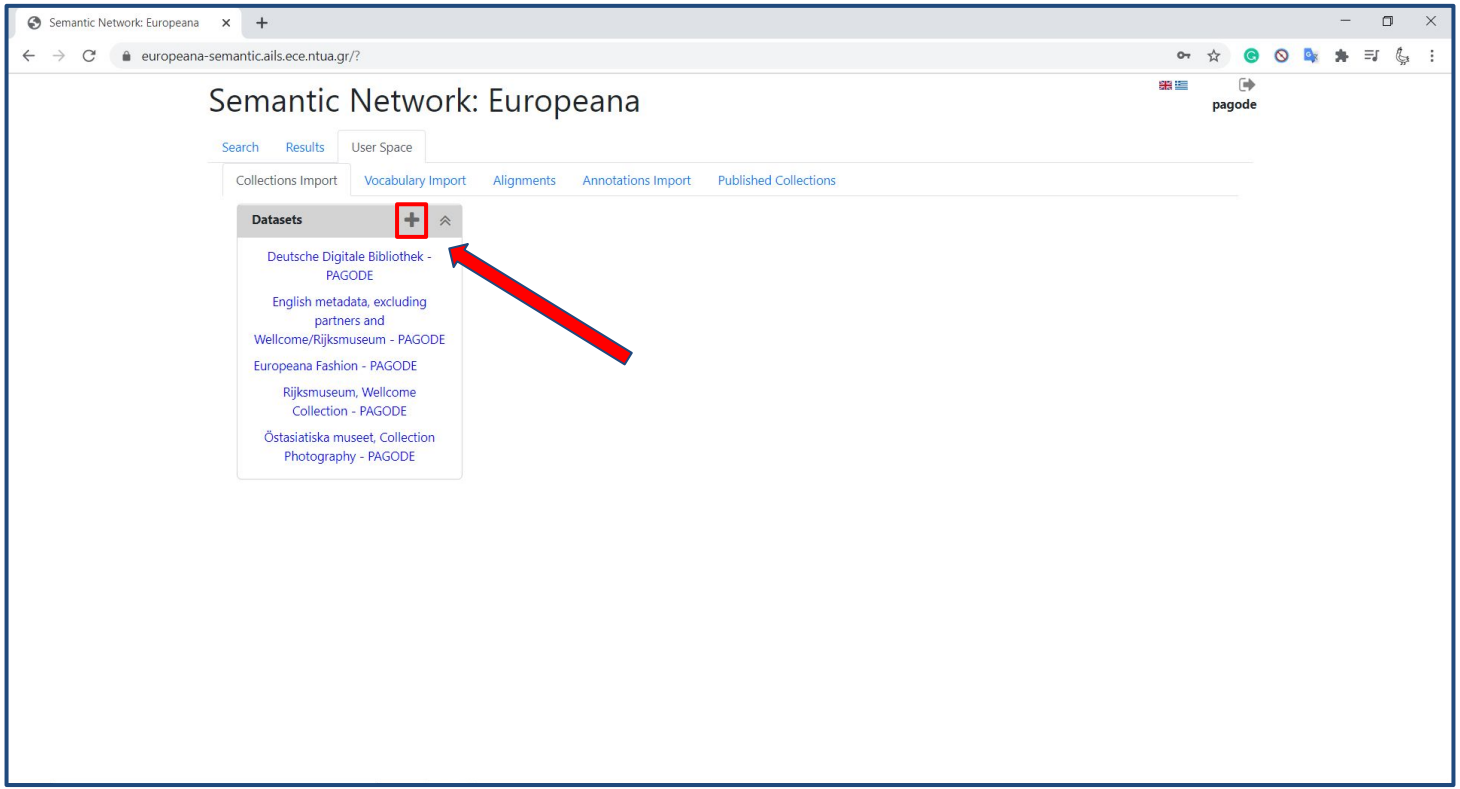
User Space - Collections Import



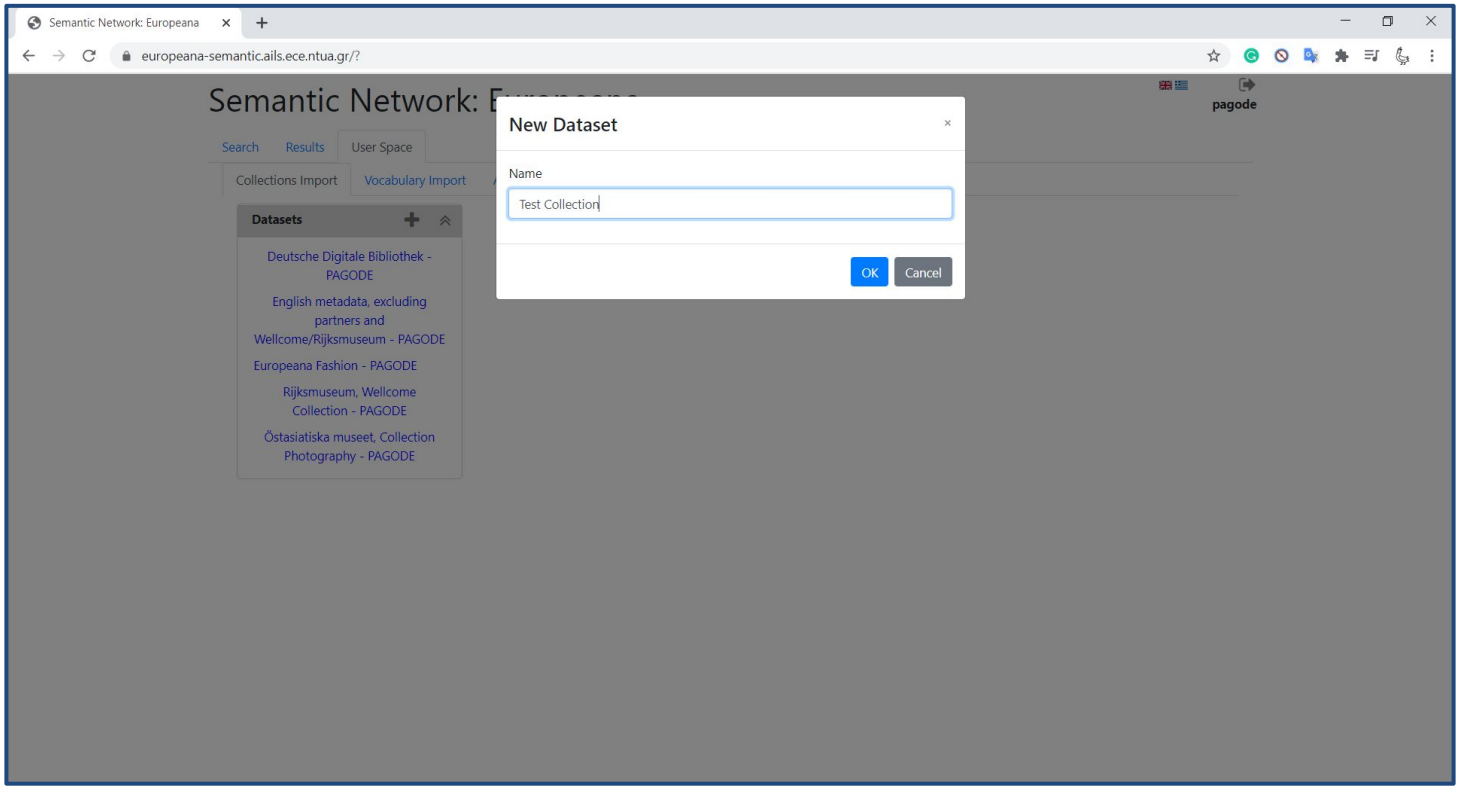
My Imported Collections



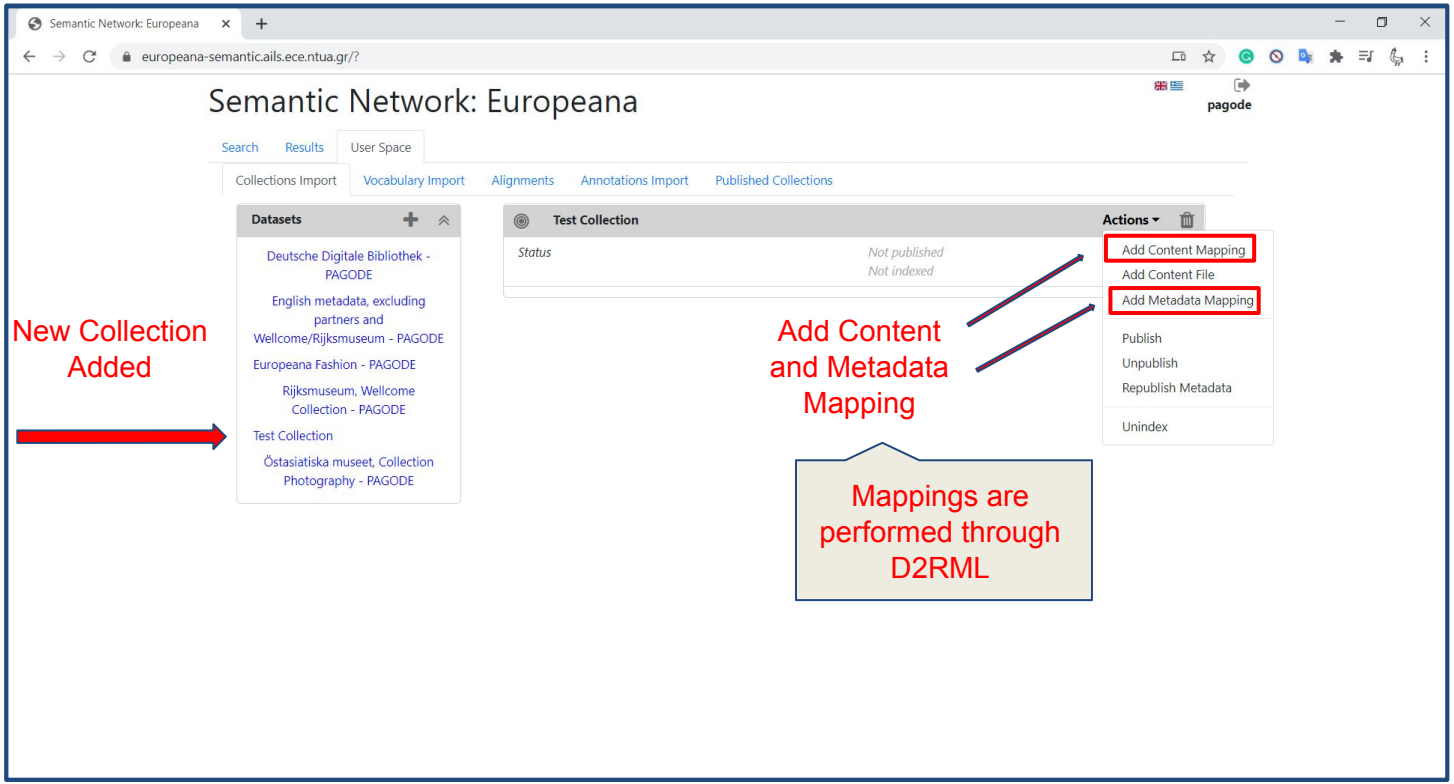
Add new Collection



Add new Collection



Import new Collection



D2RML

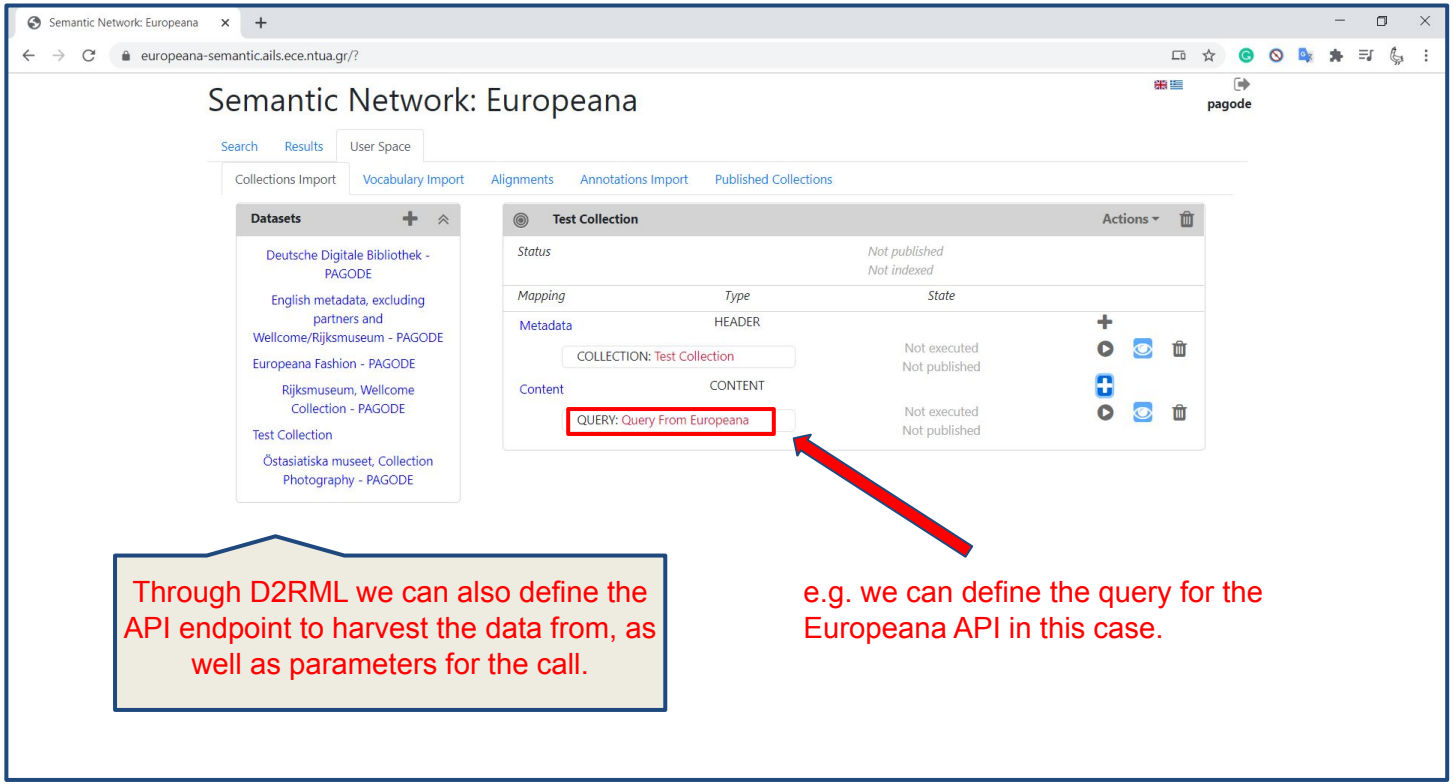
What is D2RML?

D2RML[1] is a language for transforming heterogeneous data (from various sources and of various formats) to RDF datasets.

Extension of the R2RML mapping language.

Created by NTUA.

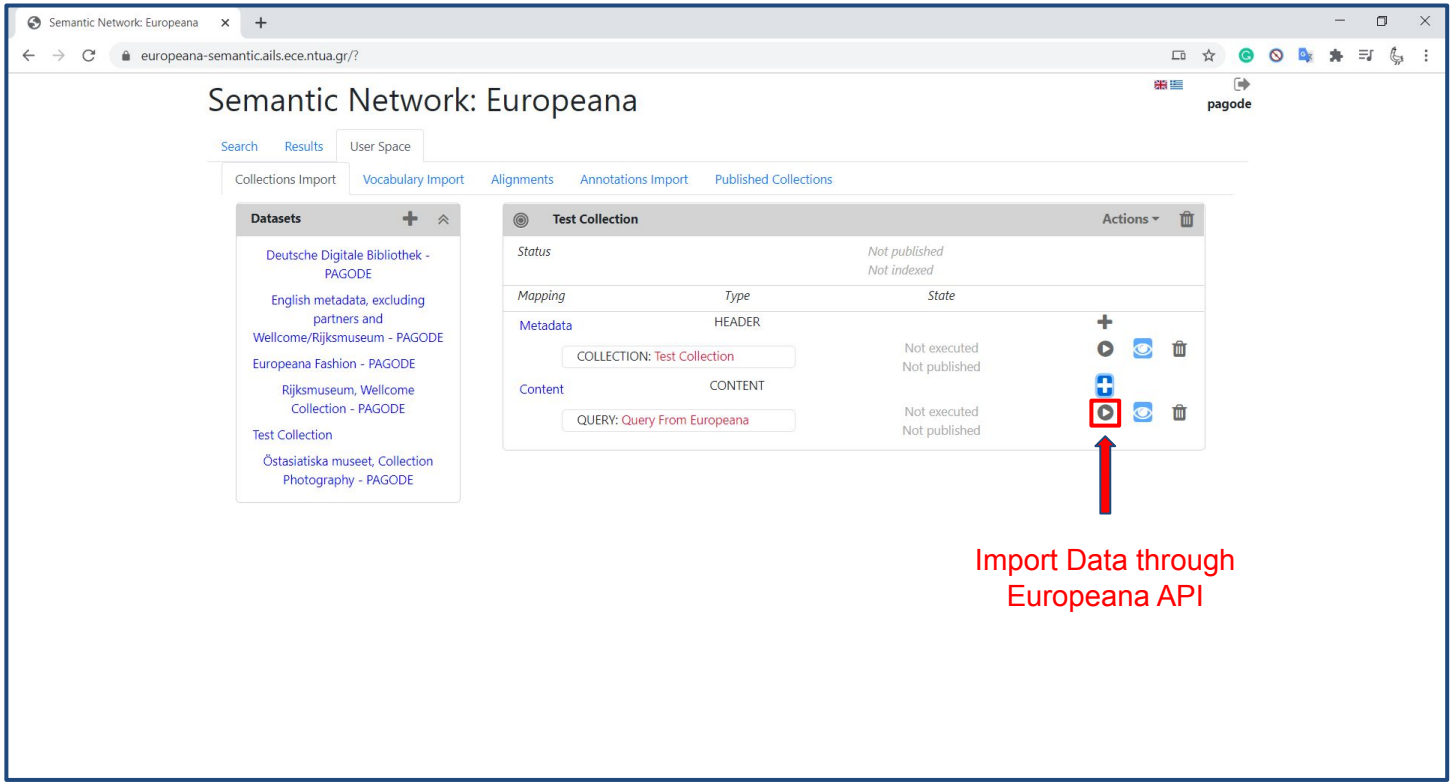
Import Data



Through D2RML we can also define the API endpoint to harvest the data from, as well as parameters for the call.

e.g. we can define the query for the Europeana API in this case.

Import Data

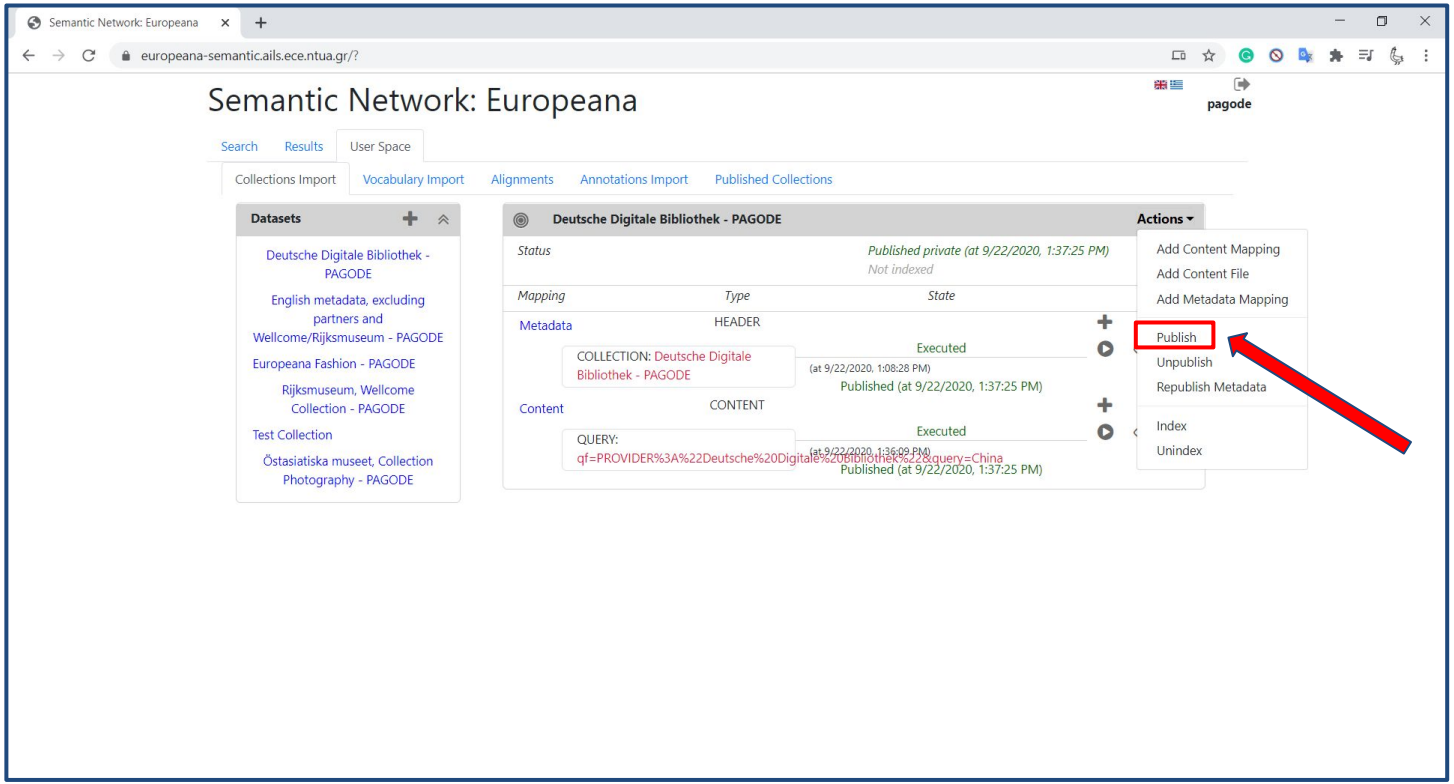


Import Data through Europeana API

made with europeana

DIGITAL CULTURE
www.digitalmeetsculture.net

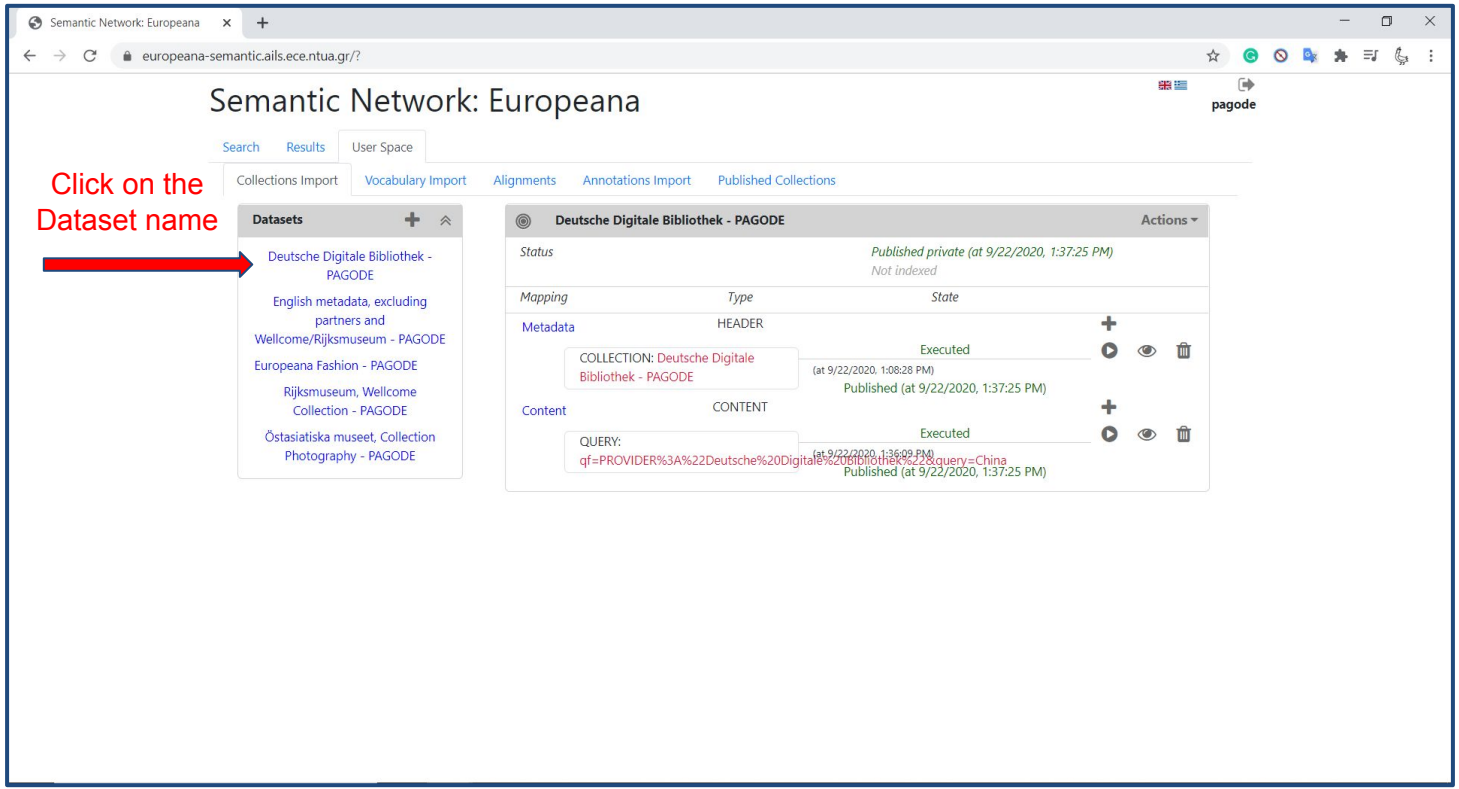
Publish Imported Collection to TripleStore



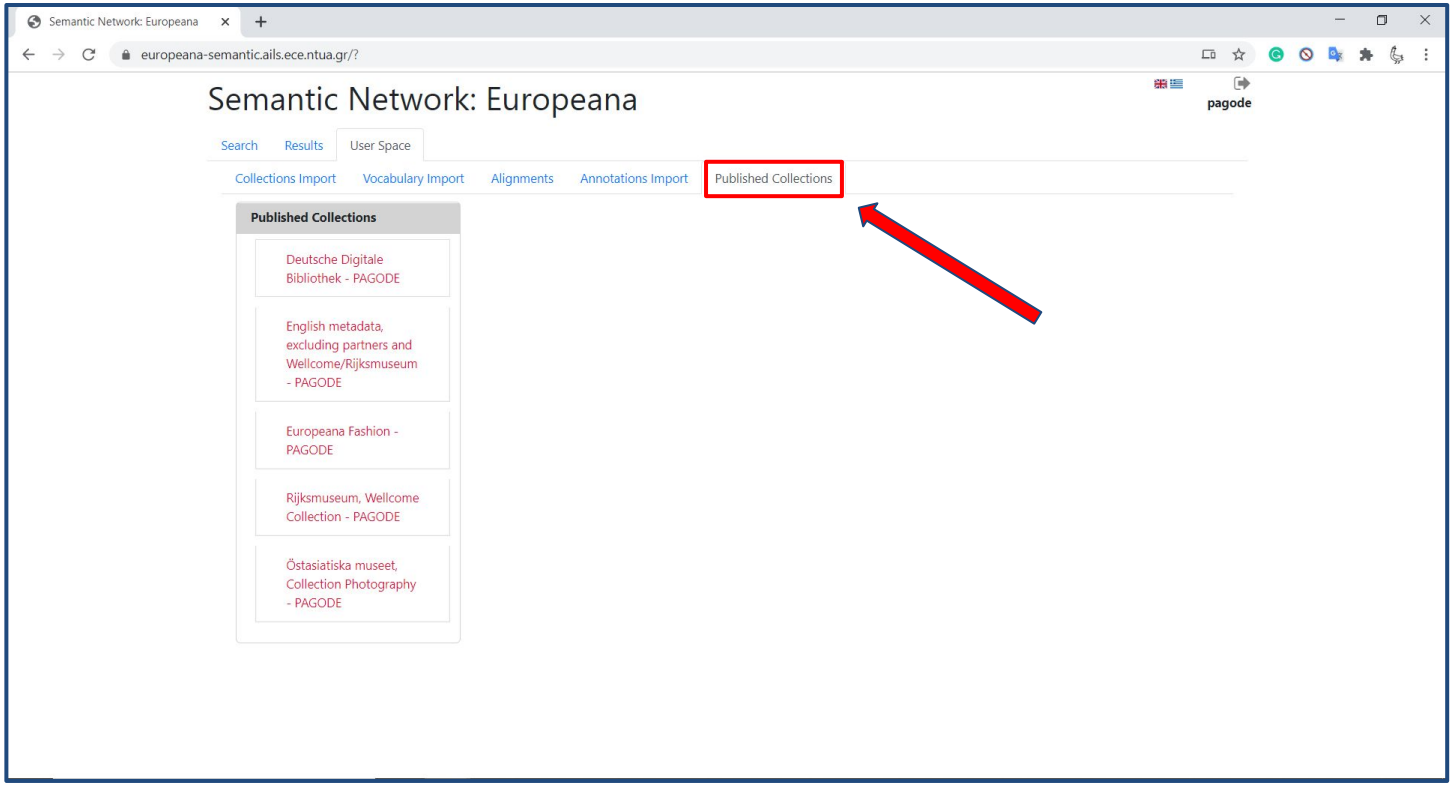
made with europeana

DIGITAL CULTURE
www.digitalmeetsculture.net

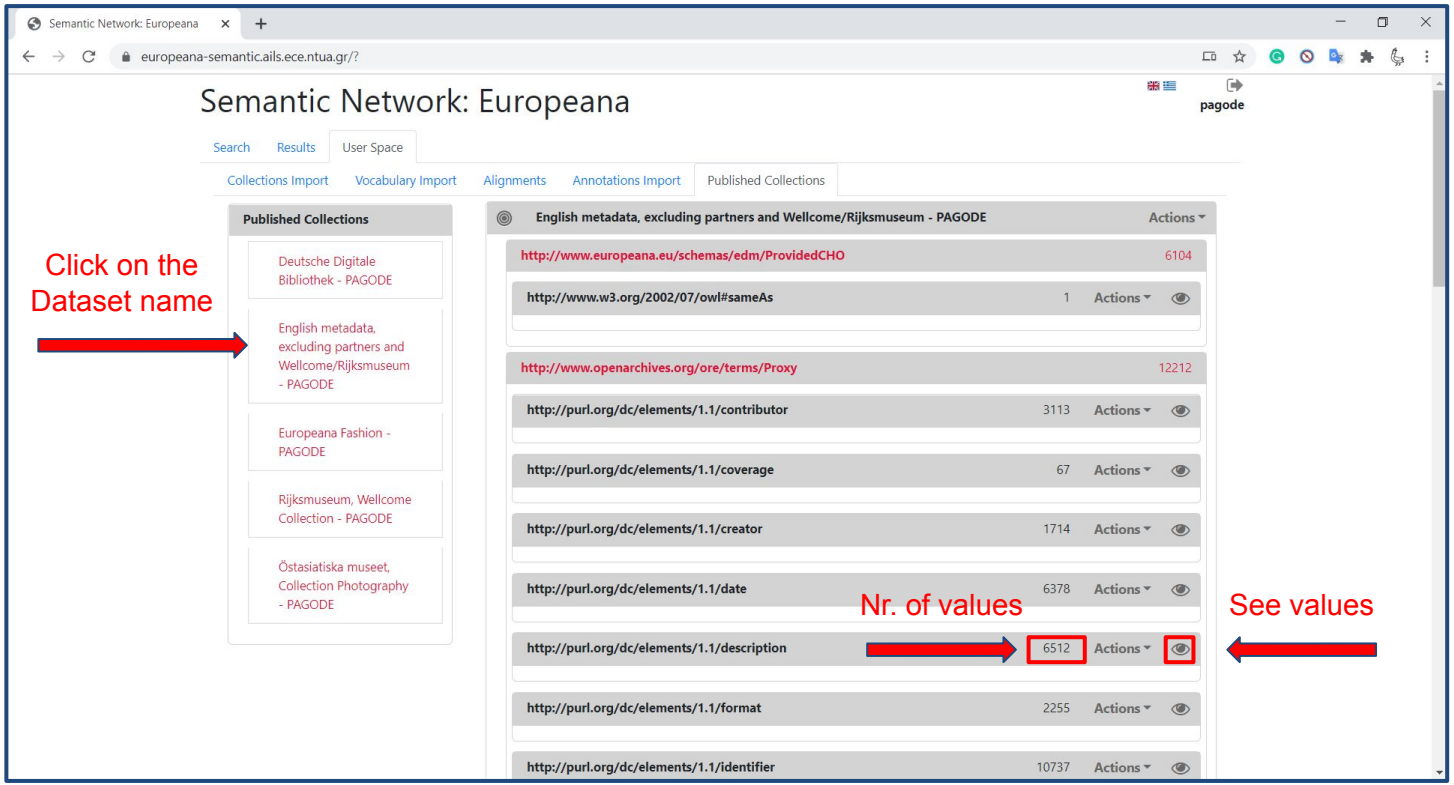
See Dataset/Collection Import Status



Published Collections



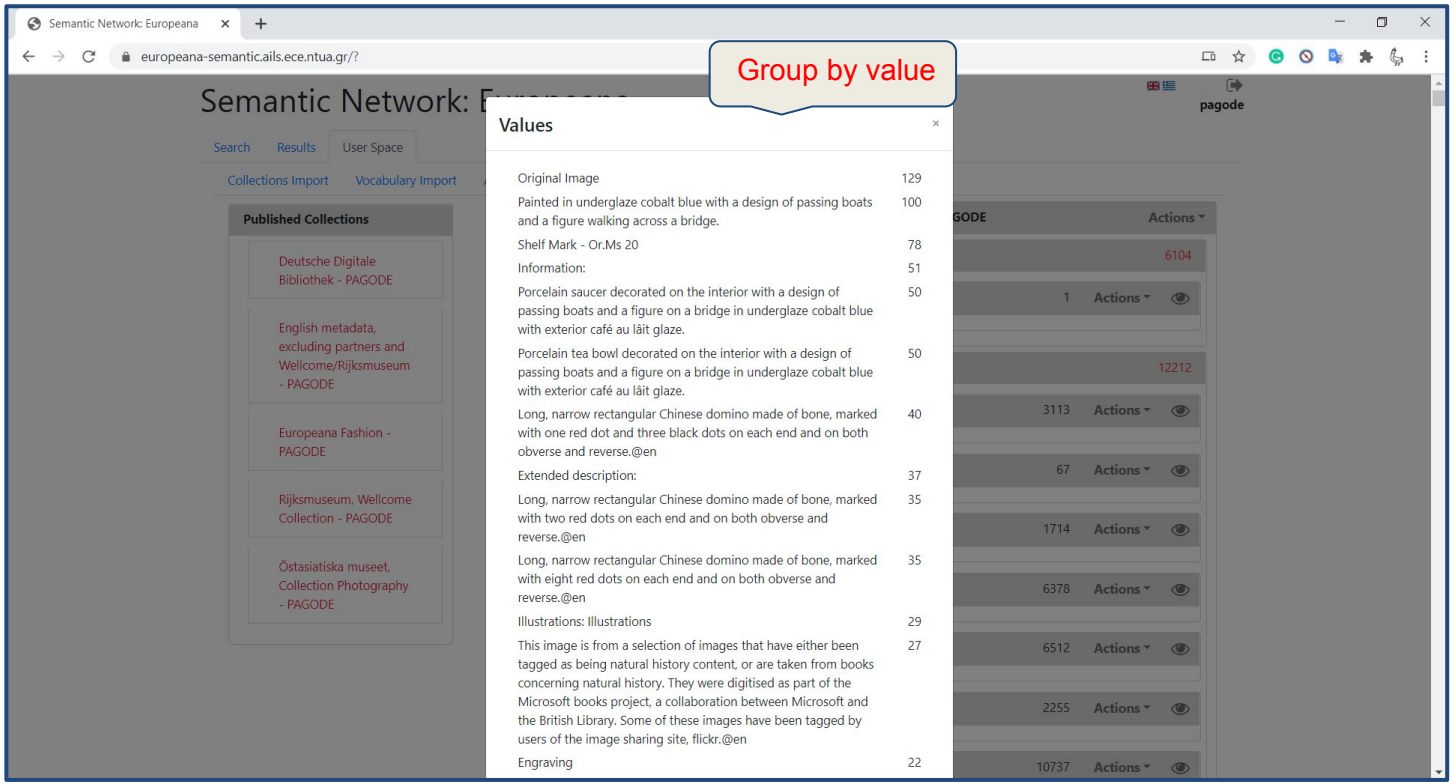
Properties of Published Collection



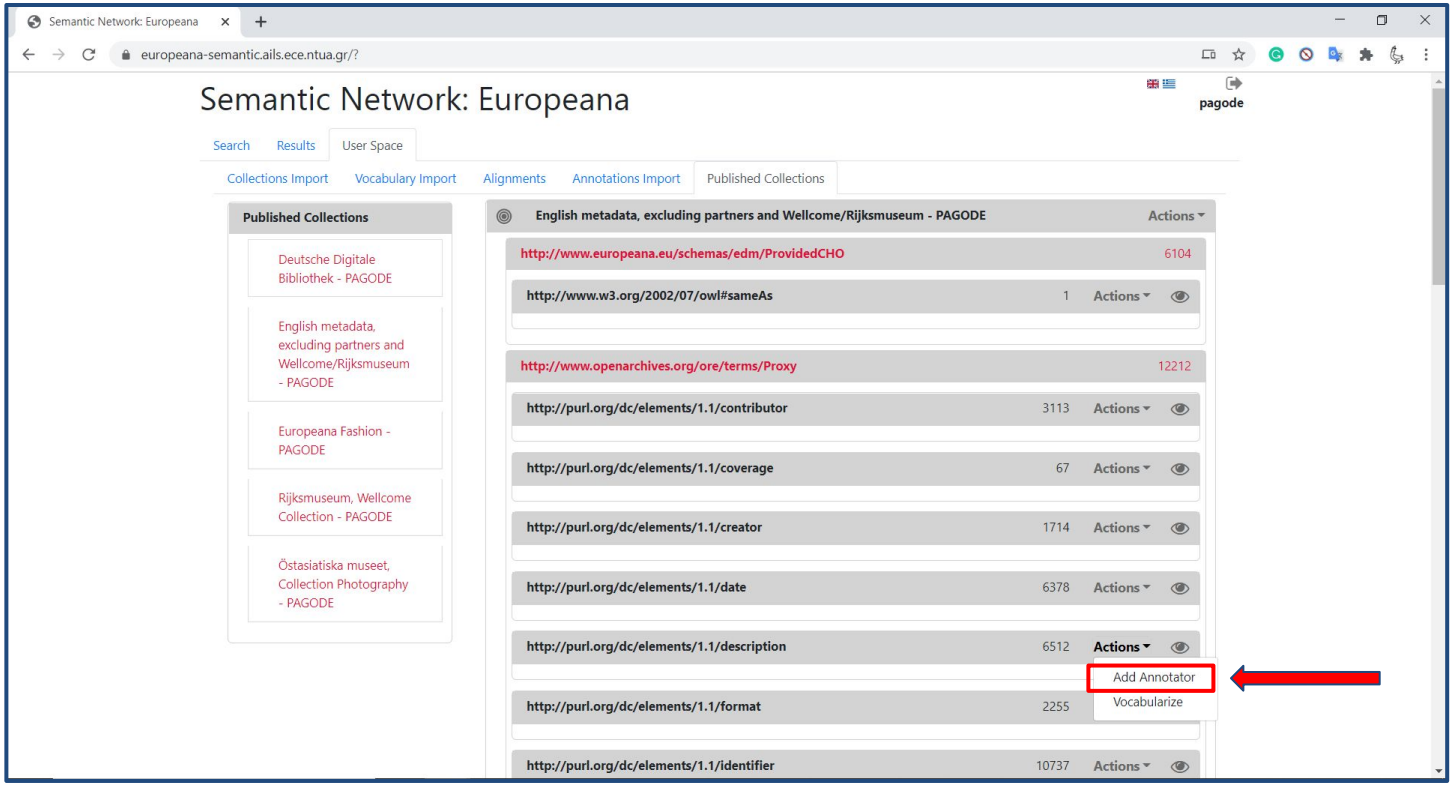
made with europeana

DIGITAL CULTURE
www.digitalmeetsculture.net

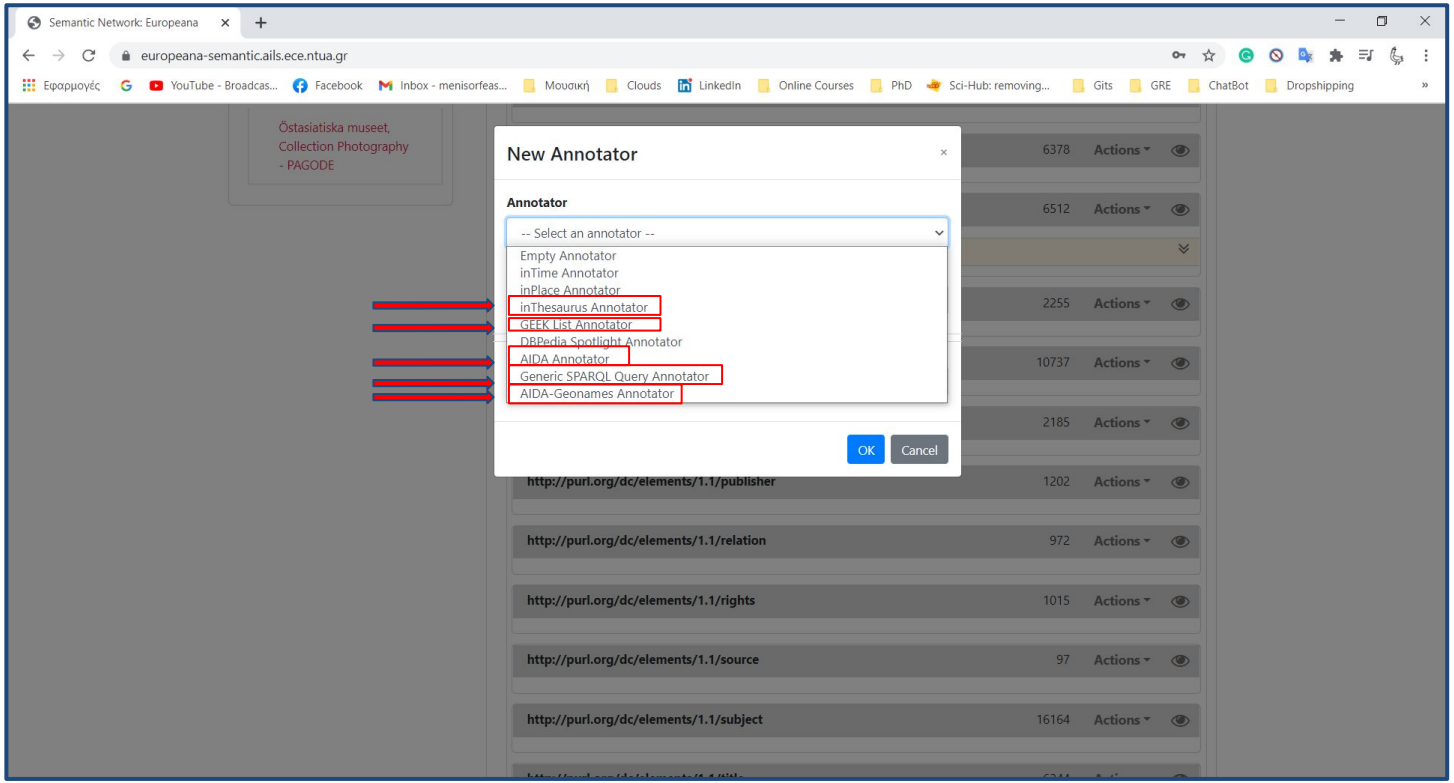
See field values



Add Annotator



Annotators



GEEK

What is GEEK?

GEEK[2] is a Named Entity Recognition and Disambiguation (NERD) system that extracts named entities in text and links them to a knowledge base using a graph-based method.

Created by NTUA.

AIDA

What is AIDA?

AIDA[3] is a framework and online tool for entity detection and disambiguation. Given a natural-language text or a Web table, AIDA maps mentions of ambiguous names onto canonical entities like people or places, registered in a knowledge base like DBpedia, Freebase, or YAGO.

Created by Max-Planck-Institut für Informatik.

Named Entities

What is a named entity?

In information extraction, a **named entity** is a real-world object, such as persons, locations, organizations, products, etc., that can be denoted with a proper name. It can be abstract or have a physical existence. Examples of named entities include Barack Obama, New York City, Volkswagen Golf, or anything else that can be named. [4]

NERD Example

"Leonardo painted Gioconda in the early 1500s, and now the painting is stored in the Louvre."

Leonardo needs to be linked to the polymath, and not the famous Hollywood actor. Gioconda needs to be linked to Mona Lisa, and not Ponchielli's opera. Louvre needs to be recognized as the French museum. The aim of a NERD tool is to eliminate such candidate entities that don't appear to fit with the context of the given text.

Result of a NERD tool:

Leonardo ⟶ https://www.wikidata.org/wiki/Q762 (Leonardo da Vinci)

Gioconda ⟶ https://www.wikidata.org/wiki/Q12418 (Mona Lisa)

Louvre ⟶ https://www.wikidata.org/wiki/Q1075988 (Louvre Palace)

SPARQL Annotator

What is SPARQL?

SPARQL is an RDF query language able to retrieve and manipulate data stored in RDF format. [5]

Wikidata provide a SPARQL endpoint, so through this annotator we can directly retrieve information from wikidata.

AIDA - Geonames

Place Annotator with Geonames URI.

2-Stage Annotator

In the future it will be implemented with GEEK as well.

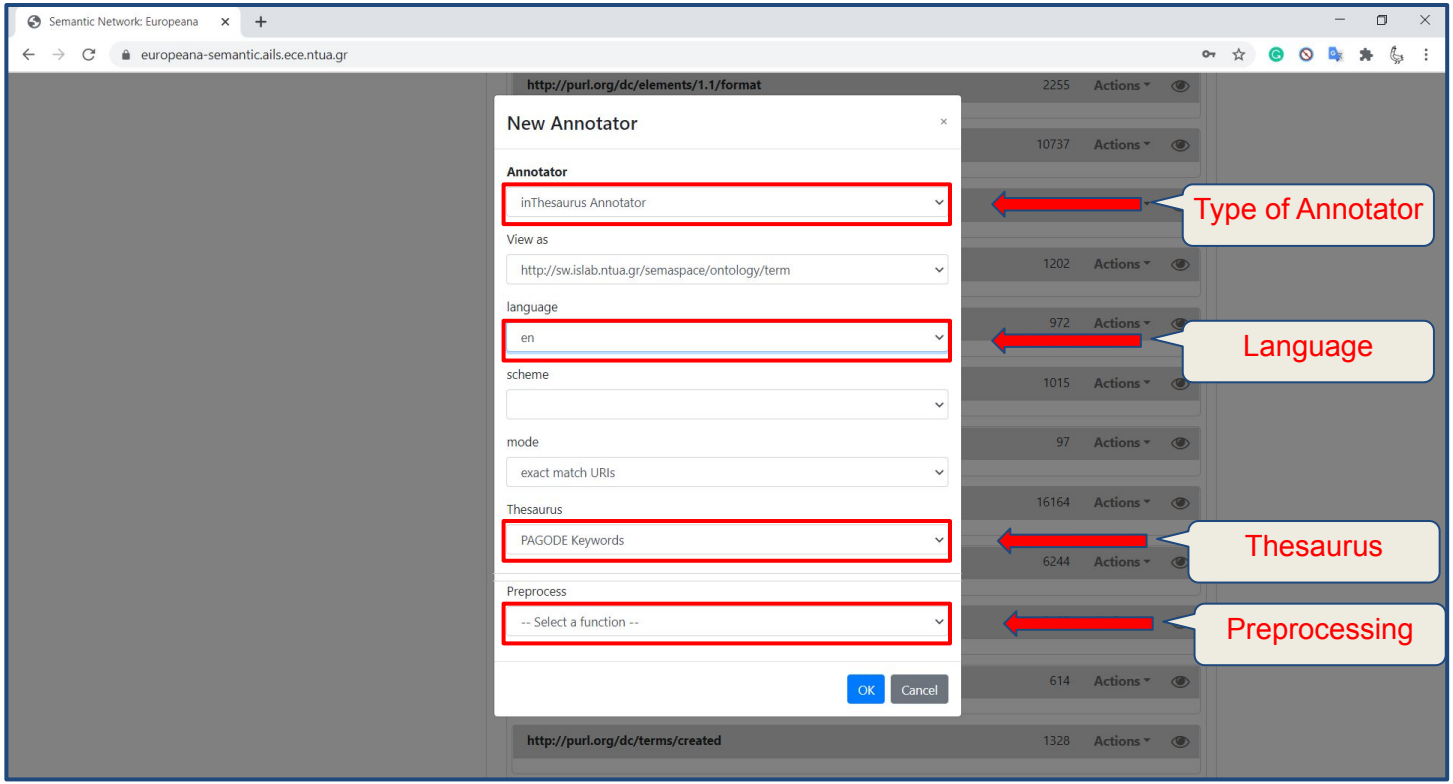
Thesaurus Annotator

Smart String Matching with thesaurus terms.

Main annotator tool for PAGODE.

Multilingual.

Annotator Specifications



made with europeana

DIGITAL CULTURE
www.digitalmeetsculture.net

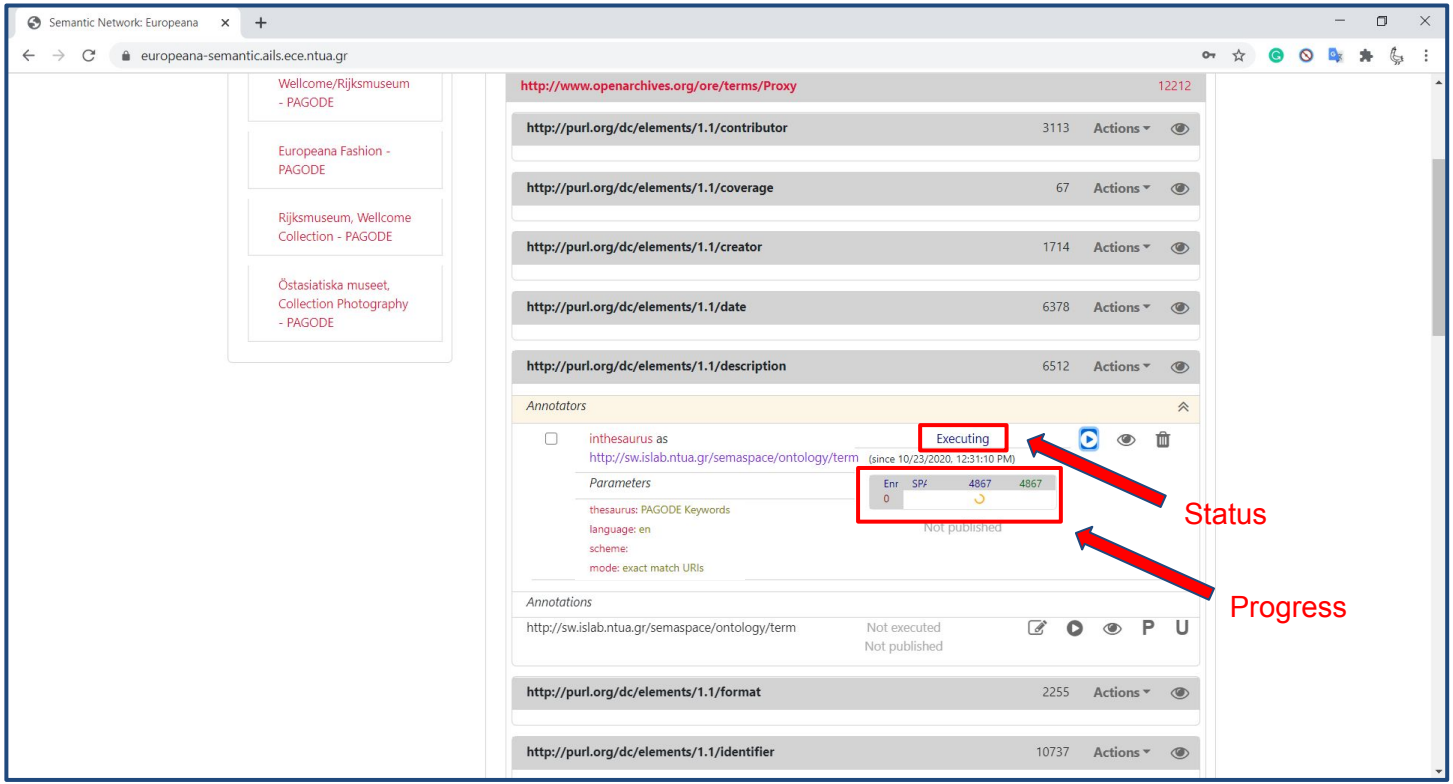
Preprocessing

Option to preprocess text before fed to the annotator.

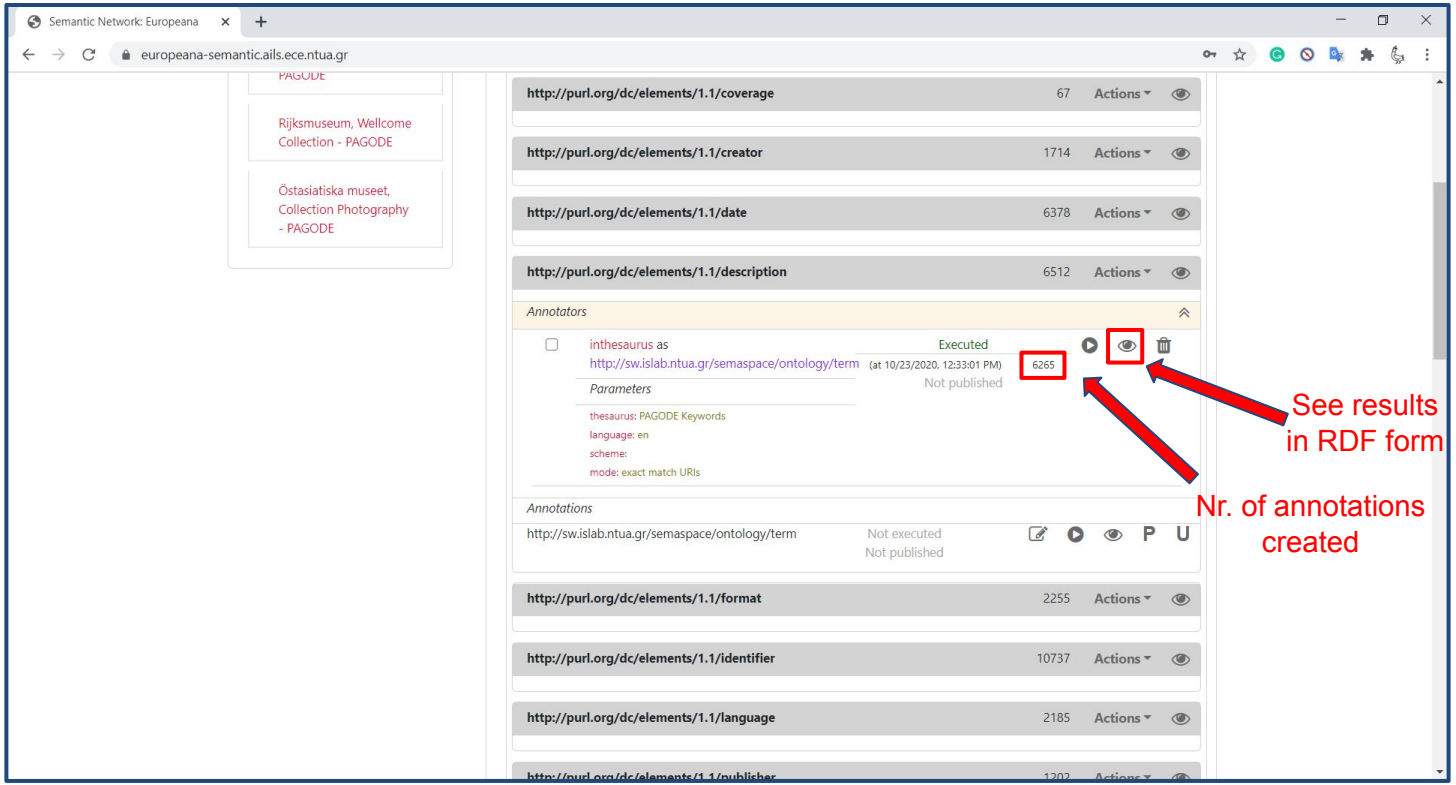
Use Regular Expressions (RegEx) to manipulate text prior to enrichment process.

Very useful feature in combination with the SPARQL annotator to query wikidata for persons and/or organizations.

Enrichment Progress



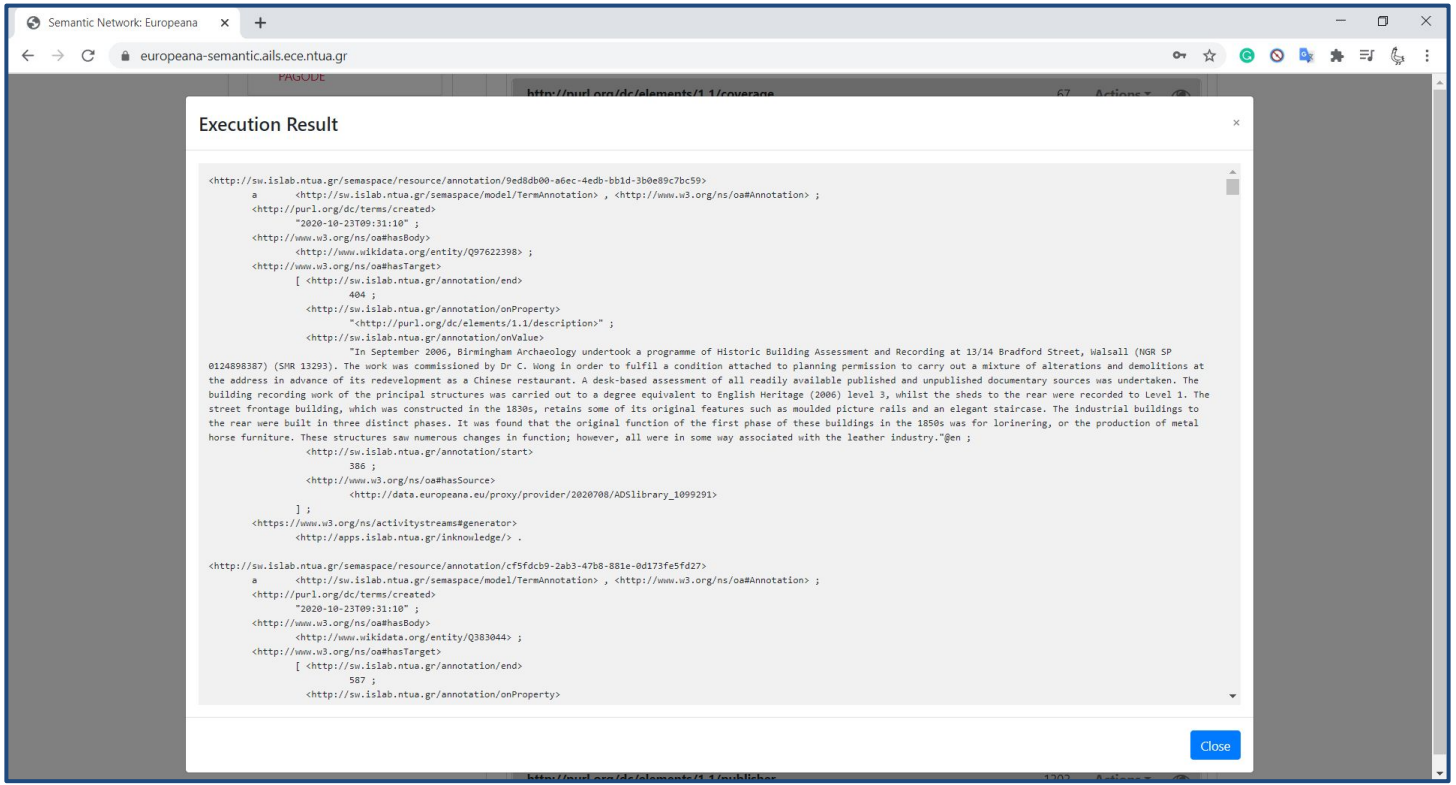
Enrichment Finished



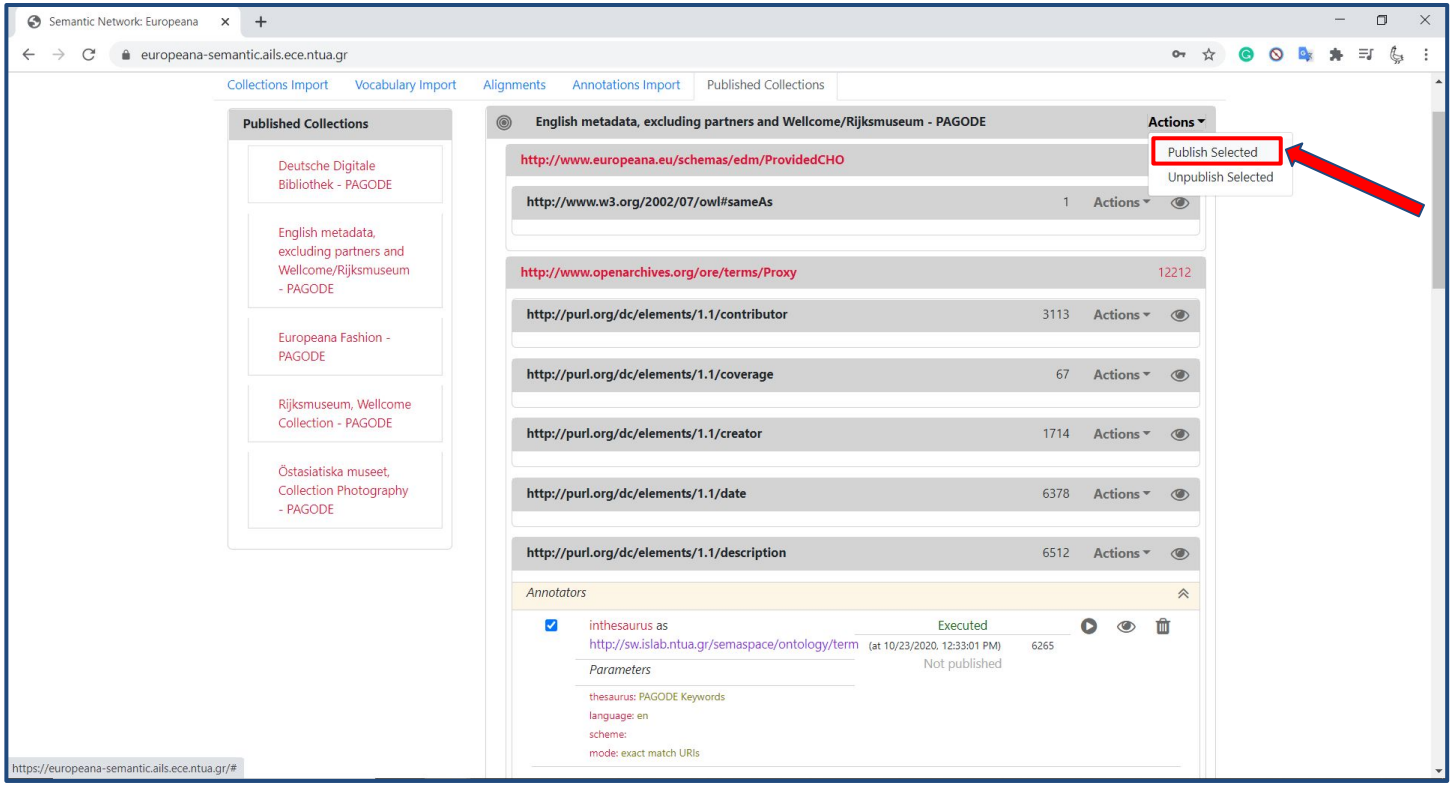
made with europeana

DIGITAL CULTURE
www.digitalmeetsculture.net

Raw RDF Enrichments



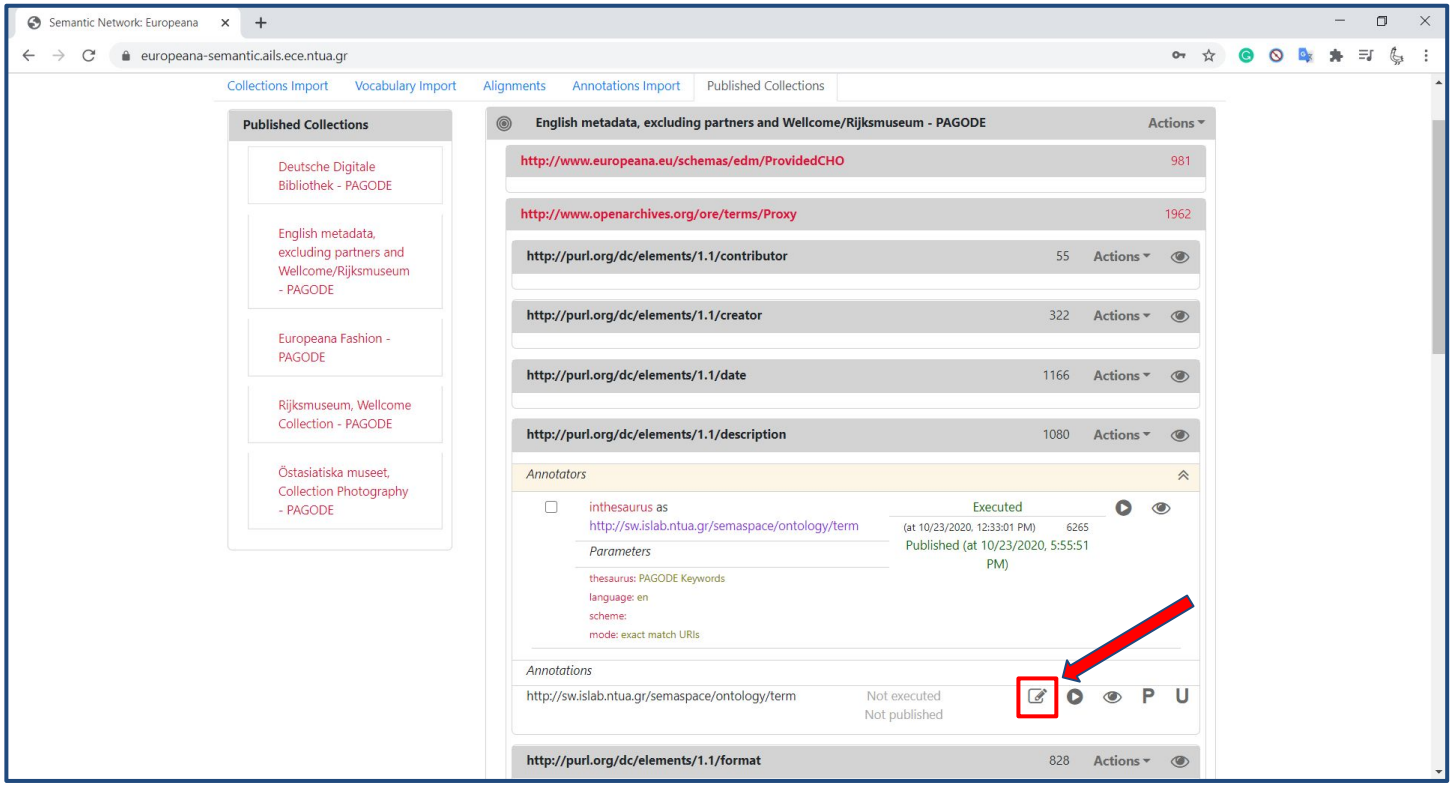
Publish Enrichments to TripleStore



made with europeana

DIGITAL CULTURE
www.digitalmeetsculture.net

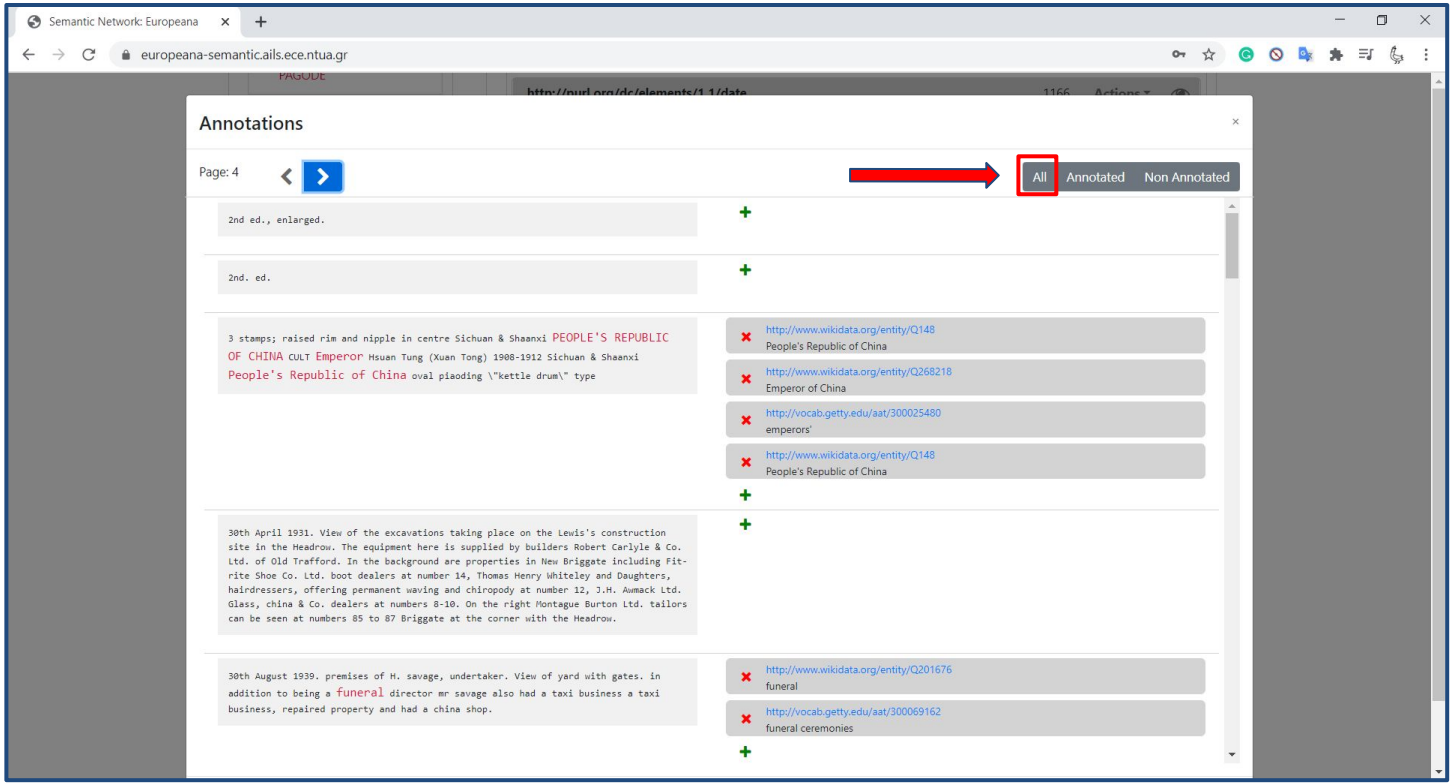
See and Edit Enrichments



made with europeana

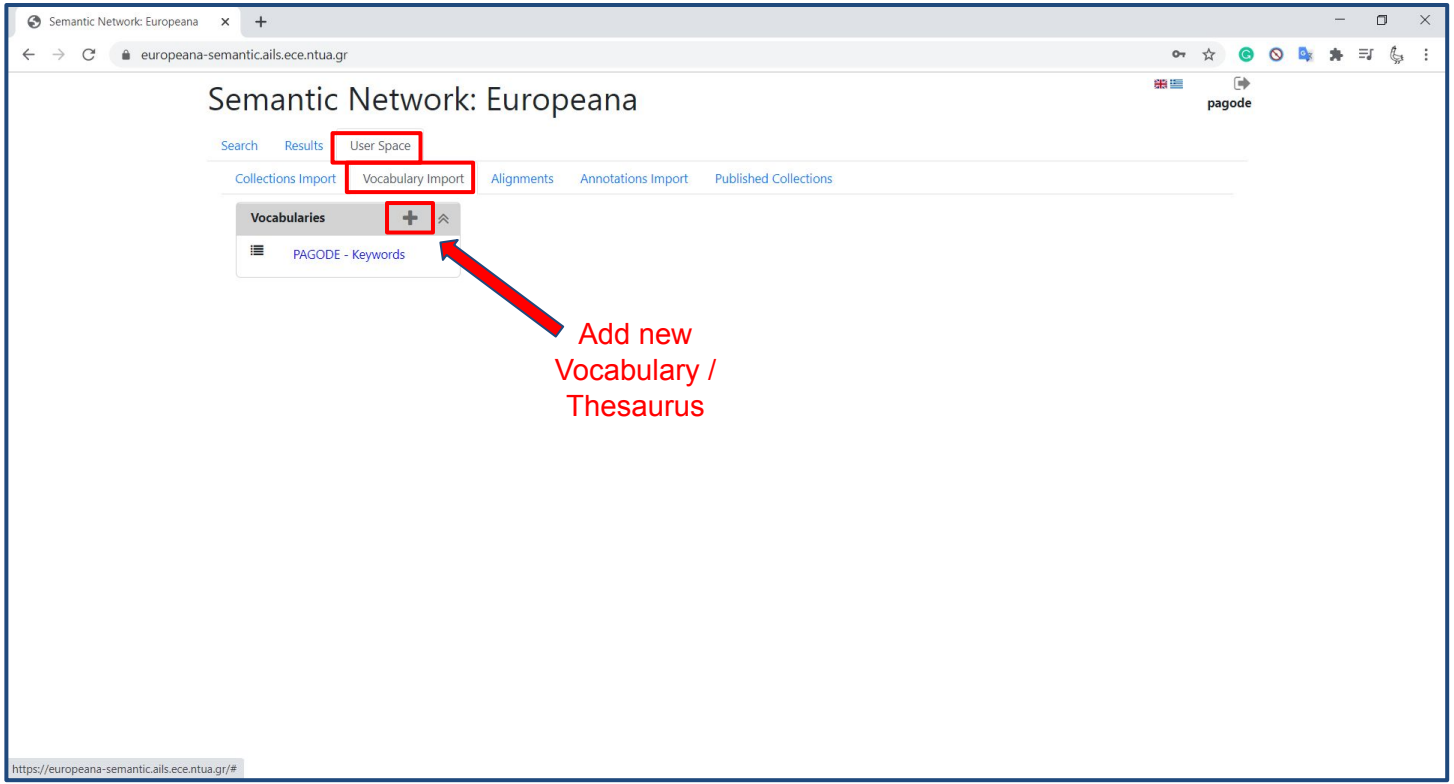
DIGITAL CULTURE
www.digitalmeetsculture.net

See and Edit Enrichments



Example of Enrichments

Import Vocabularies



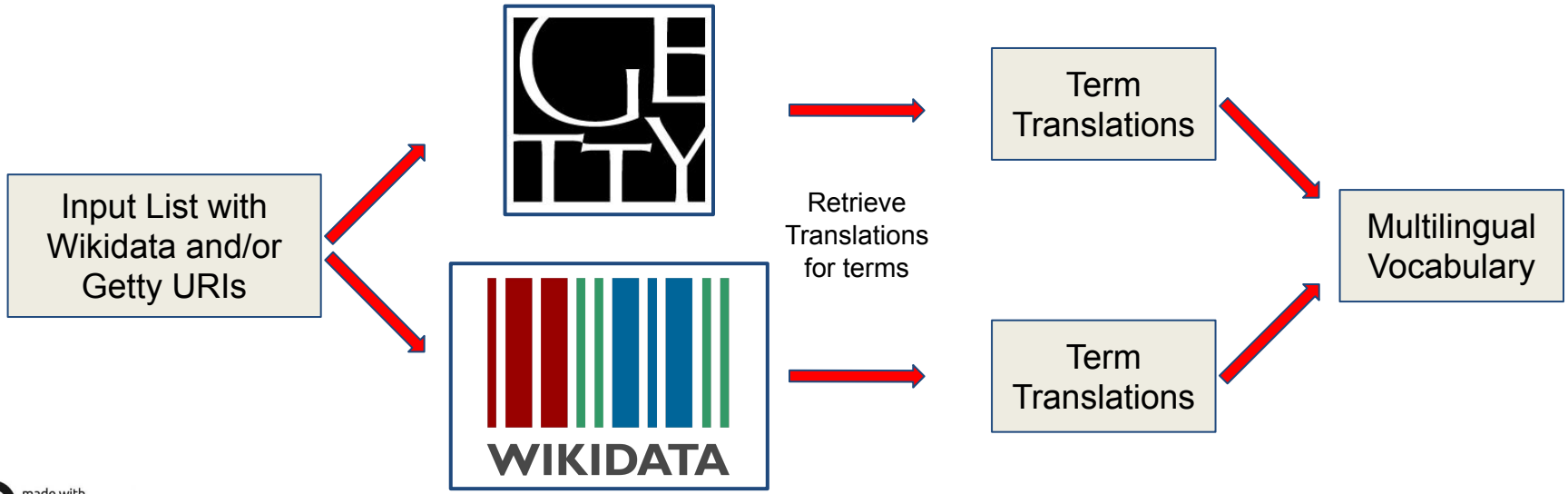
Import Existing Vocabulary
or
Create Custom Vocabulary

The user can import an existing vocabulary just by uploading the RDF file of the vocabulary.

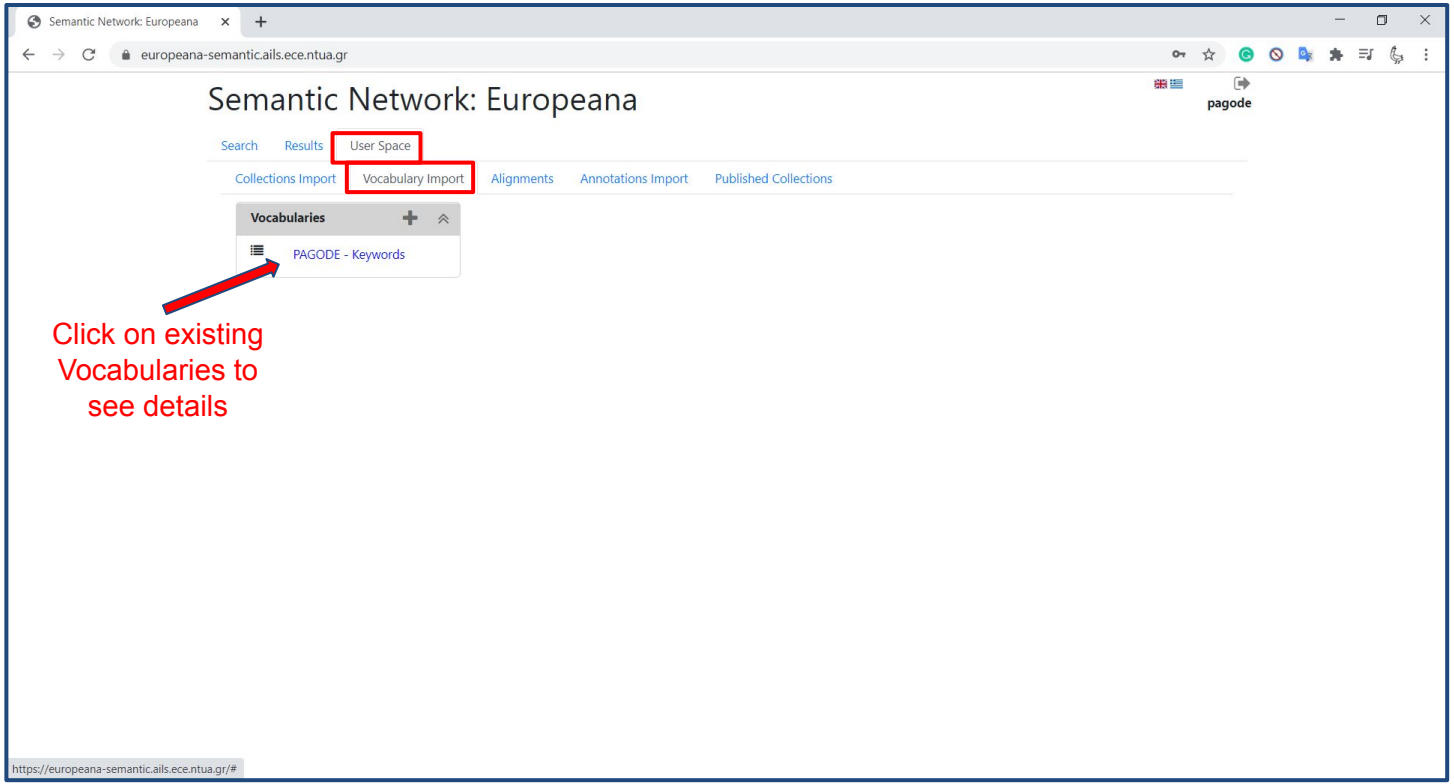
There is also the option to create a custom vocabulary, like we did for PAGODE. D2RML handles the mapping and transforms a list of terms to an RDF vocabulary / thesaurus.

Multilinguality for Custom Vocabulary

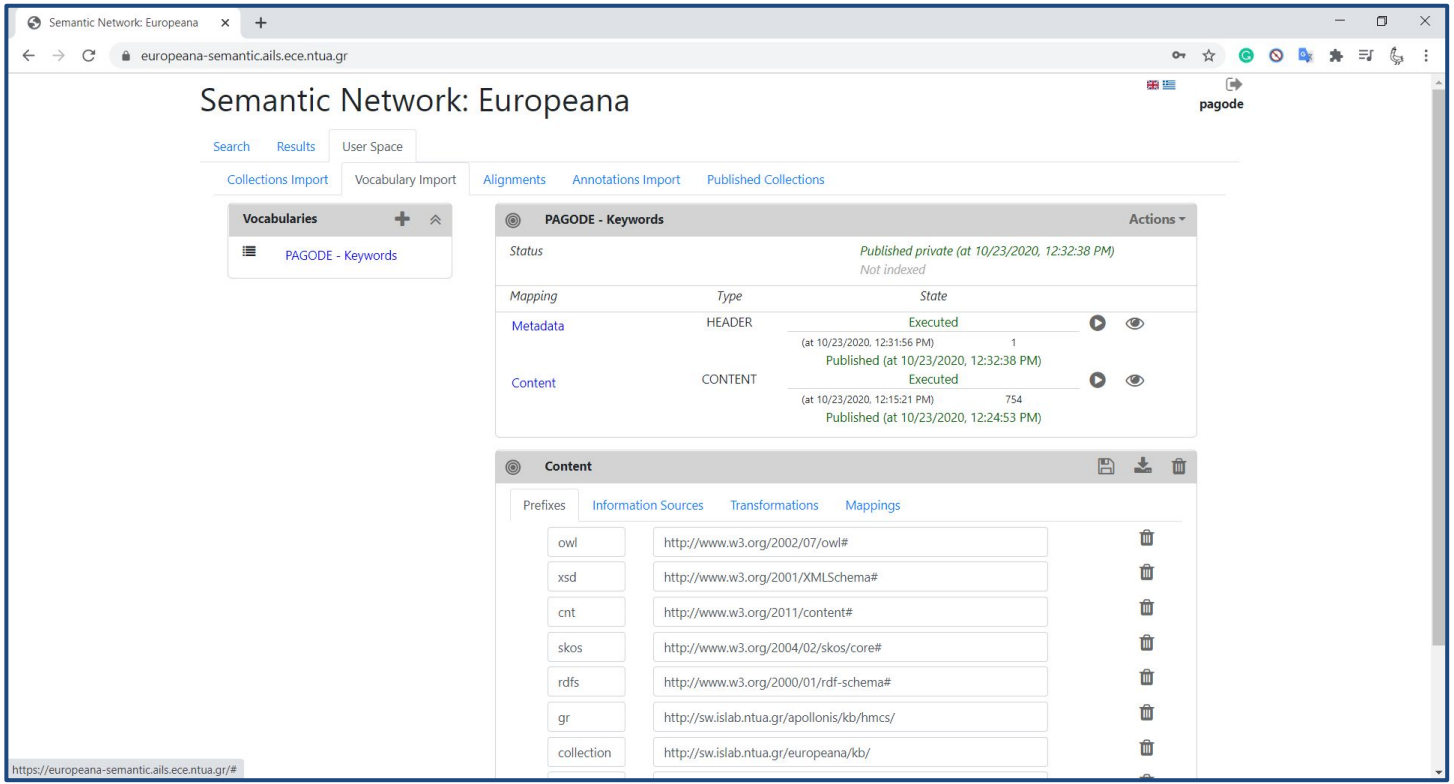
Functionality to create multilingual vocabulary from list of URIs.



Vocabularies



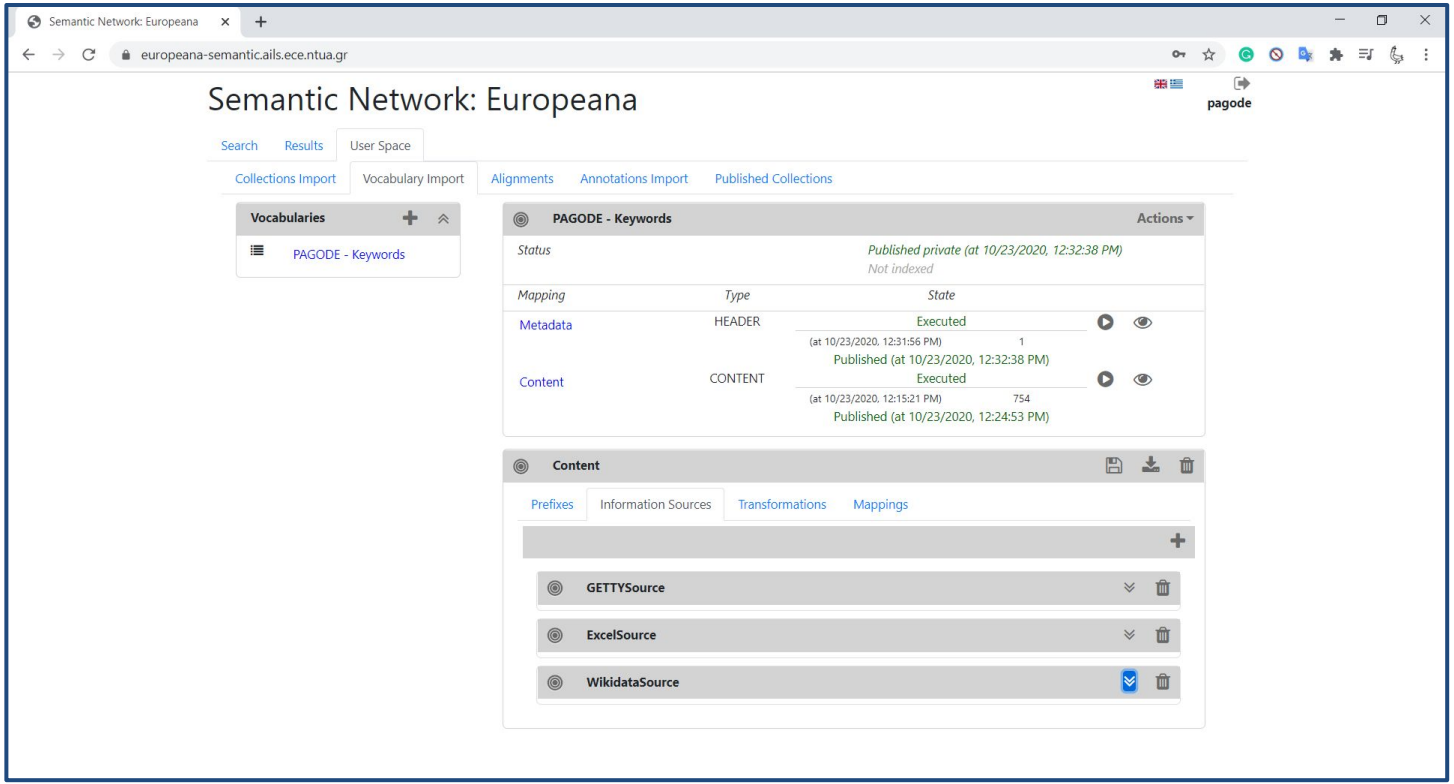
Vocabulary Properties



made with europeana

DIGITAL CULTURE
www.digitalmeetsculture.net

Vocabulary Properties



References

[1] Chortaras, A. and G. Stamou. “D2RML: Integrating Heterogeneous Data and Web Services into Custom RDF Graphs.” *LDOW@WWW* (2018).

[2] Alexios Mandalios, Konstantinos Tzamaloukas, Alexandros Chortaras, Giorgos Stamou. “GEEK: Incremental Graph-based Entity Disambiguation.” LDOW@WWW (2018)

[3] Yosef, Mohamed & Hoffart, Johannes & Bordino, Ilaria & Spaniol, Marc & Weikum, Gerhard. (2011). AIDA: An Online Tool for Accurate Disambiguation of Named Entities in Text and Tables. PVLDB. 4. 1450-1453. 10.14778/3402755.3402793.

[4] “Named entity”, *Wikipedia, The Free Encyclopedia*,12 October 2020, 19:31 (UTC). <https://en.wikipedia.org/wiki/Named_entity#:~:text=In%20information%20extraction%2C%20a%20named,or%20have%20a%20physical%20existence.>

[5] “SPARQL”, *Wikipedia, The Free Encyclopedia*,23 October 2020, 11:56 (UTC). <https://en.wikipedia.org/wiki/SPARQL#:~:text=SPARQL%20(pronounced%20%22sparkle%22%20%2F,Description%20Framework%20(RDF)%20format.>

PAGODE

Thank you

made with europeana

DIGITALMEETSCULTURE
www.digitalmeetsculture.net