PhD Hub
Semantic annotation and linking of scientific Classifications

Authors

Charalampos Bratsas, Sotirios Karampatakis, Panagiotis-Marios Filippidis, Lazaros Ioannidis[1], Jose de las Heras[2]

Executive Summary

The aim of this deliverable is the construction of a scientific knowledge graph via semantic annotation and linking of PhD research fields. First, a survey of scientific classifications has been conducted, in order to identify the ones that are the most common and complete for the knowledge graph to build upon. Classifications from different scientific domains have been used to semantically annotate their thematic topics and fields. The various scientific fields are connected based on their similarity, enlightening and creating, in this way, cross-domain research fields. A core scientific graph containing main fields of science with their relations to domain specific vocabularies and classifications has been created. This knowledge graph can be used to retrieve specific scientific fields in a related, broader or narrower, research area. Finally, use cases leveraging the semantic features of the graph are presented, indicating its usefulness for all the potential users of the PhD Hub.

[1] Aristotle University of Thessaloniki

[2] Advantic Sistemas y Servicios

Co-funded by the
Erasmus+ Programme
of the European Union

Table Of Contents

With the support of Erasmus+

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

With the support of Erasmus+

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

Glossary and Abbreviations

WP	Work Package
SKOS	Simple Knowledge Organization System
RDF	Resource Description Framework
SW	Semantic Web
TM	Text Mining
IE	Information Extraction
NER	Named Entity Recognition
NLP	Natural Language Processing
ETL	Extract Transform Load
API	Application Programming Interface
KG	Knowledge Graph
LP-ET	LinkedPipesETL
REGEX	Regular Expression
SPARQL	SPARQL Protocol and RDF Query Language

1 Introduction

There are many scientific classifications which are oriented, or limited to a particular research area, such as Mathematics, Physics, Computer Science and others. These classifications serve as an index of all domain-specific fields of the respective research area, also allowing the hierarchical structuring of the corresponding concepts.

These classifications may include over a thousand of relevant terms. Inevitably, there are some overlaps between classifications of relative scientific areas, which means that a specific research field could be part of the hierarchy of two or more distinct scientific classifications at the same time.

It is evident, thus, that without concept schemes that combine the knowledge from different scientific classifications, the usage of these classifications are limited only to their corresponding audience, limiting, as well, cross-domain research activities, originating from similar domain fields.

The aim of this deliverable is to combine the broad scope of generic classifications of research areas with the specialized knowledge and domain fields of specific scientific classifications, in order to build a unified scientific knowledge graph including all the research fields of the respective scientific areas in a common hierarchy.

This knowledge graph will help the PhD Hub users extend their research range to more scientific domains, as they will be somehow connected, via their common fields. Graph analysis features result in a scientific area clustering and reduce the distance between seemingly different scientific fields and areas.

For example, in most cases, students searching the most suitable PhD offer in PhD research websites, are limited to search only for specific scientific areas. This may require considerable time and effort, if their research interests are diverse. When the relevant data are connected, knowledge retrieval techniques can automatically result all the corresponding information in a user's search, broadening the search scope to additional scientific fields.

This result benefits the development of the online academia hubs, since it intends to cluster the profile of platform academia users in research areas, as well as their call for proposals. This is a crucial endeavour since it will aim at providing targeted information to the users on peers using the platform and their call for cooperation without limiting the potential of information too much by providing too restrictive information.

Finally, an academia platform using this knowledge graph can be clustered into the various scientific areas clusters to facilitate interdisciplinary research activities and knowledge transfer, for any users group of the PhD Hub, such as students, professors, universities and businesses, among others.

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

1.1 Task Objective

Classifications from different scientific domains will be used to semantically annotate their thematic topics and fields, in order to connect these fields based on their similarity, enlightening and creating, in this way, cross-domain research fields. A core scientific graph containing main fields of science with their relations to domain specific vocabularies and classifications has been created and can be used to retrieve specific scientific fields in a related, broader or narrower, research area.

1.2 Task impact

This result will be helpful for the online European PhD Hub, since it intends to cluster the profile of platform users in research areas, as well as their call for proposals. This is a crucial endeavour, since it will aim at providing targeted information to the users on peers using the platform and their call for cooperation, without limiting the potential of information too much by providing too restrictive information. To this end, a conceptual linking of the various thematic topics that are currently classified only under certain scientific fields is needed. Thematic topics and areas can be retrieved by scientific domains classifications such as the Medical Subject Headings (MESH), the Mathematics Subject Classification (MSC2010), the ACM Computing Classification System (CCS) and libraries classifications, such as the Library of Congress Subject Headings(LCSH).

Graph analysis features will result in a scientific area clustering and will also reduce the distance between seemingly different scientific fields and areas. Finally, the PhD Pages of the PhD Hub will be clustered into the various scientific areas clusters, to facilitate interdisciplinary research activities and knowledge transfer.

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

2 Theoretical Background

The construction of the knowledge graph leads to the need for the conceptual linking of the various thematic topics that are currently classified only under certain scientific fields. Thematic topics and areas can be retrieved by classifications from various scientific domains such as Mathematics, Physics, Medicine, Computer Science and other, while libraries classifications could be valuable as well.

The connection between classifications from different scientific fields requires a common representation format, in order to facilitate the linking procedure, as well as the construction of the core knowledge graph. A list of scientific concepts and areas can thus be transformed into a richer representation and later be linked to other, similar classifications.

The representation selected for this deliverable is the Simple Knowledge Organization System (SKOS). SKOS is an RDF vocabulary, W3C recommended, designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, and generally any other type of structured controlled vocabulary. SKOS is part of the Semantic Web family of standards built upon RDF and RDFS, and its main objective is to enable easy publication and use of such vocabularies as linked data. Because SKOS is based on the Resource Description Framework (RDF), these representations are machine-readable and can be exchanged between software applications and published on the World Wide Web.

Most of the selected classifications were already published in SKOS format. Otherwise, a SKOS transformation task had to take place in order to attain the proper format for each classification. Additional modification and transformation tasks were also required where the RDF representations of the classifications were not valid or contained errors.

2.1 Methodology

2.1.1 Search classifications of scientific fields

We conducted a survey of scientific classifications including potential PHD research fields. Priority was granted to scientific classifications that can be considered as complete as possible for their scope, generic or specific, containing a plethora of research areas. Classifications also had to come from recognizable sources as scientific organizations and competent authorities. Additionally, while the survey may include classifications with restricted access, there was clear preference to open classifications that could be later used for the entities linking and knowledge graph creation tasks.

2.1.2 Retrieve and store metadata and classifications in repository

The next step was to retrieve the identified classifications and store them, along with additional information and metadata in a Git repository. The RDF format of the classifications, if any, was selected from all available formats in the initial sources.

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

2.1.3 Identify classifications to be used

Only a part of the retrieved classifications was finally selected to participate in the linking procedure, taking into account their specific attributes, the information wealth, their format and structure, as well as their possible overlap with other classifications. This was also the case for selecting the backbone classification, where the various classifications of specific scientific fields would be linked. The backbone classification had to be a generic list including terms of many scientific areas. To facilitate the linking process, the generic classification should also be either a rather small list of scientific terms, or a large hierarchical list with well-separated terms in broader levels. In this way, the specific scientific classifications could connect more easily with the generic classification in the broader level.

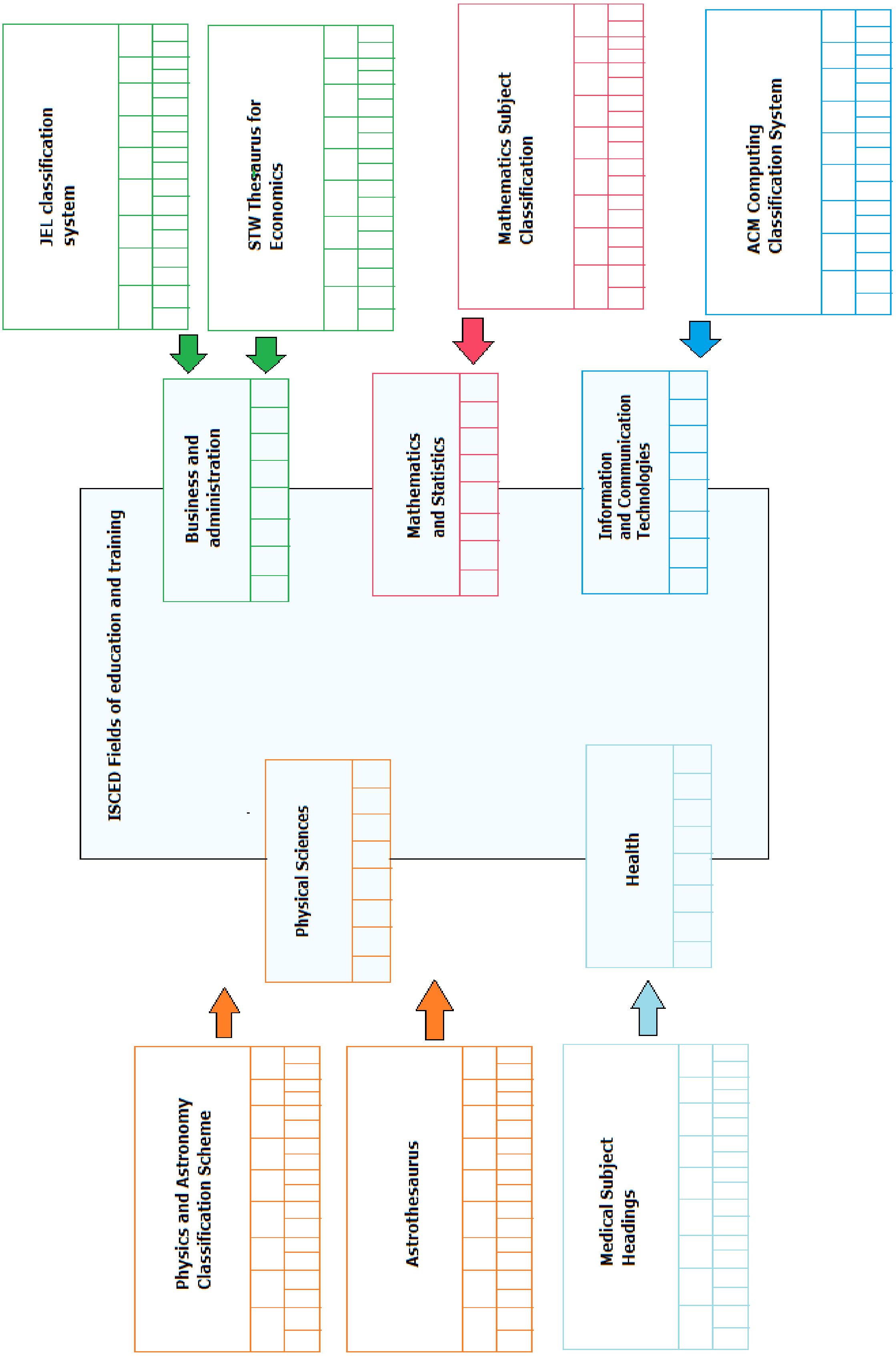
2.1.4 Semantic Annotation-SKOSification

Semantic Annotation tasks were performed for the selected classifications that were not already in the appropriate SKOS format. The SKOS-ifying task was implemented with the LinkedPipes ETL tool, requiring the construction of a components pipeline for the conversion of the classification into RDF format.

2.1.5 Create-Import linksets

Having all the selected classifications into the appropriate SKOS format, semi-automated linking processes between the core and the specific classifications were performed, using the Alignment tool. The core knowledge graph ends up containing terms from various specific classifications.

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220


JEL classification system
STW Thesaurus for Economics
Mathematics Subject Classification
ACM Computing Classification System
Business and administration
Mathematics and Statistics
Information and Communication Technologies
ISCED Fields of education and training
Physical Sciences
Health
Physics and Astronomy Classification Scheme
Astrothesaurus
Medical Subject Headings

With the support of Erasmus+

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

2.1.6 Results

Finally, based on the created linksets between the SKOS-ified classifications, further conclusions were drawn as far the need for experts contribution in the linking process is concerned.

Examples of the usefulness of the task described in this deliverable are presented through different use cases where the resulting knowledge is exploited for performing operations through PhD Hub that could not be completely accomplished otherwise.

2.2 Tools

In order to assist and curate the results of the above tasks, we needed to deploy a variety of services in PaaS environment. Docker is one of the leading software containerization platforms. “Docker containers wrap a piece of software in a complete filesystem that contains everything needed to run: code, runtime, system tools, system libraries – anything that can be installed on a server. This guarantees that the software will always run the same, regardless of its environment”[3]. This technology supports the PhD Hub Knowledge Repository in the integration of different modules and components.

Multiple applications have to be combined and work together. Therefore, there is the need for common interfaces between the applications. Furthermore all those different applications have to be installed on a server and to be connected to each other. The Docker-technology offers a way for running multiple applications in separate environments on a single server. Additionally Docker-Compose offers a way to orchestrate multiple applications together and to make them communicate to each other.

The prototype is based on the Docker-technology: https://www.docker.com Each application is contained within a Docker-Container. Docker containers are built from so-called Docker-Images which are either downloaded directly from DockerHub https://hub.docker.com or build from a Dockerfile during the deployment. Mostly only a single application of the PhD Hub Knowledge Repository Stack is contained within a Docker container.

The whole stack is currently developed and maintained in a Git repository. In order to deploy the full stack, the minimum software requirements are

- Linux kernel 3.10
- latest Docker CE version
- latest Docker-compose version

For more detailed installation instruction please consider this documentation and the instructions on the Git repository.

[3] https://www.docker.com

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

The server is aparted of several services, based on the requirements of the task.

Alignment: a collaborative, system-aided, user-driven ontology/vocabulary matching application. It is used to manually create linksets where automation fails and to validate linksets, by engaging domain experts.

LinkedPipesETL: LinkedPipes ETL is an RDF-based, lightweight ETL tool. There are cases where the classifications does not exist in SKOS representation so they have to be described using the SKOS vocabulary. Linkedpipes enables experts to build reusable pipelines that perform the whole ETL procedure from tabular data to RDF in a controllable and sustainable manner.

Silk: The Linked Data Integration Framework will be used to create automated links, where possible. It can be integrated with LinkedPipesETL, to make automated linking part of the ETL procedure, if possible.

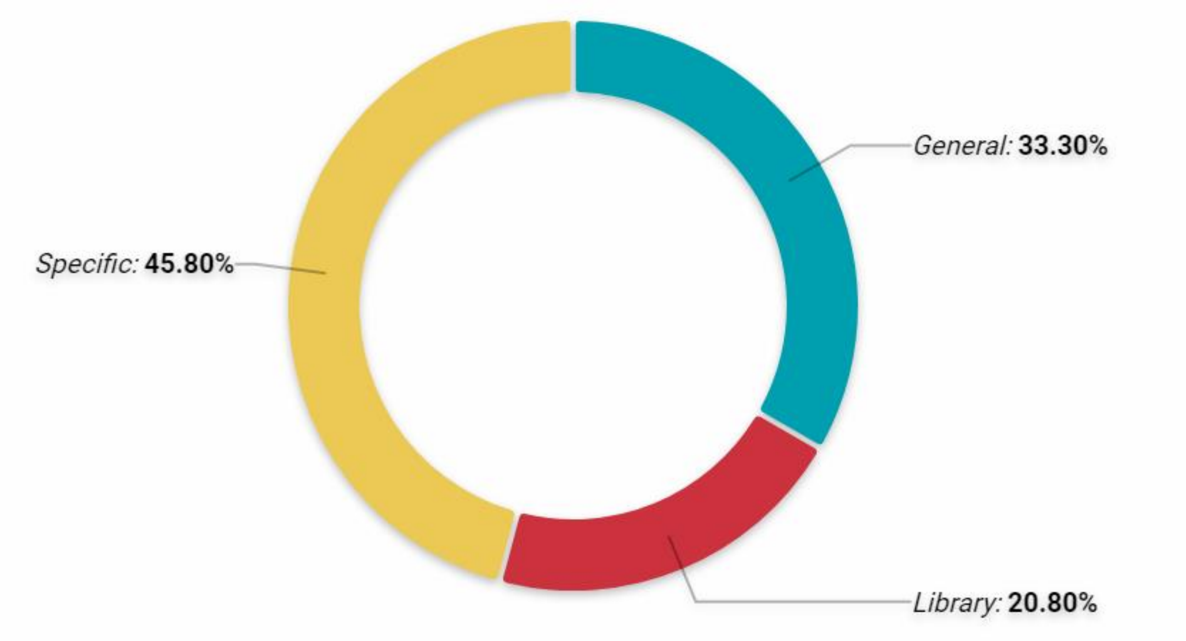
VocBench: VocBench 3 (or, simply, VB3) offers a powerful editing environment, with facilities for management of OWL ontologies, SKOS/SKOS-XL thesauri, OntoLex lexicons and any sort of RDF dataset. It aims to set new standards for flexibility, openness and expressive power as a free and open source RDF modelling platform. Funded by the European Commission ISA² programme, the development of Vocbench 3 (VB3) is managed by the Publications Office of the EU, to curate KG of the EU like EuroVoc. VocBench will be used to curate the whole KG of the PhDHub.

Virtuoso: Virtuoso is a scalable cross-platform server that combines Relational, Graph, and Document Data Management with Web Application Server and Web Services Platform functionality. It will be used to store the the KGs as triples and build services upon.

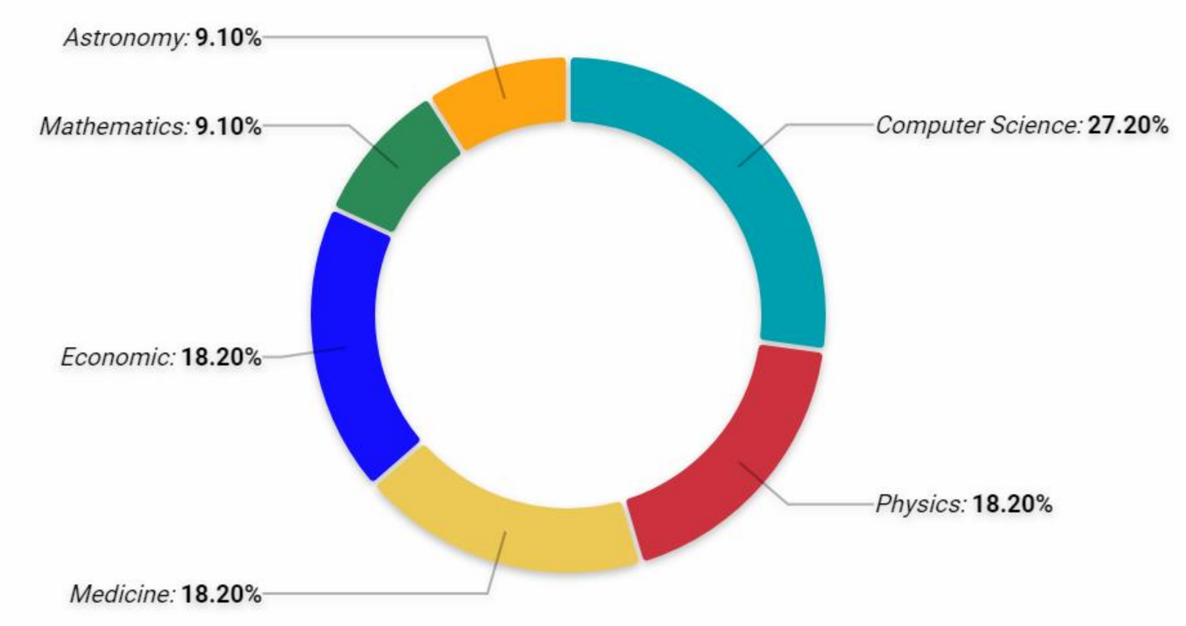
With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

3 Classifications Overview

For the purposes of the deliverable, a survey of scientific classifications was conducted, in order to identify the ones that are the most common and complete for the knowledge graph to build upon. A total of 24 classifications were identified, where eight of them are general classifications, five of them are library classifications and 11 of them are classifications of a specific scientific field.



Specific classifications split up into several distinct scientific fields, such as Computer Science and Engineering (3), Physics (2), Astronomy (1), Mathematics (1), Medicine (2) and Economy (2).



With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

3.1 Classifications Features

An overview of the classifications based on four key features is shown in the next table. These key features are boolean attributes (true of false), such as if the classification is:

- in RDF format
- open
- hierarchical
- available in multiple languages

For many of the classifications with no open access, it was unclear whether some of their specific features are true or false.

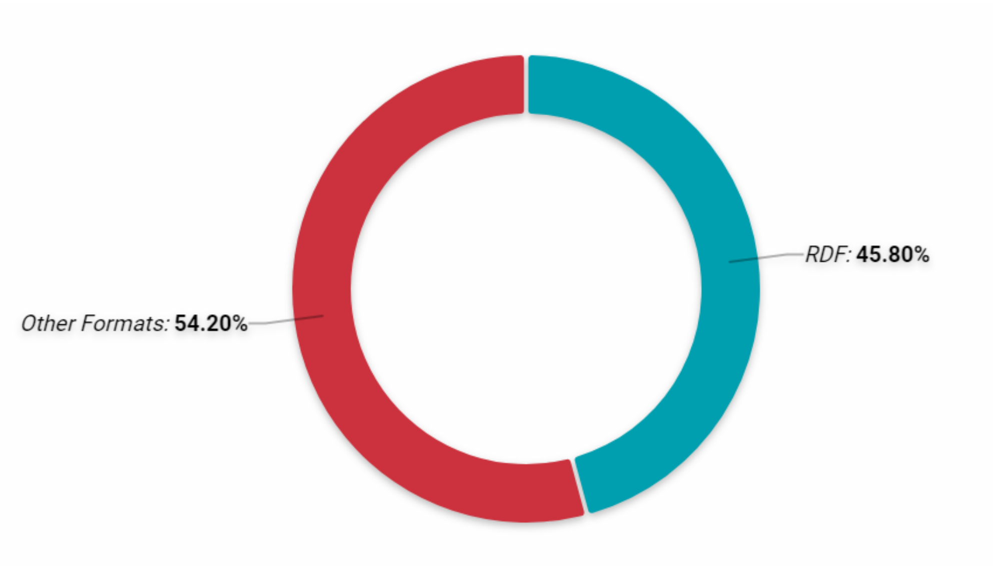
3.1.1 All Classifications

All Classifications	True	False	Unclear
RDF	11	13	-
Open	19	5	-
Multiple Languages	12	8	4
Hierarchical	18	2	4

Format (RDF of others) of all classifications

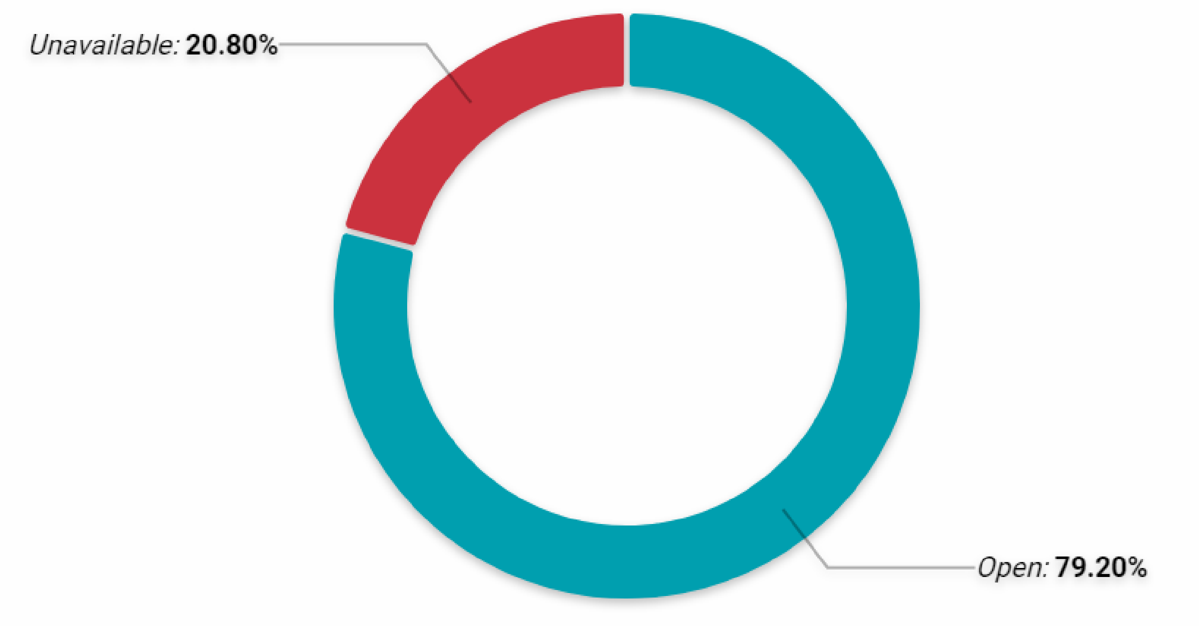
11 out of 24 classifications are available in RDF format, while the rest 13 are available only to other formats (XLS, PDF and others).

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220



Open and Unavailable classifications

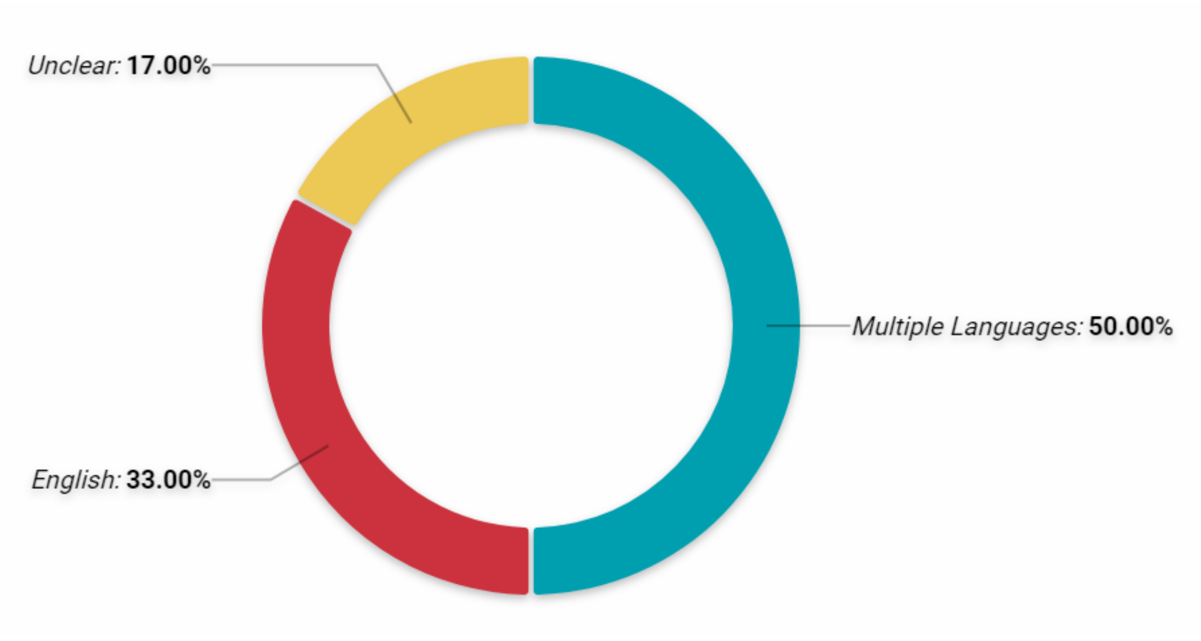
19 out of 24 classifications are open and accessible as full lists of concepts, while the rest five have restricted access to their contents (only broader research areas or browsing features).



Language (only English or multiple languages) of all classifications

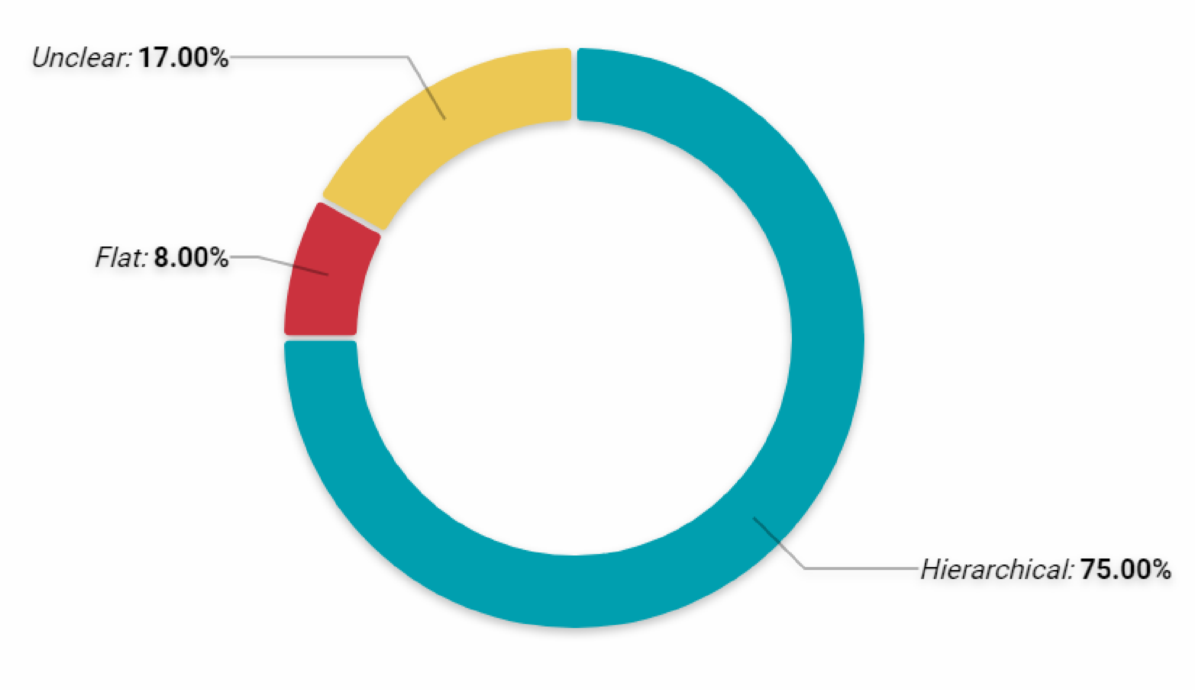
Half of the classifications are available in additional languages, besides English. Eight of them are available only in English, while the restricted access to four classifications did not allow identify multiple languages (unclear status).

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220



Structure (hierarchical or flat) of all classifications

18 out of 24 classifications have an hierarchical structure, while only two of them are flat. The restricted access to four classifications did not allow identify their structure (unclear status).



3.1.2 General Classifications

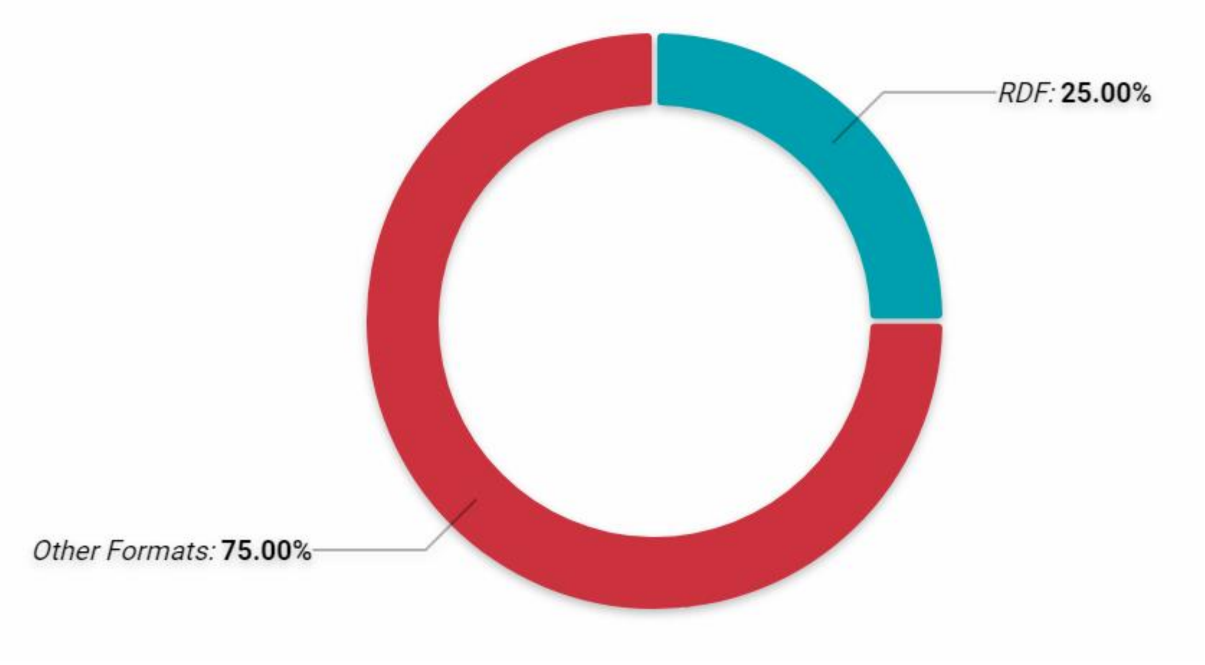
With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

An overview of the general classifications, based on the four key features (RDF, open, multiple languages, hierarchical), is shown in the next table.

General Classifications	True	False	Unclear
RDF	2	6	-
Open	8	-	-
Multiple Languages	4	4	-
Hierarchical	8	-	-

Format (RDF of others) of general classifications

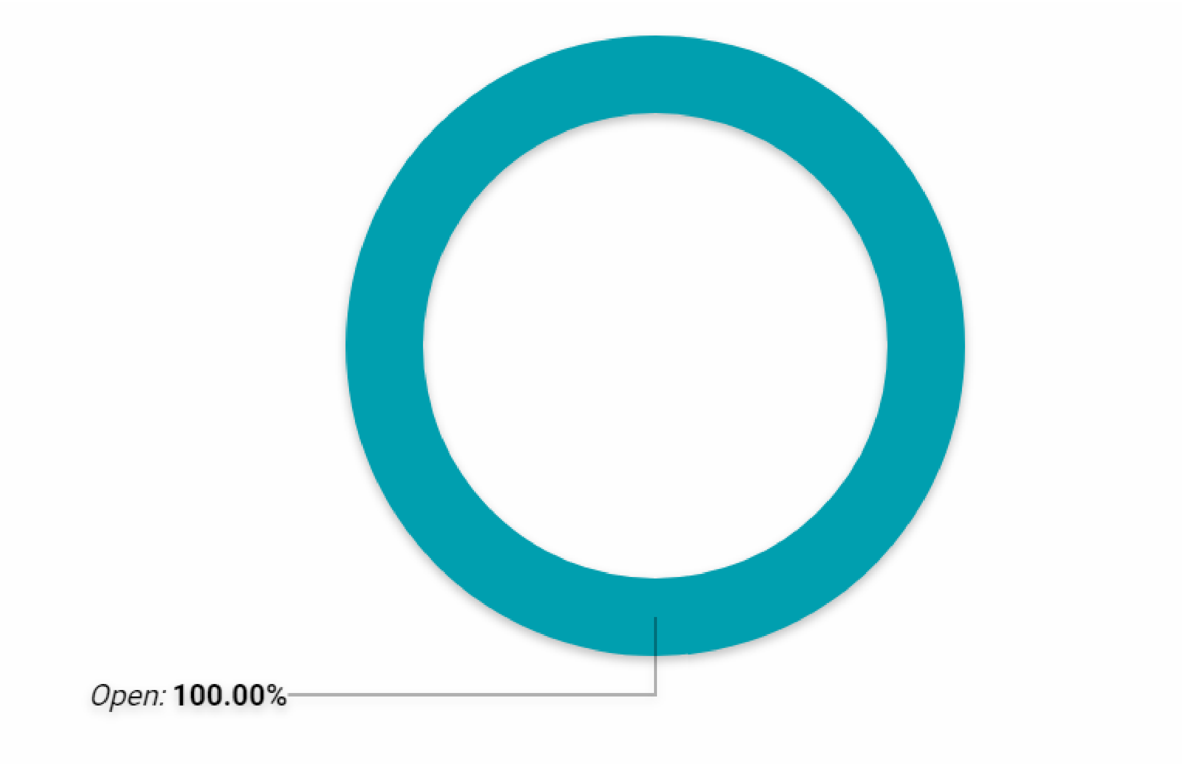
Only two out of eight general classifications are available in RDF format, while six are available only to other formats (XLS, PDF and others).



Open and Unavailable general classifications

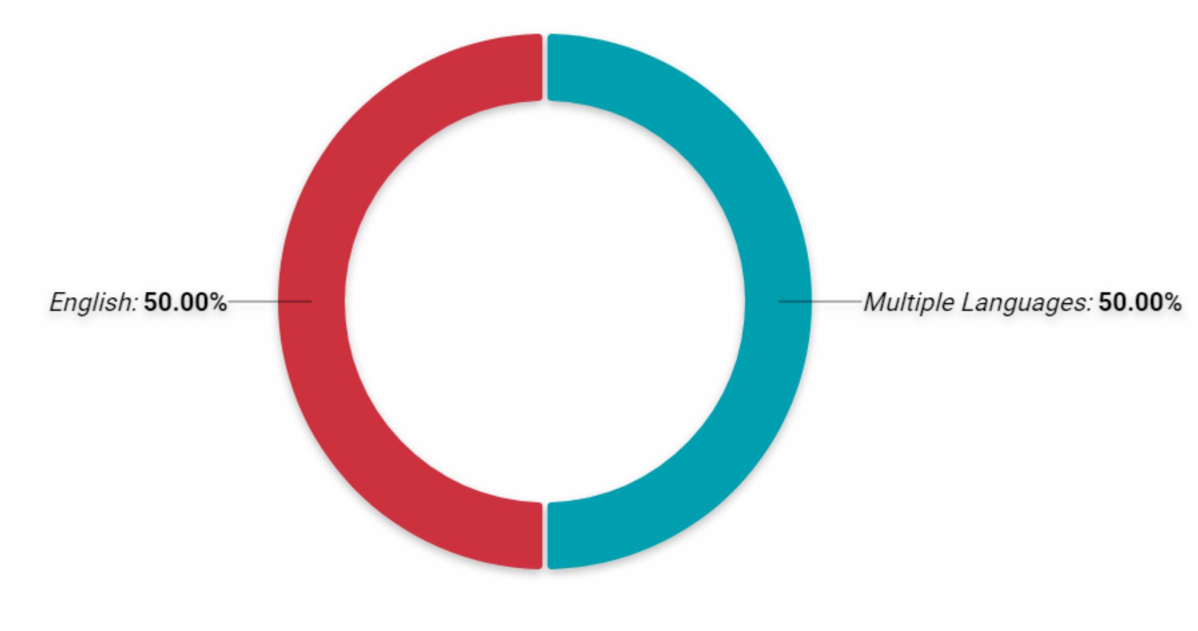
All eight general classifications are open and accessible as full lists of concepts.

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220



Language (only English or multiple languages) of general classifications

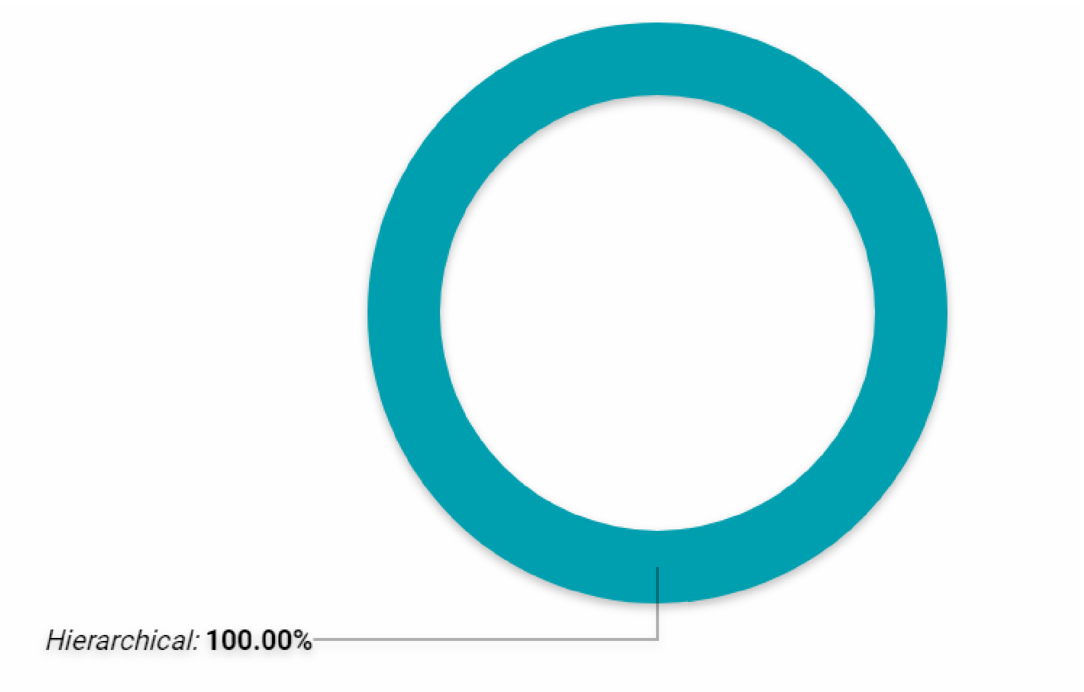
Half of the general classifications are available in multiple languages, while the other half are available only in English.



Structure (hierarchical or flat) of general classifications

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

All eight general classifications have an hierarchical structure.



3.1.3 Library Classifications

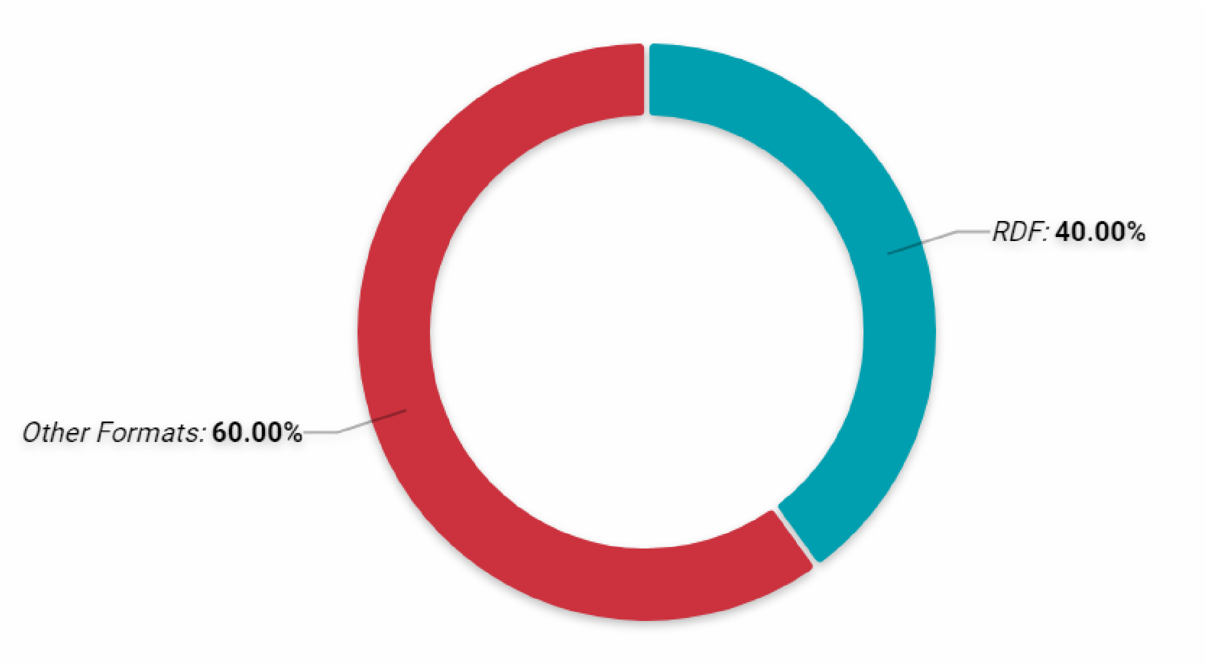
An overview of the library classifications based on the four key features (RDF, open, multiple languages, hierarchical) is shown in the next table.

Library Classifications	True	False	Unclear
RDF	2	3	-
Open	2	3	-
Multiple Languages	2	1	2
Hierarchical	2	1	2

Format (RDF of others) of all classifications

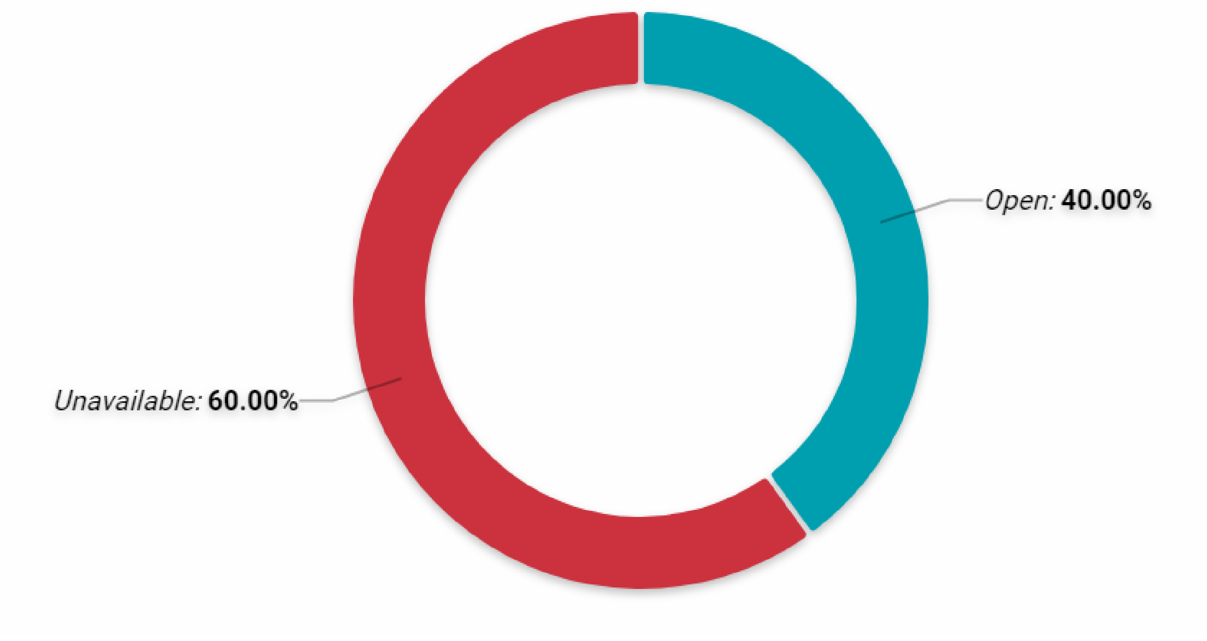
Two out of five library classifications are available in RDF format, while the rest three are available only to other formats, or unavailable at all.

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220



Open and Unavailable classifications

Two out of five classifications are open and accessible as full lists of concepts, while the remaining three provide restricted or no access to their contents (only broader research areas or browsing features).



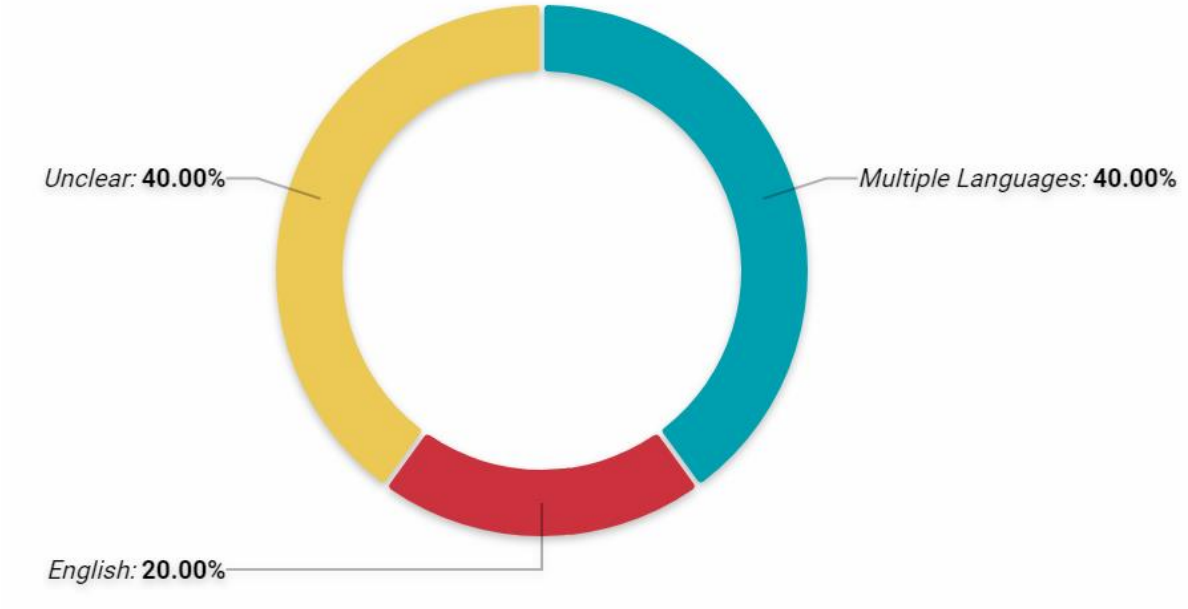
Language (only English or multiple language) of all classifications

With the support of Erasmus+

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

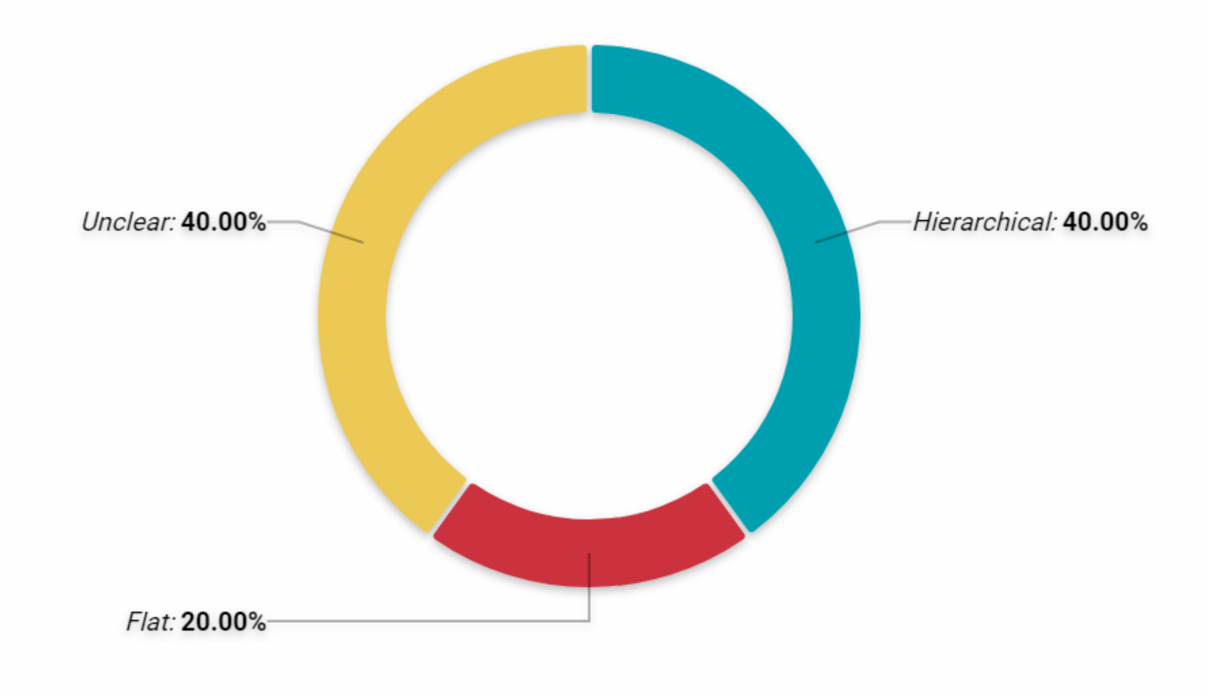
Two out of the library classifications are available in additional languages, besides English. One of them is available only in English, while the restricted access to the remaining two classifications did not allow identify multiple languages.



Structure (hierarchical or flat) of all classifications

Two out of five library classifications have an hierarchical structure, while one of them is flat. The restricted access to the remaining two classifications did not allow identify their structure (unclear status).

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220



3.1.4 Specific Classifications

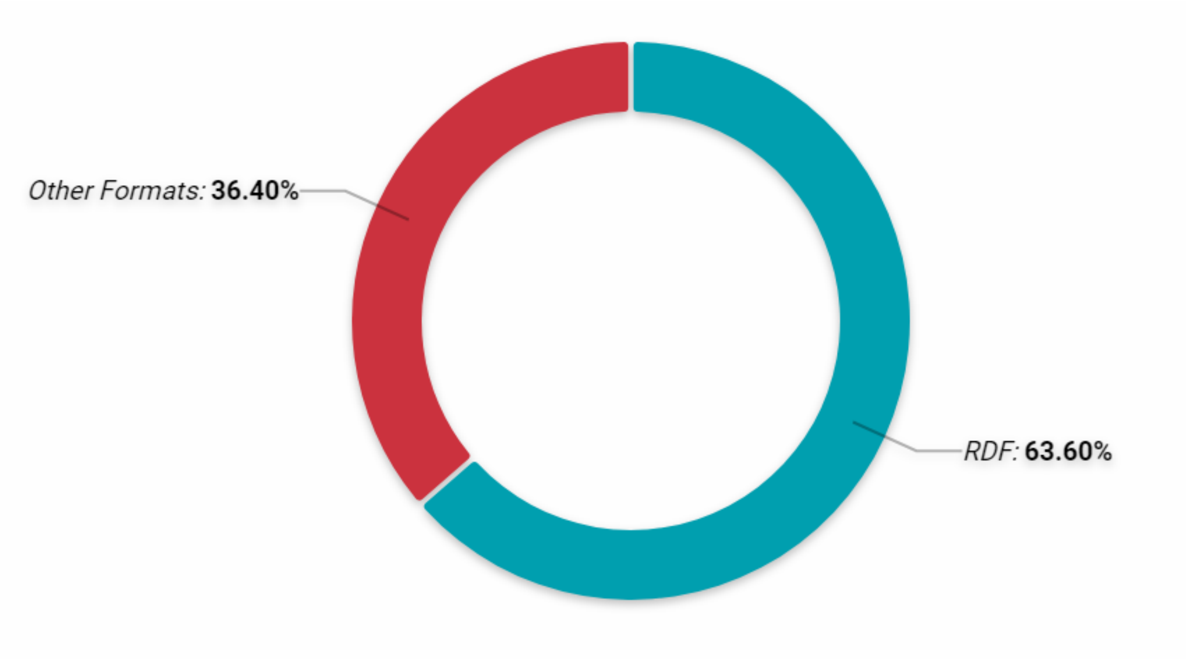
An overview of the specific classifications based on the four key features (RDF, open, multiple languages, hierarchical) is shown in the next table.

Specific Classifications	True	False	Unclear
RDF	7	4	-
Open	9	2	-
Multiple Languages	6	3	2
Hierarchical	8	1	2

Format (RDF of others) of specific classifications

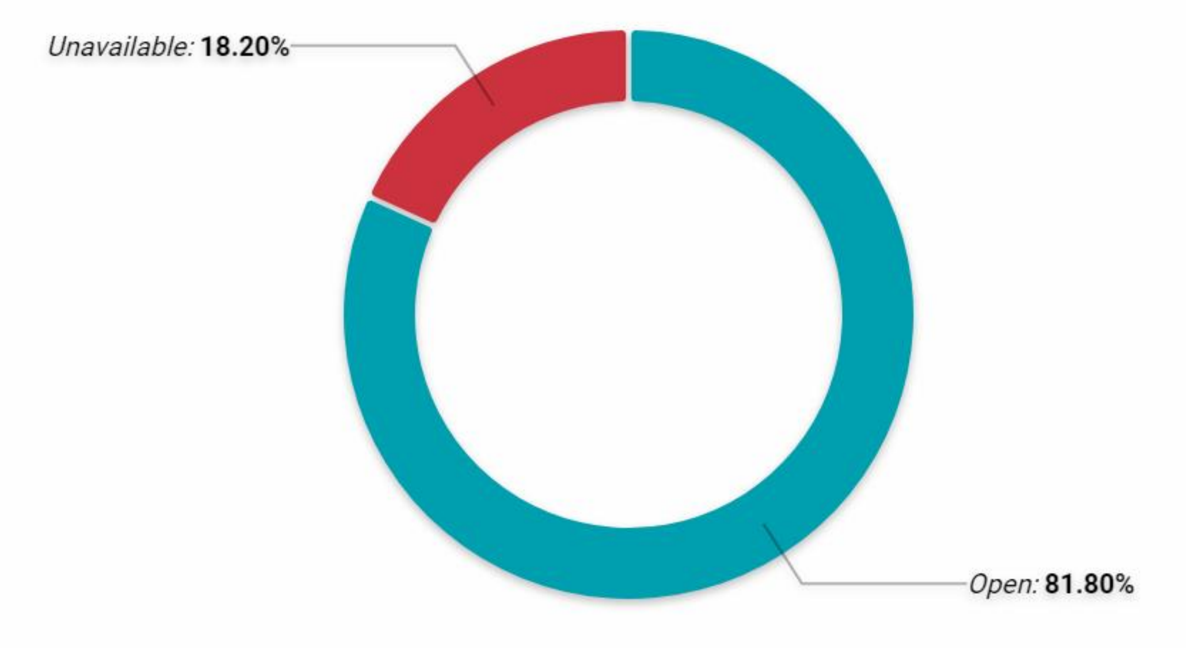
Seven out of 11 specific classifications are available in RDF format, while the rest four are available only to other formats (XLS, PDF and others), or unavailable at all.

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220



Open and Unavailable specific classifications

Nine out of 11 specific classifications are open and accessible as full lists of concepts, while the remaining two provide restricted access to their contents (only broader research areas or browsing features).



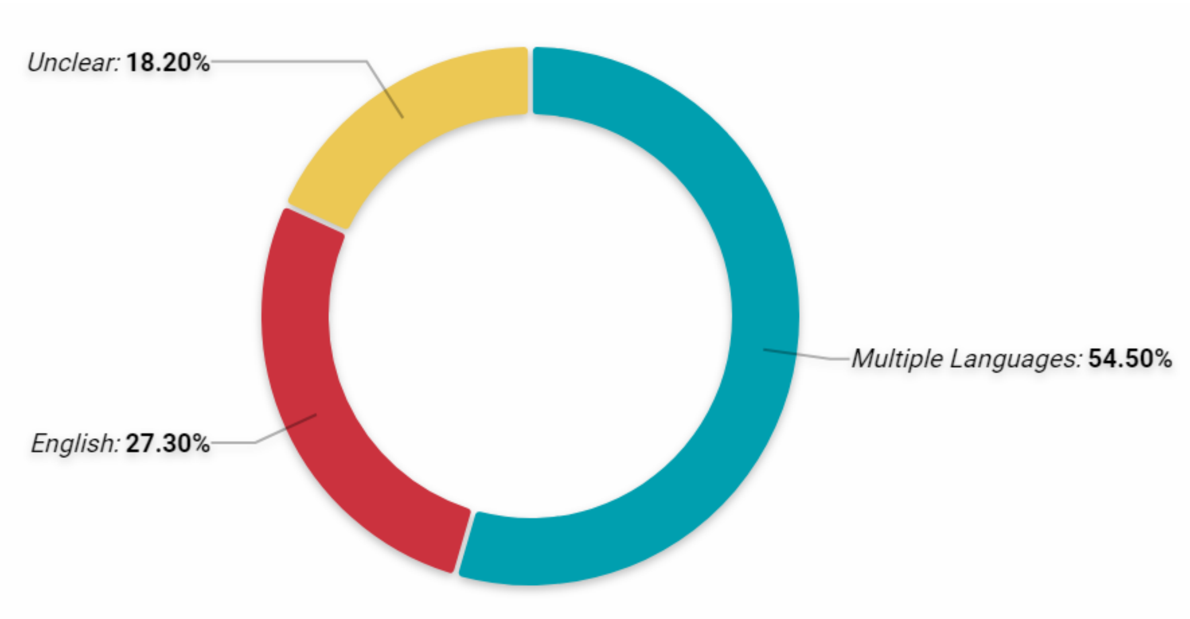
Language (only English or multiple language) of specific classifications

With the support of Erasmus+

Co-funded by the Erasmus+ Programme of the European Union

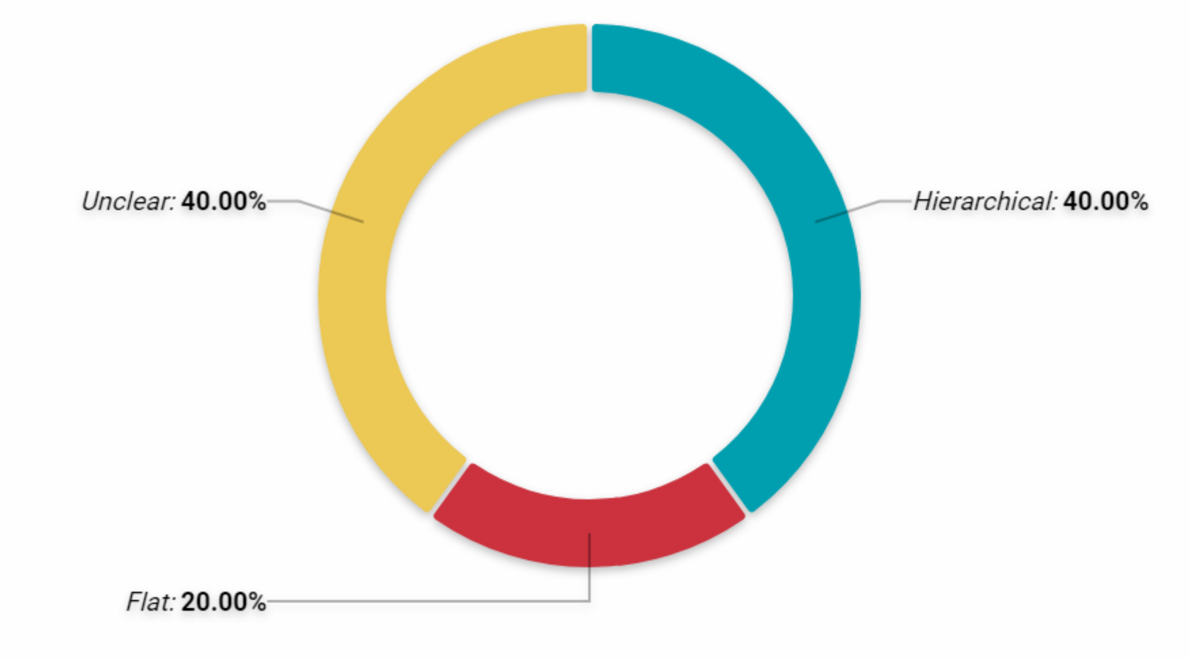
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

Six out of 11 scientific classifications are available in additional languages, besides English. Three of them are available only in English, while the restricted access to the remaining two classifications did not allow identify multiple languages.



Structure (hierarchical or flat) of specific classifications

Eight out of 11 specific classifications have an hierarchical structure, while only one of them is flat. The restricted access to the remaining two classifications did not allow identify their structure (unclear status).



With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

3.2 Classification Attributes

The identified classifications are presented in detail, through the following tables, which include information such as their name, abbreviations, category, source, source link, format and whether they are available in SKOS, structure, number of concepts, languages, additional related links and notes, where available.

3.2.1 General Classifications

3.2.1.1 UNESCO nomenclature for fields of science and technology

Field	Value
Name	UNESCO nomenclature for fields of science and technology
Abbreviation	
Category	General
Source	UNESCO
Source Link	SKOS
Format	Turtle, RDF/XML
original in SKOS	true
Structure	hierarchical
No. of concepts	2504
Other Links	Overview
Notes	2007 version. Same concepts as the Fields of Science and Technology classification
Language	EN, FR, ES

3.2.1.2 Fields of Science and Technology

Field	Value

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

Name	Fields of Science and Technology
Abbreviation	FOS
Category	General
Source	OECD
Source Link	Eurostat
Format	XML, CSV, XLS, PDF
original in SKOS	false
Structure	hierarchical
No. of concepts	52
Other Links	Manual, Concepts
Notes	Published 2002, Revised 2007. Same concepts as the UNESCO nomenclature for fields of science and technology which is older (1973, 1974)
Language	EN, FR, DE

3.2.1.3 Classification of fields of education and training

Field	**Value**
Name	Classification of fields of education and training (1999)
Abbreviation	
Category	General
Source	CSO
Source Link	Eurostat
Format	XML, CSV, XLS, PDF
original in SKOS	false
Structure	hierarchical
No. of concepts	127
Other Links	SKOSified ISCED, Description, Concepts

Notes	A more generic classification. This classification was derived from a joint project with Eurostat and UNESCO. It is based on the International Standard Classification of Education - ISCED 1997 (Fields of Education). In 1999, Eurostat and UNESCO further elaborated on the ISCED 1997 and provided a third level of about 80 categories called 'Detailed Fields'. This additional level was added to allow the EUROSTAT Classification of Training developed in 1997-98 to be integrated with ISCED 1997 and features subjects which would normally be covered as part of vocational education. Same with or similar to ISCED-F 2013.
Language	EN, FR, DE

3.2.1.4 ISCED Fields of education and training

Field	Value
Name	ISCED Fields of education and training
Abbreviation	ISCED-F 2013
Category	General
Source	UNESCO
Source Link	Eurostat
Format	XML, CSV, XLS, PDF
original in SKOS	false
Structure	hierarchical
No. of concepts	229
Other Links	Manual, Mappings
Notes	With the revision of ISCED 1997 in ISCED 2011, it was agreed that the fields of education should be examined in a separate process establishing an independent but related classification called the ISCED Fields of Education and Training (ISCED-F). Same with or similar to Classification of fields of education and training.
Language	EN, FR, DE

With the support of Erasmus+

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

3.2.1.5 Joint Academic Coding System

Field	Value
Name	Joint Academic Coding System
Abbreviation	JACS
Category	General
Source	HESA and UCAS
Source Link	RDF
Format	XML, CSV, XLS, PDF
original in SKOS	true
Structure	hierarchical
No. of concepts	over 1600
Other Links	Info, Classification - Related Files
Notes	The Joint Academic Coding System (JACS) system is used by the Higher Education Statistics Agency (HESA) and the Universities and Colleges Admissions Service (UCAS) in the United Kingdom to classify academic subjects. It is due to be replaced by the Higher Education Coding System (HECoS) and the Common Aggregation Hierarchy (CAH) for the 2019/20 academic year.
Language	EN

3.2.1.6 Classification of Instructional Programs

Field	Value
Name	Classification of Instructional Programs
Abbreviation	CIP
Category	General
Source	U.S. Department of Education's National Center for Education Statistics (NCES)
Source Link	Classification files
Format	XLS

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

original in SKOS	false
Structure	hierarchical
No. of concepts	over 1K
Other Links	
Notes	The purpose of the Classification of Instructional Programs (CIP) is to provide a taxonomic scheme that will support the accurate tracking, assessment, and reporting of fields of study and program completions activity. CIP was originally developed by the U.S. Department of Education's National Center for Education Statistics (NCES) in 1980, with revisions occurring in 1985 and 1990. The 2000 edition (CIP-2000) is the third revision of the taxonomy and presents an updated taxonomy of instructional program classifications and descriptions.
Language	EN

3.2.1.7 Australian Standard Classification of Education

Field	Value
Name	Australian Standard Classification of Education
Abbreviation	ASCED
Category	General
Source	Australian bureau of statistics
Source Link	Classification
Format	XLS
original in SKOS	false
Structure	hierarchical
No. of concepts	439 (Field of Education part)
Other Links	Related Info
Notes	ASCED comprises two classifications: Level of Education and Field of Education.
Language	EN

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

3.2.1.8 Australian and New Zealand Standard Research Classification

Field	Value
Name	Australian and New Zealand Standard Research Classification
Abbreviation	ANZSRC
Category	General
Source	Australian bureau of statistics
Source Link	Downloads
Format	PDF
original in SKOS	false
Structure	hierarchical
No. of concepts	
Other Links	Description
Notes	
Language	EN

3.2.2 Library Classifications

3.2.2.1 Spinger SciGraph Subjects

Field	Value
Name	Spinger Subjects
Abbreviation	
Category	Library
Source	Springer
Source Link	SciGraph Subjects, Subject Codes

With the support of Erasmus+

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

Format	TTL, XLS
original in SKOS	true
Structure	hierarchical
No. of concepts	1470
Other Links	Info, SciGraph
Notes	
Language	EN, DE

3.2.2.2 Library of Congress Classification

Field	Value
Name	Library of Congress Classification
Abbreviation	LCSH
Category	Library
Source	Library of Congress
Source Link	datahub
Format	RDF, N-Triples
original in SKOS	true
Structure	hierarchical
No. of concepts	
Other Links	Info, Outline
Notes	The Library of Congress Classification (LCC) is a system of library classification developed by the Library of Congress. It is used by most research and academic libraries in the U.S. and several other countries.
Language	EN

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

3.2.2.3 Dewey Decimal Classification

Field	Value
Name	Dewey Decimal Classification
Abbreviation	
Category	Library
Source	OCLC
Source Link	datahub
Format	
original in SKOS	false
Structure	
No. of concepts	
Other Links	Info
Notes	Website http://dewey.info/ not currently working.
Language	

3.2.2.4 arXiv

Field	Value
Name	arXiv
Abbreviation	arXiv
Category	Library
Source	Cornell University
Source Link	datahub
Format	
original in SKOS	false
Structure	
No. of concepts	

With the support of Erasmus+

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

Other Links	arxiv.org/, Wiki Info
Notes	
Language	

3.2.2.5 Cambridge University Library Classification

Field	Value
Name	Cambridge University Library Classification
Abbreviation	
Category	Library
Source	Cambridge University
Source Link	datahub, outline PDF
Format	PDF
original in SKOS	false
Structure	false
No. of concepts	
Other Links	Classification Scheme
Notes	
Language	EN

3.2.3 Specific Classifications

3.2.3.1 ACM Computing Classification System

Field	Value
Name	ACM Computing Classification System
Abbreviation	CCS
Category	Computer Science
Source	Association for Computing Machinery (ACM)

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

Source Link	Full Classification
Format	XML, HTML, Word
original in SKOS	true
Structure	hierarchical
No. of concepts	over 2400
Other Links	Flat List. Wiki Info
Notes	
Language	EN

3.2.3.2 Computing Research Repository

Field	Value
Name	Computing Research Repository
Abbreviation	CoRR
Category	Computer Science
Source	Through a partnership of ACM, the arXiv.org e-print archive, and NCSTRL(Networked Computer Science Technical Reference Library), an online Computing Research Repository (CoRR) has been established.
Source Link	[Subject Classes](https://arxiv.org/corr/subjectclasses)
Format	HTML
original in SKOS	false
Structure	flat
No. of concepts	40
Other Links	About, FAQ
Notes	Papers in CoRR are classified in two ways: by subject area from a list of subjects listed below and by using the 1998 ACM Computing Classification System. The ACM classification scheme provides us with a relatively stable scheme that covers all of computer science. The subject areas are not mutually

	exclusive, nor do they (yet) provide complete coverage of the field. On the other hand, we hope that they better reflect the active areas of research in CS. We expect to add more subject areas and subdivide current subject areas according to demand. Authors who cannot find an appropriate subject area should use subject area Other. Through a partnership of ACM, the LANL (Los Alamos National Laboratory) e-Print archive, and NCSTRL (Networked Computer Science Technical Reference Library), an online Computing Research Repository (CoRR) was established
Language	EN

3.2.3.3 Institute of Electrical and Electronics Engineers Taxonomy

Field	**Value**
Name	Institute of Electrical and Electronics Engineers Taxonomy
Abbreviation	IEEE
Category	electrical engineering, computer science
Source	Merger of the American Institute of Electrical Engineers and the Institute of Radio Engineers
Source Link	PDF, Thesaurus
Format	PDF
original in SKOS	false
Structure	hierarchical
No. of concepts	over 6500
Other Links	Wiki Info
Notes	In PDF format following NISO format. It contains a list in two column format, terms are abbreviated. Also the license does not allow any modification without permission. It is an extensive enough list but has a lot of limitations. We could request official permission to transform and publish the taxonomy as RDF in SKOS if it is needed
Language	EN

3.2.3.4 Mathematics Subject Classification

Field	Value
Name	Mathematics Subject Classification
Abbreviation	MSC
Category	Mathematics
Source	Mathematical Reviews and Zentralblatt MATH
Source Link	RDF, PDF
Format	RDF, Turtle, N-Triples, PDF
original in SKOS	true
Structure	hierarchical
No. of concepts	over 8000
Other Links	Related Info, Wiki Info
Notes	Relation to other classification schemes: For physics papers the Physics and Astronomy Classification Scheme (PACS) is often used. Due to the large overlap between mathematics and physics research it is quite common to see both PACS and MSC codes on research papers, particularly for multidisciplinary journals and repositories such as the arXiv. The ACM Computing Classification System (CCS) is a similar hierarchical classification scheme for computer science. There is some overlap between the AMS and ACM classification schemes, in subjects related to both mathematics and computer science, however the two schemes differ in the details of their organization of those topics. The classification scheme used on the arXiv is chosen to reflect the papers submitted. As arXiv is multidisciplinary its classification scheme does not fit entirely with the MSC, ACM or PACS classification schemes. It is common to see codes from one or more of these schemes on individual papers.
Language	EN, ZH, IT

3.2.3.5 Physics Subject Headings

Field	Value

With the support of Erasmus+

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

Name	Physics Subject Headings
Abbreviation	PhySH
Category	Physics
Source	American Physical Society (APS)
Source Link	
Format	
original in SKOS	false
Structure	
No. of concepts	
Other Links	Wiki Info, Browser
Notes	"Since 2016, American Physical Society introduced the PhySH (Physics Subject Headings) system instead of PACS. PhySH is copyrighted with all rights reserved by the American Physical Society. We are still considering what license we would use for any public release of PhySH."
Language	EN

3.2.3.6 Physics and Astronomy Classification Scheme

Field	**Value**
Name	Physics and Astronomy Classification Scheme
Abbreviation	PACS
Category	Physics, Astronomy
Source	American Institute of Physics (AIP)
Source Link	RDF/XML, HTML
Format	RDF/XML
original in SKOS	true
Structure	hierarchical

No. of concepts	over 3700
Other Links	Regular Edition, Alphabetical Index, Wiki Info
Notes	
Language	EN

3.2.3.7 Astrothesaurus

Field	Value
Name	Astrothesaurus
Abbreviation	
Category	Astronomy
Source	American Astronomical Society (AAS)
Source Link	GitHub
Format	RDF, JSON, CSV
original in SKOS	true
Structure	hierarchical
No. of concepts	2706
Other Links	Astrothesaurus Project
Notes	Community-based. This Unified Astronomy Thesaurus (UAT) is an open, interoperable and community-supported thesaurus which unifies the existing divergent and isolated Astronomy & Astrophysics thesauri into a single high-quality, freely-available open thesaurus formalizing astronomical concepts and their inter-relationships. The UAT builds upon the existing IAU Thesaurus with major contributions from the Astronomy portions of the thesauri developed by the Institute of Physics Publishing and the American Institute of Physics. We expect that the Unified Astronomy Thesaurus will be further enhanced and updated through a collaborative effort involving broad community participation. While the AAS has assumed formal ownership of the UAT, the work will be available under a Creative Commons License, ensuring its widest use while protecting the intellectual property of the contributors.

With the support of Erasmus+

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

Language	EN

3.2.3.8 Medical Subject Headings

Field	Value
Name	Medical Subject Headings
Abbreviation	MeSH
Category	Medicine
Source	US National Library of Medicine
Source Link	Files
Format	RDF, XML, PDF
original in SKOS	true
Structure	hierarchical
No. of concepts	
Other Links	Files & Info, Extra Files, treeView
Notes	
Language	EN

3.2.3.9 Unified Medical Language System

Field	Value
Name	Unified Medical Language System
Abbreviation	UMLS
Category	Medicine
Source	US National Library of Medicine
Source Link	
Format	
original in SKOS	false

Structure	
No. of concepts	
Other Links	RDFising Tool, UMLS Home, Knowledge Sources, Terminology Service
Notes	
Language	EN

3.2.3.10 JEL classification system

Field	Value
Name	JEL classification system
Abbreviation	JEL
Category	Economics
Source	Journal of Economic Literature (JEL)
Source Link	datahub, Classification Tree
Format	RDF, XML
original in SKOS	true
Structure	hierarchical
No. of concepts	1000
Other Links	JEL Codes, Guide
Notes	The JEL classification system was developed for use in the Journal of Economic Literature (JEL), and is a standard method of classifying scholarly literature in the field of economics. The system is used to classify articles, dissertations, books, book reviews, and working papers in EconLit, and in many other applications. For descriptions and examples, see the JEL Codes Guide.
Language	EN, FR, DE, ES

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

3.2.3.11 STW Thesaurus for Economics

Field	Value
Name	STW Thesaurus for Economics
Abbreviation	STW
Category	Economics
Source	German National Library of Economics
Source Link	ZBW
Format	RDF, Ntriples, Turtle
original in SKOS	true
Structure	hierarchical
No. of concepts	6000
Other Links	info, mappings
Notes	The STW Thesaurus for Economics is the world's most comprehensive bilingual thesaurus for representing and searching for economics-related content. With its almost 6,000 subject headings in English and German and more than 20,000 synonyms it covers all economics-related subject areas and, on a broader level, the most important related subject fields. The STW is published and continuously further developed by the ZBW according to the latest changes in the economic terminology.
Language	EN, DE

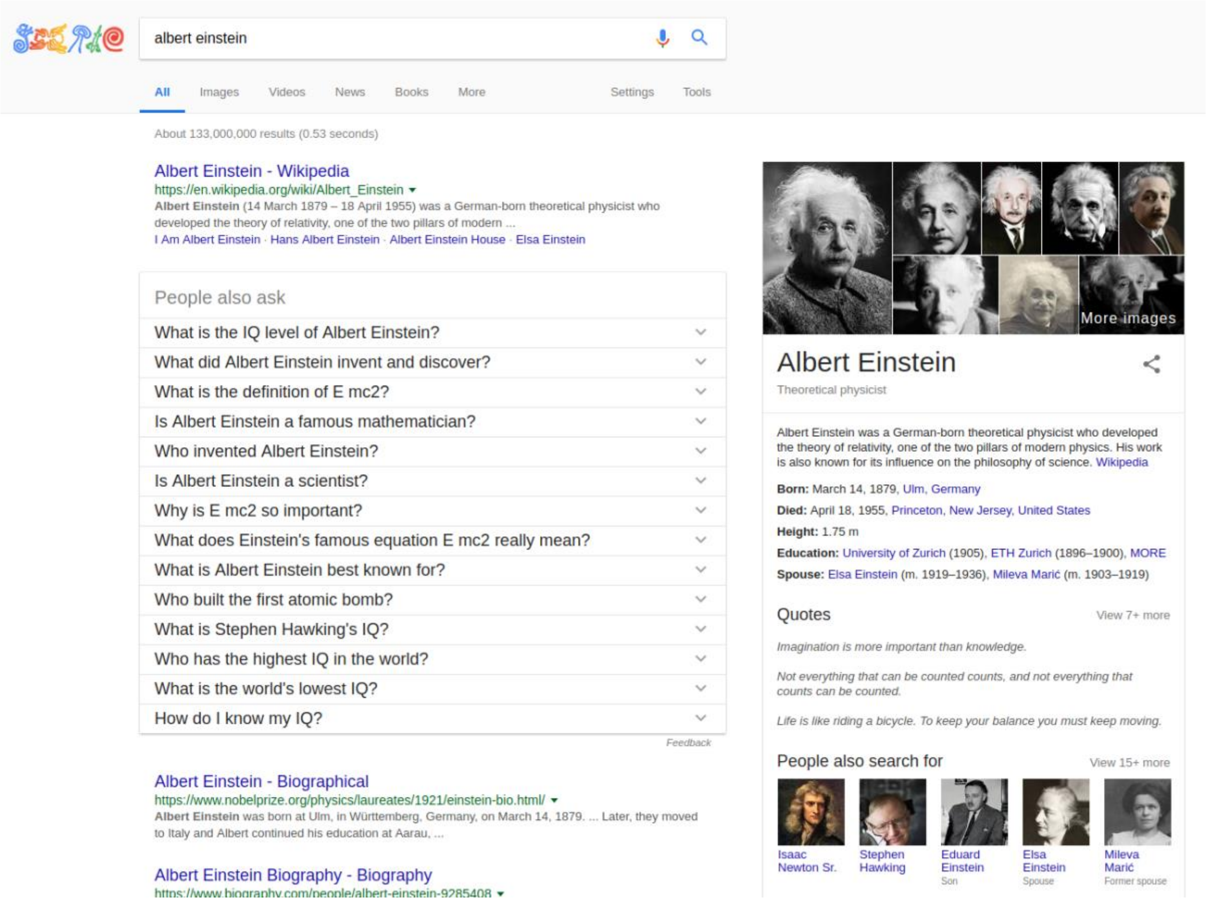
With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

4 Building the Core Knowledge Graph

4.1 What is a Knowledge Graph

An Ontology in computer science is a representation of Knowledge, in a machine readable format, thus parsable from computer agents. According to Wikipedia[4] "...an ontology encompasses a representation, formal naming, and definition of the categories, properties, and relations of the concepts, data, and entities that substantiate one, many, or all domains."

Ontologies are also known as Knowledge Graphs, or more recently Knowledge Vaults. The term appeared when Google presented an initiative called Knowledge Graph, a knowledge base used by Google and other in-house services, in order to enhance its search engine results, with information gathered from a variety of sources, most notably Wikipedia, DBpedia and Freebase[5]. For instance, you can search on the Google Search Engine about the term "albert einstein" and we get the following results page:



About 133,000,000 results (0.53 seconds)

Albert Einstein - Wikipedia
https://en.wikipedia.org/wiki/Albert_Einstein
Albert Einstein (14 March 1879 – 18 April 1955) was a German-born theoretical physicist who developed the theory of relativity, one of the two pillars of modern ...
I Am Albert Einstein · Hans Albert Einstein · Albert Einstein House · Elsa Einstein

People also ask

What is the IQ level of Albert Einstein?
What did Albert Einstein invent and discover?
What is the definition of E mc2?
Is Albert Einstein a famous mathematician?
Who invented Albert Einstein?
Is Albert Einstein a scientist?
Why is E mc2 so important?
What does Einstein's famous equation E mc2 really mean?
What is Albert Einstein best known for?
Who built the first atomic bomb?
What is Stephen Hawking's IQ?
Who has the highest IQ in the world?
What is the world's lowest IQ?
How do I know my IQ?

Albert Einstein - Biographical
https://www.nobelprize.org/physics/laureates/1921/einstein-bio.html/
Albert Einstein was born at Ulm, in Württemberg, Germany, on March 14, 1879. ... Later, they moved to Italy and Albert continued his education at Aarau, ...

Albert Einstein Biography - Biography
https://www.biography.com/people/albert-einstein-9285408

Albert Einstein
Theoretical physicist

Albert Einstein was a German-born theoretical physicist who developed the theory of relativity, one of the two pillars of modern physics. His work is also known for its influence on the philosophy of science. Wikipedia

Born: March 14, 1879, Ulm, Germany
Died: April 18, 1955, Princeton, New Jersey, United States
Height: 1.75 m
Education: University of Zurich (1905), ETH Zurich (1896–1900), MORE
Spouse: Elsa Einstein (m. 1919–1936), Mileva Marić (m. 1903–1919)

Quotes — View 7+ more

Imagination is more important than knowledge.

Not everything that can be counted counts, and not everything that counts can be counted.

Life is like riding a bicycle. To keep your balance you must keep moving.

People also search for — View 15+ more

Isaac Newton Sr.
Stephen Hawking
Eduard Einstein — Son
Elsa Einstein — Spouse
Mileva Marić — Former spouse

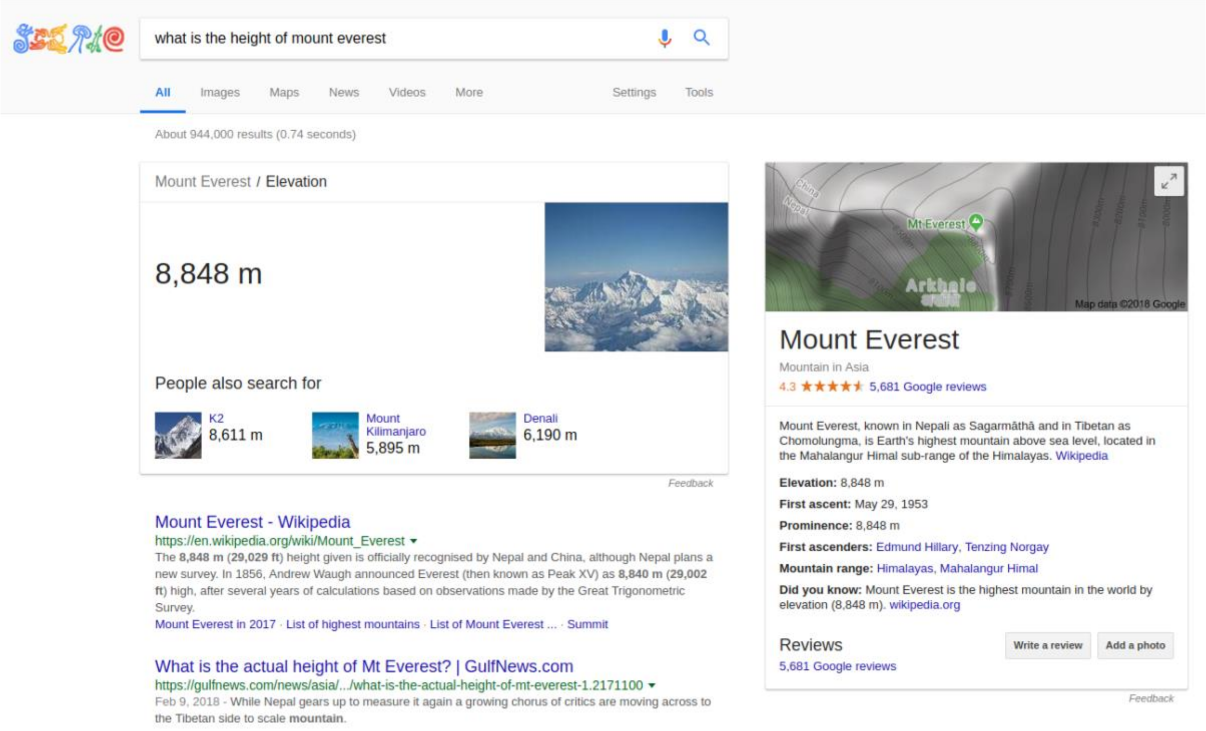
On the left side of the page, the organic results of the query are presented - usually web pages or documents that contain the query term. On the right side of the page, there is a box,

[4] https://en.wikipedia.org/wiki/Ontology_(information_science)

[5] https://en.wikipedia.org/wiki/Knowledge_Graph

presenting information, in a semi-structured format about the People, named "Albert Einstein". This is information found on Google's Knowledge Graph, collected from a variety of sources. One can see information such as the birth/death date and place of Albert Einstein, his height, spouses, photos of him, a short biography and some of his famous quotes.

A more interesting effect of the utilization of the Knowledge Graph in Google's search engine is that it can actually return an answer on specific queries. For instance if your query is "what is the height of mount everest", the following result page is retrieved:



what is the height of mount everest

About 944,000 results (0.74 seconds)

Mount Everest / Elevation

8,848 m

People also search for

K2 8,611 m · Mount Kilimanjaro 5,895 m · Denali 6,190 m

Mount Everest - Wikipedia
https://en.wikipedia.org/wiki/Mount_Everest
The **8,848 m** (**29,029 ft**) height given is officially recognised by Nepal and China, although Nepal plans a new survey. In 1856, Andrew Waugh announced Everest (then known as Peak XV) as **8,840 m** (**29,002 ft**) high, after several years of calculations based on observations made by the Great Trigonometric Survey.
Mount Everest in 2017 · List of highest mountains · List of Mount Everest ... · Summit

What is the actual height of Mt Everest? | GulfNews.com
https://gulfnews.com/news/asia/.../what-is-the-actual-height-of-mt-everest-1.2171100
Feb 9, 2018 - While Nepal gears up to measure it again a growing chorus of critics are moving across to the Tibetan side to scale **mountain**.

Mount Everest
Mountain in Asia
4.3 ★★★★★ 5,681 Google reviews

Mount Everest, known in Nepali as Sagarmāthā and in Tibetan as Chomolungma, is Earth's highest mountain above sea level, located in the Mahalangur Himal sub-range of the Himalayas. Wikipedia

Elevation: 8,848 m
First ascent: May 29, 1953
Prominence: 8,848 m
First ascenders: Edmund Hillary, Tenzing Norgay
Mountain range: Himalayas, Mahalangur Himal
Did you know: Mount Everest is the highest mountain in the world by elevation (8,848 m). wikipedia.org

Reviews
5,681 Google reviews

Except of the organic results of pages containing information about the elevation of Mount Everest you get an answer on the exact elevation, along with a box presenting a description about the Mount Everest. Both are powered by the Google Knowledge Graph.

4.2 The PhD Hub Knowledge Graph

A part of the activities of T2.5 of the PhD hub was to identify the classifications that exist, containing knowledge about Fields of Sciences and domain specific Scientific Classifications for each respective Science Field.

Semantically empowered search engines can exploit these vocabularies and improve search results by expanding terms at query or document indexing time. The scope of the use of the KG, is basically to classify PhD offers. A first option could be the ad-hoc classification of the offers, during their creation. When an offer author creates a new offer, he could choose to tag the offer with a set of selected classification instances of the PhD Hub KG. Then, these tags could be stored within the offer as a subjects field/property. Another option would be to analyze the title and the body of the offer using NLP techniques and then extract entities

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

(classifications) that lay within the text. The recognized entities can then be used to classify the offer and further annotate the text (using RDFa for instance). A hybrid approach would be to use both, with the auto-tagging feature acting as a suggestion provider to the author of the offer. With this use case in mind, some issues arise:

Issue 1: As shown on the previous section, there does not exist a global classification of scientific fields. The generic classifications are not extensive enough to cover specific needs. However we could integrate them to build a Core Knowledge Graph, covering the needs of the PhD Hub.

Issue 2: Scientific classifications are not regularly revised. For instance, the ACM's CCS was last revised back in 2012. Computer Science field is a highly dynamic research field with new areas of research emerging with exponential pace. Concepts like “Deep Learning” or “Blockchain” are not included on the CCS classification. Thus, we need the infrastructure to update the classifications with new terms.

Issue 3: Finally, a number of the classifications we identified to use, do not have a representation in SKOS, or their representation requires re-engineering, in order to match the PhD Hub needs. For instance, the CCS classification does not have dereferenceable IRIs, or fail to pass a number of quality measures[4]. Thus, we rebuilt the classification using LinkedPipes ETL.

We engineered a core KG for the PhD Hub, a mixture of the classifications we found, containing the classifications and relationships between them, with the option to extend the KG, according to the project's needs. The KG is build on the SKOS vocabulary, a common method used to represent hierarchical classifications on the Semantic Web.

In order to re-engineer the Classification we used LinkedPipes ETL, a tool that offers a sustainable approach of transforming and updating data on the SW. The procedure is further explained on the next section.

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

5 SKOSifying Scientific Classifications with LinkedPipes ETL

LinkedPipes ETL(LP-ETL) is integrated on the development server of PhD Hub. It runs on a Docker container as a part of the services that the server offers. LP-ETL does not have a complete and secure user management system. For this reason, it should not be accessed through a public endpoint. Only authorized users have access permissions. In order to access the GUI of LP-ETL the user has to connect through SSH on the server with the appropriate tunneling options. From a local *NIX system, follow these steps.

1. Open a terminal and edit the SSH config file:

```
nano ~/.ssh/config #or whatever text editor you prefer
```

2. Copy the following configuration at the end of the config file[6]:

```
Host phdhub
  LocalForward 8181 phdhub.math.auth.gr:8181
  LocalForward 2221 phdhub.math.auth.gr:2221
  LocalForward 2222 phdhub.math.auth.gr:2222
  LocalForward 2223 phdhub.math.auth.gr:2223
  LocalForward 2224 phdhub.math.auth.gr:2224
  LocalForward 2225 phdhub.math.auth.gr:2225
  LocalForward 2226 phdhub.math.auth.gr:2226
  LocalForward 2227 phdhub.math.auth.gr:2227
  LocalForward 2228 phdhub.math.auth.gr:2228
  LocalForward 2229 phdhub.math.auth.gr:2229
  LocalForward 2230 phdhub.math.auth.gr:2230
  LocalForward 8890 phdhub.math.auth.gr:8890
  HostName phdhub.math.auth.gr
  User {username} #replace {username} with your account name
  AddressFamily inet
```

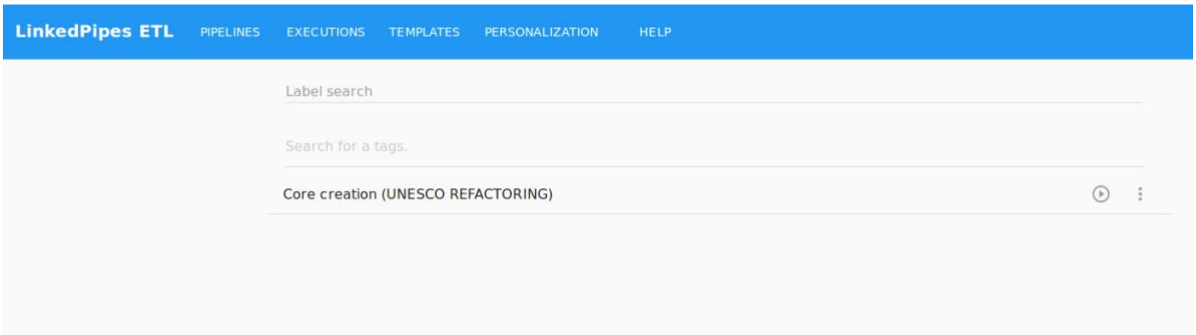
3. Save and exit from the editor, ie for nano that would be Ctrl + X
4. Connect to the server:

```
ssh phdhub
```

5. Provide your password in order to achieve a connection to the server. (you need an active account on the server)
6. Now you are connected through SSH on the server. The tunneling configuration will redirected the required local ports to the remote server. Now you can access the LP-ETL GUI with your browser.

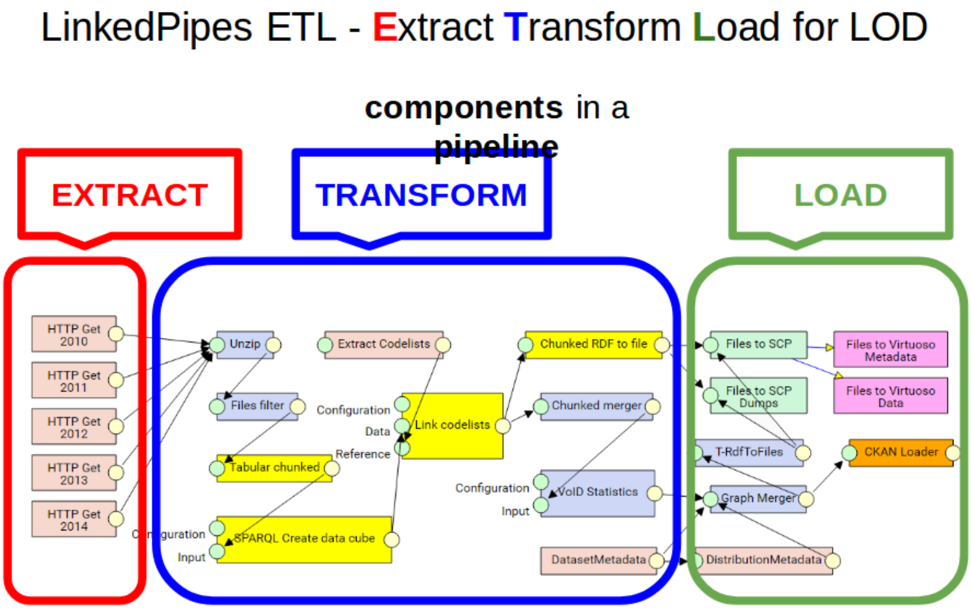
[6] If you have any of the referenced ports captured by another service on your local machine, please terminate the service before you connect to the server. Otherwise the tunneling process will not succeed and you will not be able to access the LP-ETL GUI

7. Open your browser and visit http://localhost:8181
8. If there are no errors at any step you will access the following page:



Core creation (UNESCO REFACTORING)

From here the user can get a list of the available pipelines or create, edit, delete and run a pipeline. A pipeline is a chainable procedure of single and autonomous components. There are three basic types of componentes, **E**xtractors, **T**ransformers and **L**oaders[7].

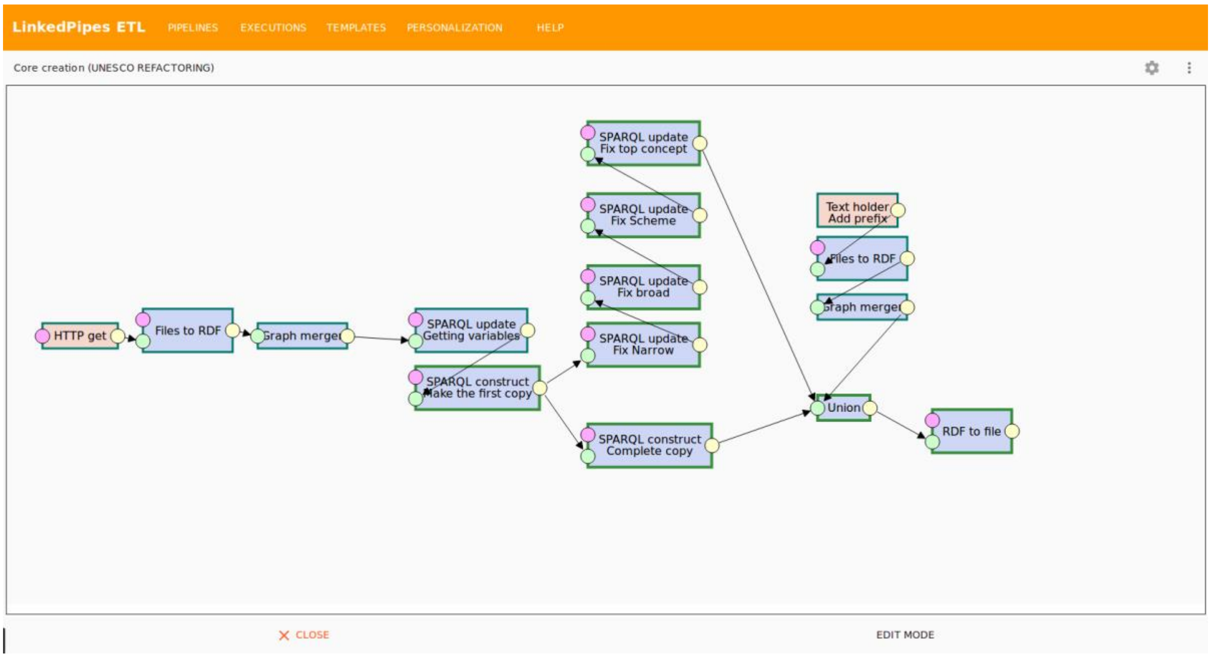


Each component can perform a discrete action. For instance, there is a component where you can retrieve a file from the web, or make an HTTP request to a remote API. There is a component that can parse tabular data from a CSV file and transform into RDF data format. Finally, there are components that can load the payload to a triplestore like Virtuoso or to an FTP server. Each component passes it's output to the next component on the chain, thus creating a pipeline. Once the pipeline is triggered, each component executes in the defined order. If there is any error on any of the components, you can access the debug logs and data

[7] https://docs.google.com/presentation/d/1UgEc2k2EuvHT9CPEtNKN1DUDW2iffaWtJ9PVEN56XMM/edit#slide=id.g28bc530bfd_0_1

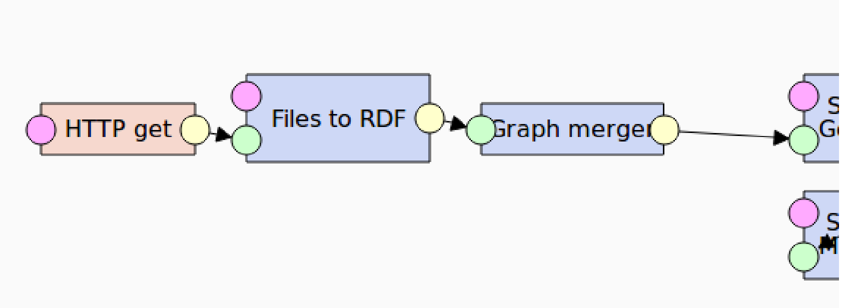
and fix any issues. For a more detailed description of components, features and HOWTOs you can refer to the homepage[8] of LinkedPipesETL.

Through the LP-ETL GUI, the user can see the various processing steps of the pipeline, in order to refactor a classification. On this example we are refactoring the UNESCO Field of Science Classification.



The pipeline starts on the left and concludes on the right side of the screen. Actually the order of the components is not defined by the way the pipeline is rendered. It could have any order visually but an organized pipeline as the above can be debugged more efficiently. The execution of each component is defined by the order the components are chained through the arrows.

The Extraction procedure is highlighted below.



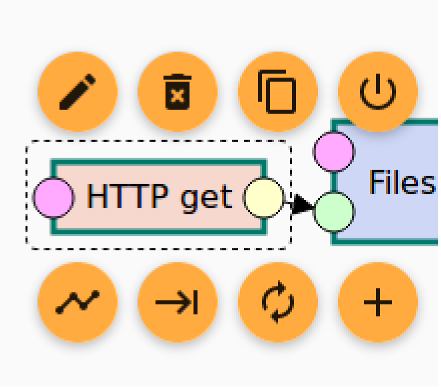
[8] https://etl.linkedpipes.com

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union

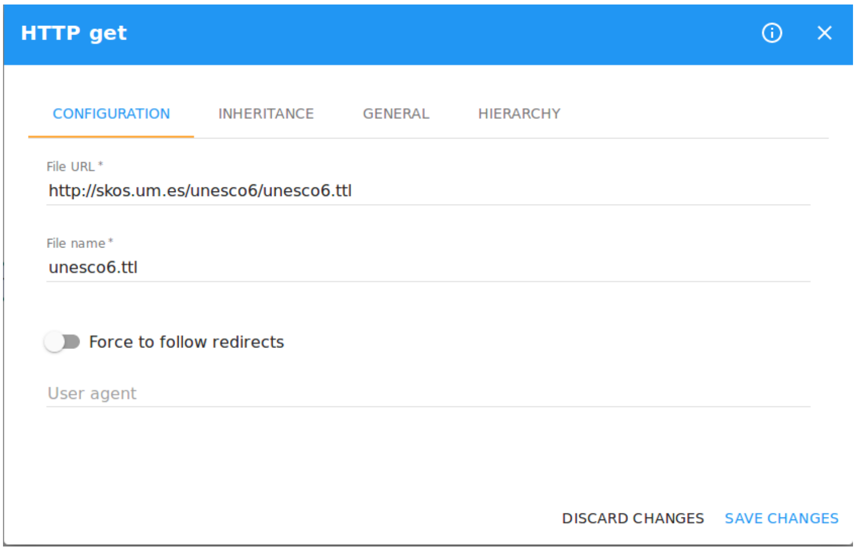
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

The first component retrieves the classification from the official repository of the Classification in http://skos.um.es/unesco6/downloads.php. The user can click on the "HTTP get" component to see more details of its functionality.

1. Click on the component
2. You will get the following



3. Click on the edit button, the one with the "pencil" icon to get the following menu.



File URL *
http://skos.um.es/unesco6/unesco6.ttl

File name *
unesco6.ttl

Force to follow redirects

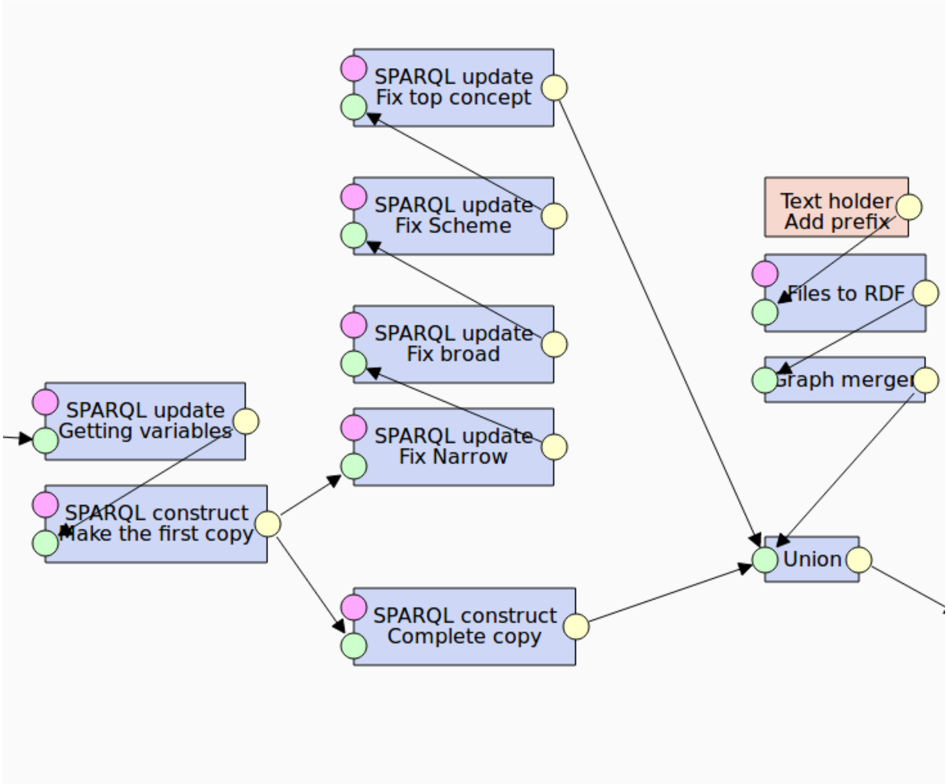
User agent

4. On the configuration tab the user can change the URL where the file to extract exists. Just edit the field "File URL"
5. On the "File name" field the user can define the name of the file that will be extracted. Be cautious when you have multiple "HTTP get" components on the same pipeline, if there are non unique filenames in all components the pipeline will fail.
6. The user has also two additional options. One boolean switch in order to turn on "Follow redirects" and the last one is the "User agent" header. Usually you will have to ignore both of them. For more details please check the LP-ETL documentation.
7. Finally the user can click on the "SAVE CHANGES" button and close this dialog.

The following two components parse the files in a local triplestore for further processing.

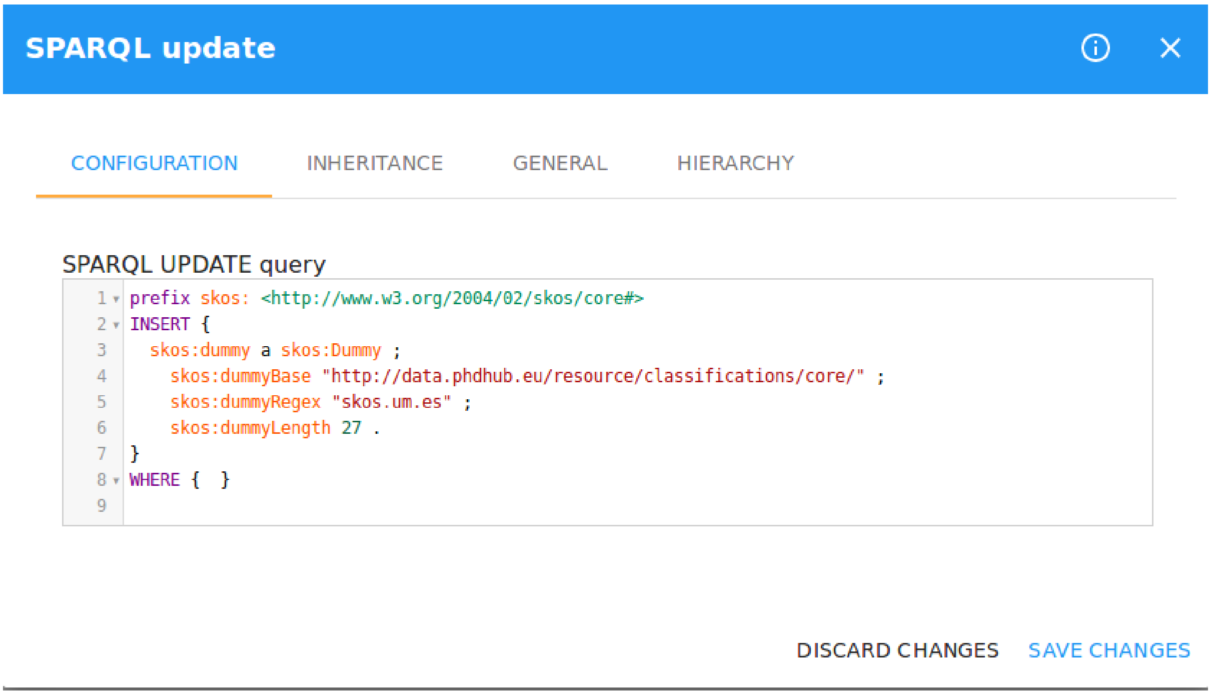
Next is the Transformation phase:

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220



The first two components components create an exact copy of each SPO triple pattern and rewrites the subjects URI with the pattern of the PhD Hub Core KG, which is http://data.phdhub.eu/resource/classifications/unesco_fos/{concept_notation}. The rest rewrite the OPS triples and add a relation to the original concept of the UNESCO FOS Classification. Each of these components executes a SPARQL Update or Construct query in order to produce or remove triples from the original graph. For instance the "Getting variables" component executes the following query.

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

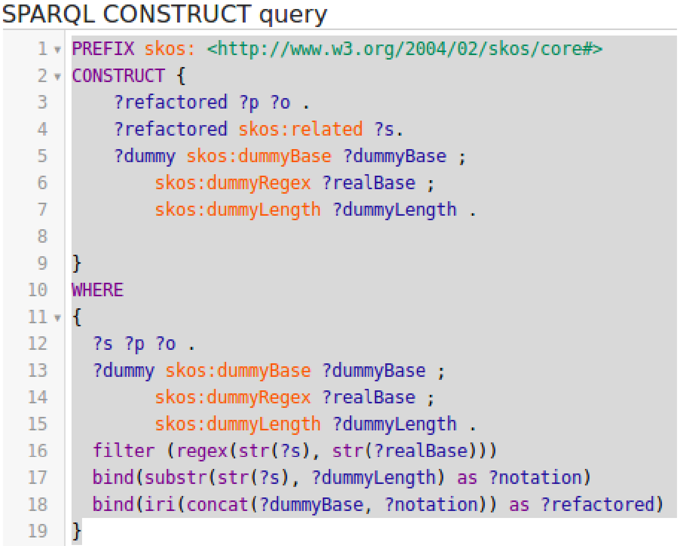


SPARQL UPDATE query

```
prefix skos: <http://www.w3.org/2004/02/skos/core#>
INSERT {
  skos:dummy a skos:Dummy ;
    skos:dummyBase "http://data.phdhub.eu/resource/classifications/core/" ;
    skos:dummyRegex "skos.um.es" ;
    skos:dummyLength 27 .
}
WHERE {  }
```

This query adds an instance of class skos:Dummy, a non-existent SKOS class, in order to pass parameters for the refactoring procedure on the graph. The “skos:dummyBase” parameter defines the URI which will be the base URI for all refactored concepts, the “skos:dummyRegex” parameter defines a REGEX rule, against which the refactoring is executed and the last parameter, the “skos:dummyLength” defines the length of the base URI to be refactored.

The next component executes the first phase copy of SPO triples. This is formulated as a SPARQL construct query shown below.



SPARQL CONSTRUCT query

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
CONSTRUCT {
    ?refactored ?p ?o .
    ?refactored skos:related ?s.
    ?dummy skos:dummyBase ?dummyBase ;
        skos:dummyRegex ?realBase ;
        skos:dummyLength ?dummyLength .

}
WHERE
{
  ?s ?p ?o .
  ?dummy skos:dummyBase ?dummyBase ;
        skos:dummyRegex ?realBase ;
        skos:dummyLength ?dummyLength .
  filter (regex(str(?s), str(?realBase)))
  bind(substr(str(?s), ?dummyLength) as ?notation)
  bind(iri(concat(?dummyBase, ?notation)) as ?refactored)
}
```

The next “Fix ...” components execute SPARQL Update queries on the graph, to create scheme and relation triples. The pipeline concludes on the “RDF to file” component, where you can serialize the produced graph into a file using a variety of RDF serializations or store it directly to the triplestore. You can always dump the file to a location on the server or through SCP, by using the respective component.

This is the procedure followed in order to modify the pipeline, and extract a file. The most common case is that the user have to copy the original pipeline, rename it, and change a number of parameters as shown above.

With the support of Erasmus+

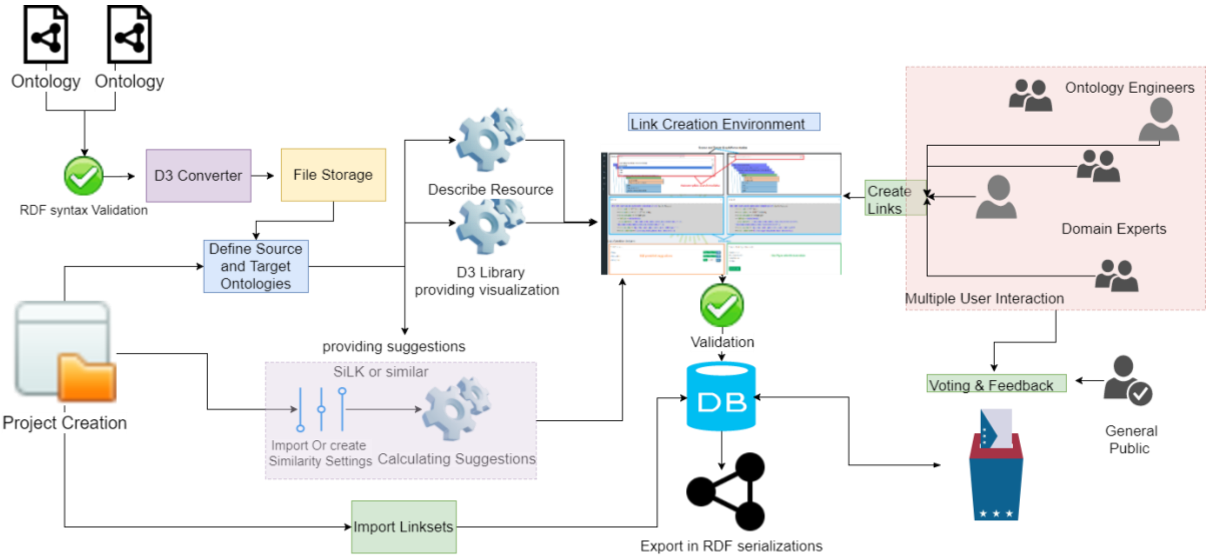
Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

6 Creating Links Manually with Alignment

Alignment is a collaborative, system-aided, user-driven ontology matching platform. It offers a simple GUI environment for matching two ontologies based on a default or user defined configuration of similarity measures and algorithms. Users can select one of the suggested links for each entity, or they can choose any other link to the target ontology, based on their domain knowledge.

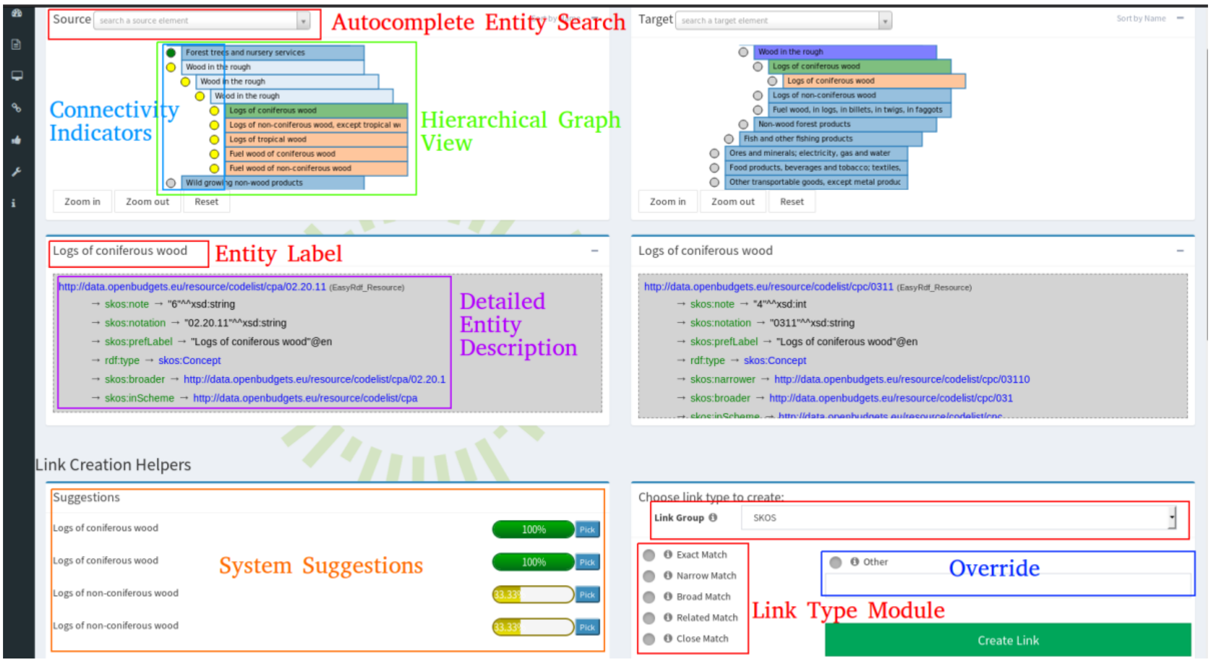
Users can also customize the similarity variables that will be used for the comparison of the two ontologies and result in the suggested links, based on their preferences. Multiple users can work on the same project and provide their own links simultaneously and interactively. The platform also offers evaluation and social features, as users can cast a positive or negative vote or comment on a specific link between two entities, providing feedback on the produced linksets. The produced linksets are then available through both a SPARQL endpoint and an API. A typical workflow of a use case is shown below.



Alignment Workflow

A user has to create a project within the platform. First, it is needed to upload the ontologies he wants to produce a linkset. The ontologies get validated and stored on the platform. Then, the user has to define the source and target ontology. Also he needs to define which similarity algorithm configurations will be used for the system provided suggestions. The user can also choose if the project will be private or public. Then, upon the creation of the project, the platform calculates the similarities between the entities of the ontologies and renders the GUI. None of the suggestions provided by the system is realized as a valid link, unless some user decides to actually create the link. Finally, the produced linksets can be exported, or sent for crowd-sourced validation, through the Voting service.

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220


Source
search a source element
Autocomplete Entity Search
Target
search a target element
Sort by Name
Connectivity Indicators
Forest trees and nursery services
Wood in the rough
Wood in the rough
Wood in the rough
Logs of coniferous wood
Logs of non-coniferous wood, except tropical w
Logs of tropical wood
Fuel wood of coniferous wood
Fuel wood of non-coniferous wood
Wild growing non-wood products
Hierarchical Graph View
Wood in the rough
Logs of coniferous wood
Logs of coniferous wood
Logs of non-coniferous wood
Fuel wood, in logs, in billets, in twigs, in faggots
Non-wood forest products
Fish and other fishing products
Ores and minerals; electricity, gas and water
Food products, beverages and tobacco; textiles,
Other transportable goods, except metal produc
Zoom in
Zoom out
Reset
Zoom in
Zoom out
Reset
Entity Label
Detailed Entity Description
Link Creation Helpers
Suggestions
System Suggestions
100%
Pick
100%
Pick
33.33
Pick
33.33
Pick
Choose link type to create:
Link Group
Override
Link Type Module
Create Link

Logs of coniferous wood

http://data.openbudgets.eu/resource/codelist/cpa/02.20.11 (EasyRdf_Resource)
→ skos:note → "6"^^xsd:string
→ skos:notation → "02.20.11"^^xsd:string
→ skos:prefLabel → "Logs of coniferous wood"@en
→ rdf:type → skos:Concept
→ skos:broader → http://data.openbudgets.eu/resource/codelist/cpa/02.20.1
→ skos:inScheme → http://data.openbudgets.eu/resource/codelist/cpa

Logs of coniferous wood

http://data.openbudgets.eu/resource/codelist/cpc/0311 (EasyRdf_Resource)
→ skos:note → "4"^^xsd:int
→ skos:notation → "0311"^^xsd:string
→ skos:prefLabel → "Logs of coniferous wood"@en
→ rdf:type → skos:Concept
→ skos:narrower → http://data.openbudgets.eu/resource/codelist/cpc/03110
→ skos:broader → http://data.openbudgets.eu/resource/codelist/cpc/031
→ skos:inScheme → http://data.openbudgets.eu/resource/codelist/cpc

Suggestions:
Logs of coniferous wood 100%
Logs of coniferous wood 100%
Logs of non-coniferous wood 33.33
Logs of non-coniferous wood 33.33

Link Group: SKOS

○ Exact Match
○ Narrow Match
○ Broad Match
○ Related Match
○ Close Match
○ Other

Alignment GUI

With the support of Erasmus+

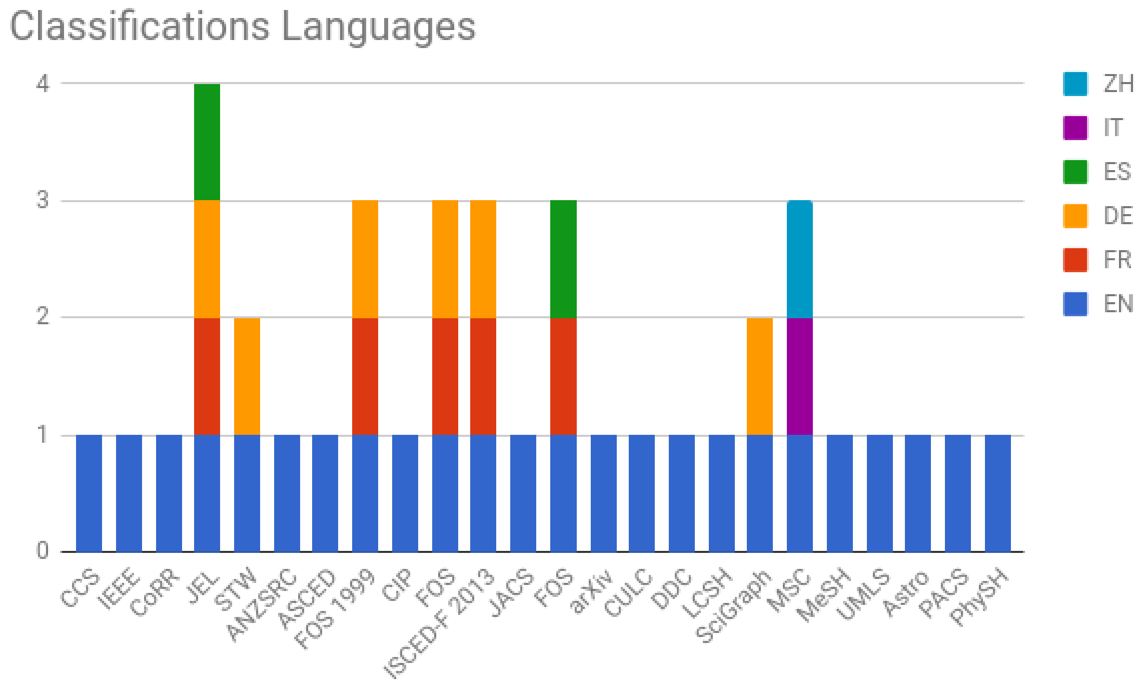
Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

7 Editing a KG with VocBench

According to the results shown in Section 3, KGs need to be maintained and updated in regular intervals. Within the scope of PhD Hub we use well-established Science classification schemes. Some of them are actively maintained by their respective community (such as MeSH), however there is a number of classifications that are not regularly updated, most notably the ACM CCS classification. With the rate of evolution of research fields in a highly dynamic field as is Computer Science, there is the issue of not covering the most recent trends. Thus, there is a need to have a service where the KGs will be maintained and updated according to the needs of PhD Hub.

Additionally, since the target audience of PhD Hub is within the EU member states, multilingual issues emerge. The EU has 24 official languages[9], however the majority of the existing Classification schemes use only the English language, and availability of translations in other languages is rather limited.



An overview of available languages on classifications.

Thus, we need a sustainable method and a user friendly environment, in order to maintain the KGs and offer an option to translate the labels of concepts. To this end, we integrated VocBench as a service to the Knowledge Repository of PhD Hub. VocBench is a web-based, multilingual, collaborative development platform for managing OWL ontologies, SKOS(XL) thesauri and generic RDF datasets. It is designed to meet the needs of semantic web and linked data environments. VocBench development has also been driven by the feedback

[9] https://en.wikipedia.org/wiki/Languages_of_the_European_Union

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

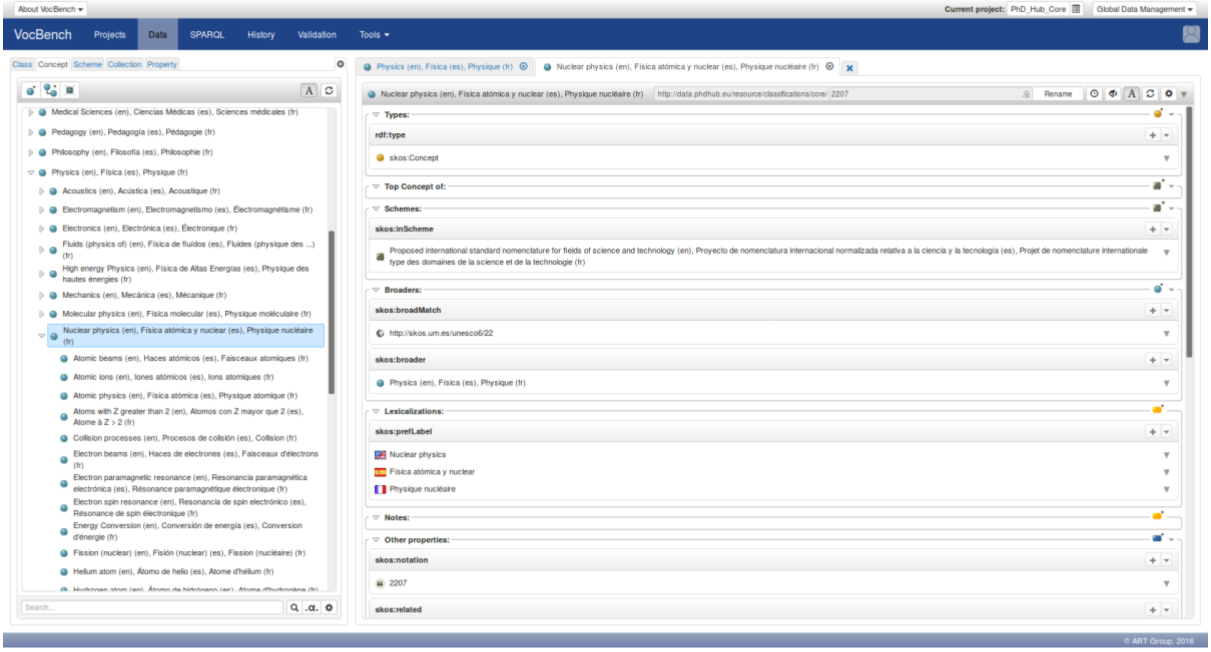
gathered from a community of users made of public organizations, companies and independent users looking for open source solutions for maintaining their ontologies, thesauri, code lists and authority resources.

VocBench business and data access layers are realized by Semantic Turkey, an open-source platform for Knowledge Acquisition and Management realized by the ART Research Group at the University of Rome Tor Vergata.

Funded by the European Commission ISA² programme, the development of Vocbench 3 (VB3) is managed by the Publications Office of the EU under the contract 10632 (Infeurope S.A.).

It is widely used in the SW community and most notably for the maintenance of EUROVOC and AGROVOC.

In order to access the environment of VocBench you have to connect to the Knowledge Repository server as described in Section 5. You can find a extensive documentation of the usage and administration of VocBench on the official website of the project



Current project: PhD_Hub_Core
Global Data Management

- Medical Sciences (en), Ciencias Médicas (es), Sciences médicales (fr)
- Pedagogy (en), Pedagogía (es), Pédagogie (fr)
- Philosophy (en), Filosofía (es), Philosophie (fr)
- Physics (en), Física (es), Physique (fr)
  - Acoustics (en), Acústica (es), Acoustique (fr)
  - Electromagnetism (en), Electromagnetismo (es), Électromagnétisme (fr)
  - Electronics (en), Electrónica (es), Électronique (fr)
  - Fluids (physics of) (en), Física de fluídos (es), Fluides (physique des ...) (fr)
  - High energy Physics (en), Física de Altas Energías (es), Physique des hautes énergies (fr)
  - Mechanics (en), Mecánica (es), Mécanique (fr)
  - Molecular physics (en), Física molecular (es), Physique moléculaire (fr)
  - Nuclear physics (en), Física atómica y nuclear (es), Physique nucléaire (fr)
    - Atomic beams (en), Haces atómicos (es), Faisceaux atomiques (fr)
    - Atomic ions (en), Iones atómicos (es), Ions atomiques (fr)
    - Atomic physics (en), Física atómica (es), Physique atomique (fr)
    - Atoms with Z greater than 2 (en), Atomos con Z mayor que 2 (es), Atome à Z > 2 (fr)
    - Collision processes (en), Procesos de colisión (es), Collision (fr)
    - Electron beams (en), Haces de electrones (es), Faisceaux d'électrons (fr)
    - Electron paramagnetic resonance (en), Resonancia paramagnética electrónica (es), Résonance paramagnétique électronique (fr)
    - Electron spin resonance (en), Resonancia de spin electrónico (es), Résonance de spin électronique (fr)
    - Energy Conversion (en), Conversión de energía (es), Conversion d'énergie (fr)
    - Fission (nuclear) (en), Fisión (nuclear) (es), Fission (nucléaire) (fr)
    - Helium atom (en), Átomo de helio (es), Atome d'hélium (fr)
    - Hydrogen atom (en), Átomo de hidrógeno (es), Atome d'hydrogène (fr)

Physics (en), Física (es), Physique (fr) | Nuclear physics (en), Física atómica y nuclear (es), Physique nucléaire (fr)

Nuclear physics (en), Física atómica y nuclear (es), Physique nucléaire (fr) http://data.phdhub.eu/resource/classifications/core/ 2207

Types:
rdf:type
skos:Concept

Top Concept of:

Schemes:
skos:inScheme
Proposed international standard nomenclature for fields of science and technology (en), Proyecto de nomenclatura internacional normalizada relativa a la ciencia y la tecnología (es), Projet de nomenclature internationale type des domaines de la science et de la technologie (fr)

Broaders:
skos:broadMatch
http://skos.um.es/unesco6/22
skos:broader
Physics (en), Física (es), Physique (fr)

Lexicalizations:
skos:prefLabel
Nuclear physics
Física atómica y nuclear
Physique nucléaire

Notes:

Other properties:
skos:notation
2207
skos:related

8 Use Cases

Although the potential use cases of a Knowledge Graph within the scope of the PhD Hub project might be numerous, we will highlight a very interesting use case from the field of Semantic Annotation for use in Information Retrieval.

Consider the following PhD position posted here from the University of Luxemburg. The offer contains a body of text, describing the research area of the PhD position, the qualifications a candidate should or would be desirable to have etc. You can find the sample below.

```
PhD Candidate in Automated Document Processing

University of Luxembourg (UL) is offering a fully funded PhD
student position in Automated Software Engineering to start as
soon as possible.

 Organisation

The University of Luxembourg seeks to hire an outstanding PhD
Candidate at its Interdisciplinary Centre for Security,
Reliability and Trust (SnT). SnT is a recently formed centre
carrying out interdisciplinary research in secure, reliable and
trustworthy ICT (Information and Communication Technologies)
systems and services, often in collaboration with industrial,
governmental or international partners.  SnT is active in several
international research projects funded by the EU framework
programme and the European Space Agency. For further information
you may check: www.securityandtrust.lu

PhD Candidate in Automated Document Processing (m/f)
Ref.: R-STR-5014-00-B
Fixed Term Contract up to 3 years, pending satisfaction of
progress milestones (CDD), full-time (40 hrs/week)
Number of positions: 1
Your Role
Under the direction of a PhD Supervisor, you will carry out
research activities and write a thesis, the main goal being to
obtain a PhD in the area of Automated Software Engineering. You
may be tasked with conducting literature surveys and establishing
state-of-the-art; developing necessary experimental and simulation
facilities where required; planning, executing, and analysing
experiments and simulations; conducting joint and independent
```

research activities; contributing to project deliverables, milestones, demonstrations, and meetings; disseminating results at international scientific conferences/workshops and peer reviewed scientific publications.

In addition, you will take part in a project in collaboration with a Luxembourgish FinTech company specialized in managing financial flows. One of the main goals of the project is to develop new ICT methods enabling better automation in document classification and document processing.

Your Profile

A master's degree in Computer Science.

A proven interest in Software Engineering or Artificial Intelligence is required.

Commitment, team working, a critical mind, and motivation are skills that are more than welcome.

Fluent written and verbal communication skills in English are mandatory.

Programming Skills (e.g. Java and Python),

Expertise in at least one of these topics: information retrieval, natural language processing, text analysis, clustering, document processing.

We offer

The University offers a Ph.D. study program with a Fixed Term Contract up to 3 years in total, pending satisfaction of progress milestones (CDD), on full time basis (40hrs/week). The University offers highly competitive salaries and is an equal opportunity employer. You will work in an exciting international environment and will have the opportunity to participate in the development of a newly created research centre.

Further Information

Applications, written in English should be submitted online and should include:

Curriculum Vitae (including your contact address, work experience, publications)

Cover letter indicating the research area of interest and your motivation

With the support of Erasmus+

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

Transcript of all courses and results from the highest university-level courses taken

A short description of your Master's work (max 1 page)

If possible, contact information for 3 referees

Deadline for applications: 30 June 2018

Applications will be considered on receipt therefore applying before the deadline is encouraged.

Further Information

Research Context

This PhD thesis will be done in collaboration with a Luxembourgish company which processes a lot of documents. The PhD thesis will extensively explore techniques aiming at (1) automate the processing of these documents, (2) extract data and knowledge from these documents, (3) compute smart "diff" of documents (syntactic and semantics), (4) continuously learn from past activities.

Topics

The proposed thesis topic focuses on automated techniques for automated document classification, smart document processing and ability to extract value-added intelligence from data stored (and being constantly updated) in databases.

As there is no off-the-shelf solution that may provide automation in the specific context of this PhD thesis, the above challenges will be addressed with a focused R&D approach requiring research-oriented investigations, context adaptation and prototyping.

These objectives are challenging from a research viewpoint since it requires to devise novel approaches combining Information Retrieval (including Natural Language Processing) and smart approaches (including learning and analysis) to express and translate into operations the know-how and current practice of experts.

The team you will be working with:

With the support of Erasmus+

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

Jacques Klein: Primary advisor
Tegawendé F. Bissyandé: Co-advisor

In order to efficiently classify the offer, we could annotate it with an entity from a predefined Classification. Thus, an efficient search engine could utilize the relationships within the classification and expand query terms given by users. One option could be to let the author of the offer to choose a set of entities from a general classification, such as LCSH, or a domain specific classification (ACM CCS), to annotate the offer, or populate a "Subject" field. Another option could be to automatically annotate the offer, using a Semantic Enhancer Engine (SEE). A hybrid approach would be to provide suggestions to the author.

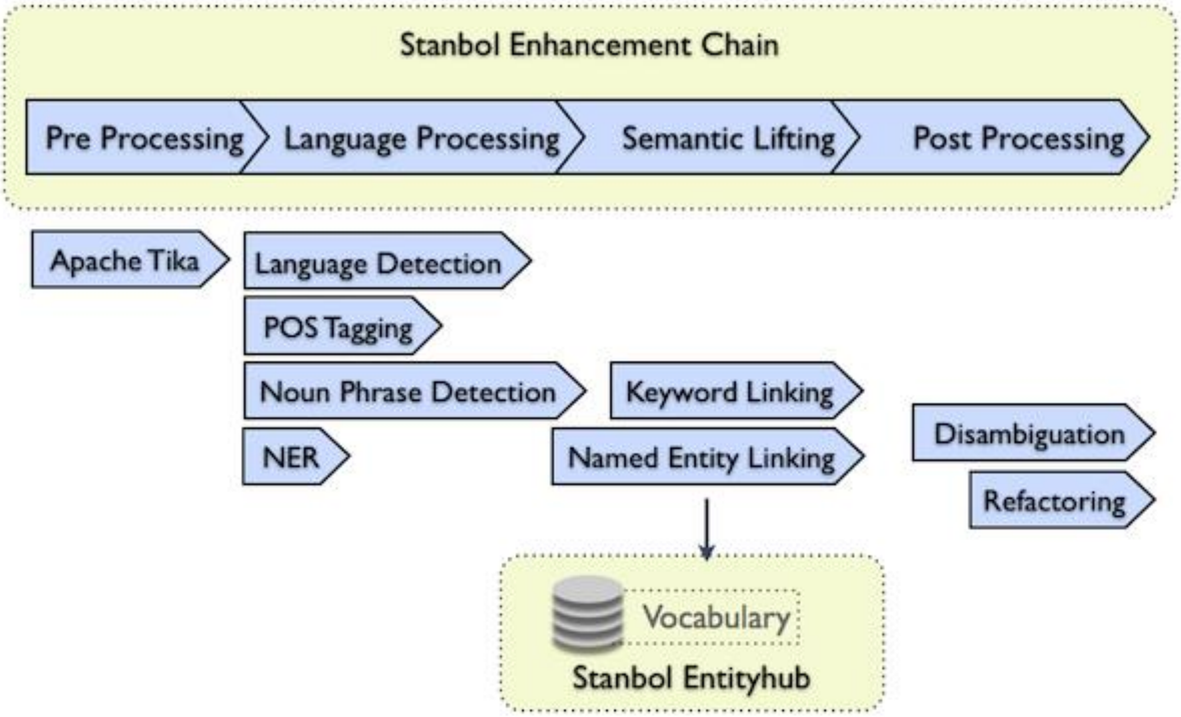
For the role of the SEE, we tested Apache Stanbol. Apache Stanbol[10] is an "open source modular software stack and reusable set of components for semantic content management. Apache Stanbol components are meant to be accessed over RESTful interfaces to provide semantic services for content management. Thus, one application is to extend traditional content management systems with (internal or external) semantic services.

Additionally, Apache Stanbol lets you create new types of content management systems with semantics at their core. The current code is written in Java and based on the OSGi component framework. Applications include extending existing content management systems with (internal or external) semantic services, and creating new types of content management systems with semantics at their core".

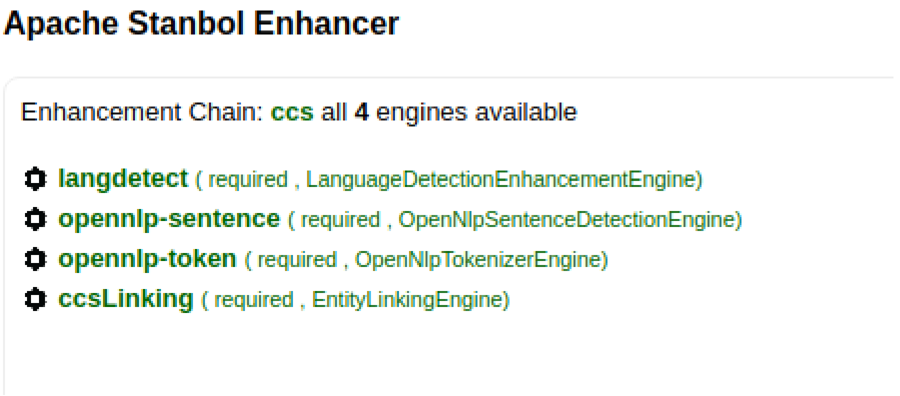
In this use case, we are using the Enhancer[11] and the EntityHub components of Apache Stanbol. We have preloaded the ACM CCS classification, a Computer Science specific classification on the EntityHub Component of Apache Stanbol. Then we created an enhancement chain. The chain executes the steps shown in the following figure:

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

[11] https://stanbol.apache.org/docs/trunk/customvocabulary.html



The CCS classification is entailed on the Keyword Linking Steps. The administrator of the Apache Stanbol can configure precisely the steps of the enhancer. You can add a numerous classifications to replace or act in parallel with the CCS classification. For the case of CCS the chain is the following:



Apache Stanbol Enhancer

Enhancement Chain: **ccs** all **4** engines available

- **langdetect** (required , LanguageDetectionEnhancementEngine)
- **opennlp-sentence** (required , OpenNlpSentenceDetectionEngine)
- **opennlp-token** (required , OpenNlpTokenizerEngine)
- **ccsLinking** (required , EntityLinkingEngine)

The first step is to detect the language of the text, followed by the text segmentation into sentences and tokens using the OpenNLP library. The last step is to perform the actual Entity Linking with the CCS Classification, by performing string similarity measures.

The chain can be triggered through the REST API of the Apache Stanbol Enhancer. For this demonstration, we use the Web View of the Enhancer, which is a web form where you can post a body of text. Once the chain is executed, you can get the results in various formats. Below, we requested the results in JSON format, a format easy to digest from common web applications. A sample of the output is shown:

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

```
{
  "@id" : "http://relative-uri.fake/#10003317",
  "http://stanbol.apache.org/ontology/entityhub/query#score" : [ {
    "@value" : "5.2062407",
    "@type" : "http://www.w3.org/2001/XMLSchema#float"
  } ],
  "@type" : [ "http://www.w3.org/2004/02/skos/core#Concept" ],
  "http://www.w3.org/2000/01/rdf-schema#label" : [ {
    "@value" : "Information retrieval",
    "@language" : "en"
  }, {
    "@value" : "ir",
    "@language" : "en"
  }, {
    "@value" : "search",
    "@language" : "en"
  }, {
    "@value" : "searching",
    "@language" : "en"
  } ]
}, {
  "@id" : "http://relative-uri.fake/#10003356",
  "http://stanbol.apache.org/ontology/entityhub/query#score" : [ {
    "@value" : "4.994456",
    "@type" : "http://www.w3.org/2001/XMLSchema#float"
  } ],
  "@type" : [ "http://www.w3.org/2004/02/skos/core#Concept" ],
  "http://www.w3.org/2000/01/rdf-schema#label" : [ {
    "@value" : "Clustering and classification",
    "@language" : "en"
  }, {
    "@value" : "classification and clustering",
    "@language" : "en"
  }, {
    "@value" : "document classification",
    "@language" : "en"
  }, {
    "@value" : "document clustering",
    "@language" : "en"
  }, {
    "@value" : "information classification",
    "@language" : "en"
  }, {
    "@value" : "information clustering",
    "@language" : "en"
  }, {
    "@value" : "text classification",
    "@language" : "en"
  }, {
    "@value" : "text clustering",
    "@language" : "en"
  } ]
}, {
  "@id" : "http://relative-uri.fake/#10003444",
  "http://stanbol.apache.org/ontology/entityhub/query#score" : [ {
    "@value" : "1.0729058",
    "@type" : "http://www.w3.org/2001/XMLSchema#float"
  } ],
  "@type" : [ "http://www.w3.org/2004/02/skos/core#Concept" ],
  "http://www.w3.org/2000/01/rdf-schema#label" : [ {
    "@value" : "Clustering",
    "@language" : "en"
  } ]
}, {
  "@id" : "http://relative-uri.fake/#10003569",
  "http://stanbol.apache.org/ontology/entityhub/query#score" : [ {
    "@value" : "0.4476598",
    "@type" : "http://www.w3.org/2001/XMLSchema#float"
  } ],
  "@type" : [ "http://www.w3.org/2004/02/skos/core#Concept" ],
  "http://www.w3.org/2000/01/rdf-schema#label" : [ {
    "@value" : "Automation",
    "@language" : "en"
  }, {
    "@value" : "automated",
    "@language" : "en"
  }, {
    "@value" : "automated processes",
    "@language" : "en"
  }, {
```

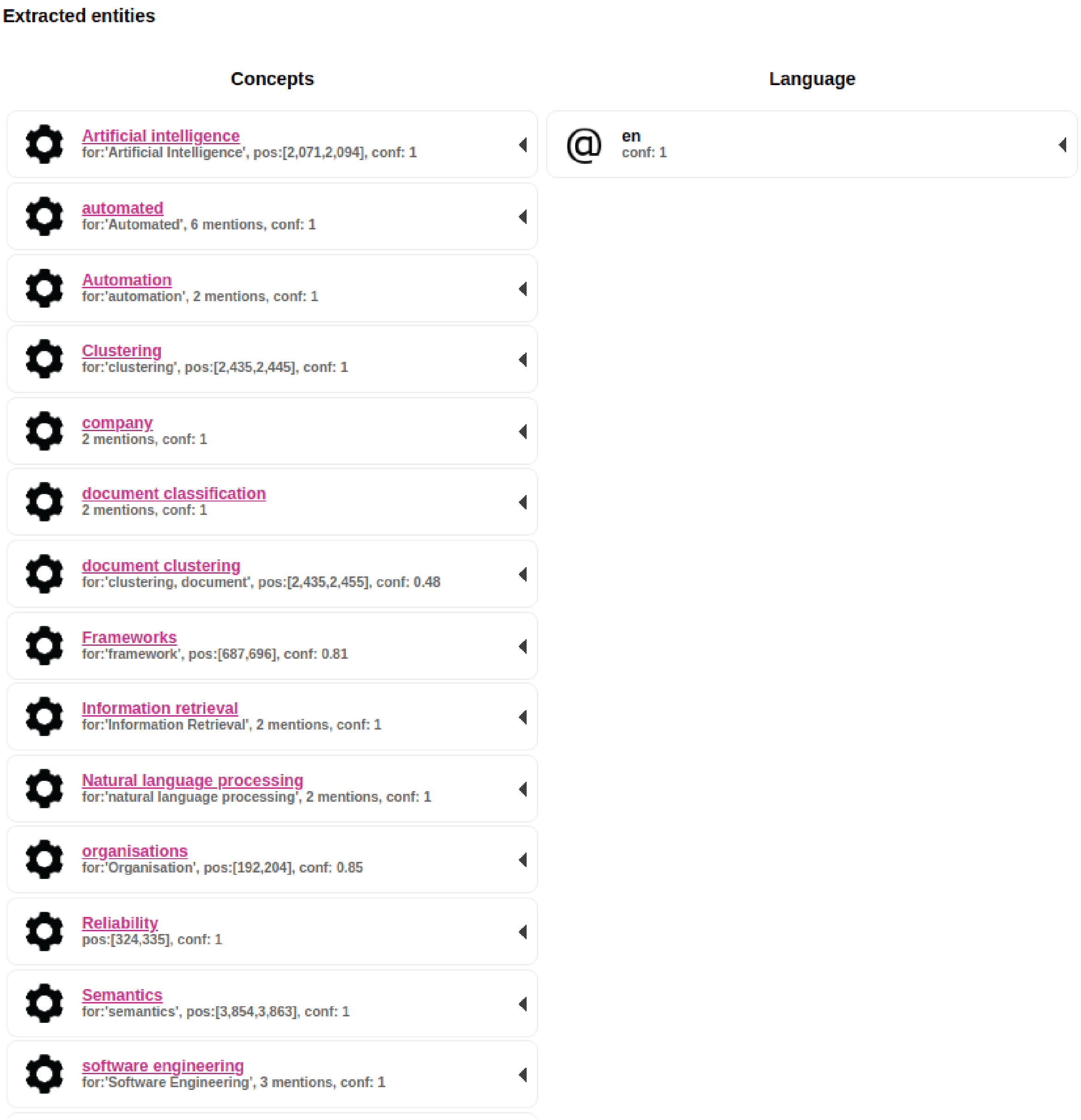
With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

```
    "@value" : "automated systems",
    "@language" : "en"
  }, {
    "@value" : "automating",
    "@language" : "en"
  }, {
    "@value" : "business automation",
    "@language" : "en"
  }, {
    "@value" : "office automation",
    "@language" : "en"
  } ]
}, {
  "@id" : "http://relative-uri.fake/#10010178",
  "http://stanbol.apache.org/ontology/entityhub/query#score" : [ {
    "@value" : "8.335619",
    "@type" : "http://www.w3.org/2001/XMLSchema#float"
  } ],
  "@type" : [ "http://www.w3.org/2004/02/skos/core#Concept" ],
  "http://www.w3.org/2000/01/rdf-schema#label" : [ {
    "@value" : "Artificial intelligence",
    "@language" : "en"
  }, {
    "@value" : "ai",
    "@language" : "en"
  } ]
},
```

As we can see, the SEE recognised a number of entities of the CCS Classification within the text, and linked them with the respective entities. These entities can be used to classify the offer and enhance search results. For instance, even if the term “text classification” is not directly referred to the body of text, a user query with this term will retrieve the offer.

With the support of Erasmus+
Co-funded by the Erasmus+ Programme of the European Union
This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220



Extracted entities

Concepts

- Artificial intelligence
  for:'Artificial Intelligence', pos:[2,071,2,094], conf: 1
- automated
  for:'Automated', 6 mentions, conf: 1
- Automation
  for:'automation', 2 mentions, conf: 1
- Clustering
  for:'clustering', pos:[2,435,2,445], conf: 1
- company
  2 mentions, conf: 1
- document classification
  2 mentions, conf: 1
- document clustering
  for:'clustering, document', pos:[2,435,2,455], conf: 0.48
- Frameworks
  for:'framework', pos:[687,696], conf: 0.81
- Information retrieval
  for:'Information Retrieval', 2 mentions, conf: 1
- Natural language processing
  for:'natural language processing', 2 mentions, conf: 1
- organisations
  for:'Organisation', pos:[192,204], conf: 0.85
- Reliability
  pos:[324,335], conf: 1
- Semantics
  for:'semantics', pos:[3,854,3,863], conf: 1
- software engineering
  for:'Software Engineering', 3 mentions, conf: 1

Language

- en
  conf: 1

With the support of Erasmus+

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220

Bibliography

[1] D. Allemang and J. Hendler, “SKOS—managing vocabularies with RDFS-Plus,” *Semantic Web for the Working Ontologist*, pp. 207–219, 2011.

[2] B. Haslhofer, F. Martins, and J. Magalhães, “Using SKOS vocabularies for improving web search,” *Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion*, 2013.

[3] “SKOS Simple Knowledge Organization System Primer,” *Extensible Markup Language (XML) 1.0 (Fifth Edition)*. [Online]. Available: https://www.w3.org/TR/skos-primer/.

[4] O. Suominen and E. Hyvönen, “Improving the Quality of SKOS Vocabularies with Skosify,” *Lecture Notes in Computer Science Knowledge Engineering and Knowledge Management*, pp. 383–397, 2012.

With the support of Erasmus+

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with the support from the European Commission. The document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Project number: 588220