Ontology Extraction and Usage in the Scholarly Knowledge Domain*

Angelo A. Salatino[0000−0002−4763−3943], Francesco Osborne[0000−0001−6557−3131], and Enrico Motta[0000−0003−0015−1952]

Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom
{angelo.salatino,francesco.osborne,enrico.motta}@open.ac.uk

Abstract. Ontologies of research areas have been proven to be useful resources for analysing and making sense of scholarly data. In this chapter, we present the Computer Science Ontology (CSO), which is the largest ontology of research areas in the field, and discuss a number of applications that build on CSO to support high-level tasks, such as topic classification, metadata extraction, and recommendation of books.

Keywords: Scholarly Data · Ontology Learning · Bibliographic Data · Scholarly Ontologies.

1 Introduction

Ontologies, as formal specifications of concepts and relations in specific domains, have become a standard solution to represent domain knowledge, integrate data from different sources, and support a variety of semantic applications [7, 10]. In the field of scholarly knowledge, ontologies are used to facilitate the integration of large datasets of research data [11], the exploration of the academic landscape [18, 9], information extraction from scientific articles [8], and so on. More specifically, ontologies representing research topics and describing their relationships, have been employed in several tasks, such as making sense of research dynamics [18], classifying research publications [25], characterising [2] and identifying [21] research communities, studying the origin of research topics [23], and forecasting research trends [24]. However, not every domain of science has an ontology that comprehensively describes all research concepts and their relations. In addition, ontologies describing research topics are typically manually crafted by domain experts which is a very time consuming process. Therefore, they usually evolve slow and become quickly outdated. A further issue is that many of these ontologies tend to be very coarse-grained, lacking the right depth that would allow them to comprehensively describe the area. All these limitations hinder the adoption of semantic technologies in several fields of science.

The question we are left with is: how do we effectively extract comprehensive, fine-grained, and up-to-date ontologies of research topics that can support a range of intelligent services?

* We would like to thank Springer Nature for partially funding this research.

In this chapter, we present the Computer Science Ontology (CSO) Framework [27, 26], which is a conceptual framework for generating a large scale ontology of Computer Science. This solution has been used to support a variety of high-level tasks, such as (i) categorising proceedings in digital libraries, (ii) enhancing semantically the metadata of scientific publications, (iii) generating recommendations, (iv) producing smart analytics, (v) detecting research trends, and others [20, 27]. Figure 1 shows the architecture of the framework, where each layer exploits the underneath layers.

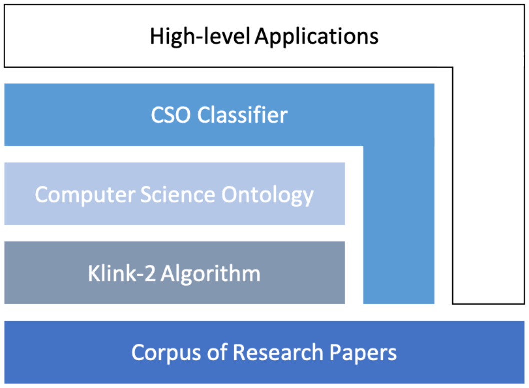


Fig. 1. The Computer Science Ontology Framework.

The first layer is the corpus of research papers. This represent the core input data from which we want to extract a granular and data-driven ontology of research areas.

Sitting on top of the data layer, we find the Klink-2 algorithm [17] that generates the large-scale ontology of research topics from publication metadata. The Computer Science Ontology (CSO), which represents the third layer, is a large-scale, granular, and automatically generated ontology of research areas. The current version includes about 14K topics and 163K semantic relationships. At the fourth layer, we find the CSO Classifier [25], a tool for automatically classifying research papers according to the topics available in the Computer Science Ontology. This system enables users to represent scientific publications in terms of CSO concepts and allows all relevant stakeholders to develop a multitude of relevant smart functionalities.

The rest of the chapter is organised as follows. In Section 2, we review the literature regarding topic detection in research papers, pointing out the existing gap. In Section 3, we describe the Klink-2 algorithm for generating the ontology, and in Sections 4 we discuss CSO with its data model. In Section 5 we describe the CSO Classifier and in Section 6 we show in more detail some high-level applications that allowed us to perform smart analytics on scholarly data. Finally, in Section 7 we summarise the main conclusions and outline future directions of research.

2 Literature Review

Some fields of research are comprehensively described by ontologies of research areas, e.g., MeSH in Biology and PhySH in Physics. These ontologies can provide support toward a multitude of tasks, such as integrating heterogeneous datasets [11], assisting users in exploring digital libraries [18], producing scholarly analytics [3], and forecasting research dynamics [24, 20]. In this section, we will review the current state of the art with regards to developing and using ontologies of research areas. In particular, we will first provide an overview of some of most well-known ontologies of research areas, then we will discuss current approaches for the generation of these ontologies, and finally, we will describe the approaches that take advantage of such ontologies to perform several high-level tasks.

2.1 Ontologies of research areas

In literature, we can find different ontologies of research areas, which are scoped to a particular branch of science. In the field of Computer Science, the most well-known taxonomy is the ACM Computing Classification System[1], developed and maintained by the Association for Computing Machinery (ACM). This taxonomy contains around 2K concepts and it is manually curated. These characteristics represent a limitation as its representation of the field lacks both depth and breadth, and its curation process is slow and expensive.

In the field of Medicine, the most popular solution is the Medical Subject Heading (MeSH)[2] maintained by the National Library of Medicine of the United States. This taxonomy is constantly updated by collecting new terms as they appear in the scientific literature.

The Physics Subject Headings (PhySH)[3] is the standard solution in the field of Physics and Astronomy. It is developed by the American Physical Society (APS) and it is constantly updated with the support of authors, reviewers, editors and organisers of scientific conferences.

In the field of Mathematics there is the Mathematics Subject Classification (MSC)[4] which is maintained by Mathematical Reviews and zbMATH. In the field of Economics we can find the JEL classification[5], created by the Journal of Economic Literature of the American Economic Association, and the STW Thesaurus for Economics[6] developed by ZBW - Leibniz Information Centre for Economics.

[1] ACM Computing Classif. System - https://www.acm.org/publications/class-2012

[2] Medical Subject Heading - https://www.ncbi.nlm.nih.gov/mesh

[3] Physics Subject Headings - https://physh.aps.org/

[4] Mathematics Subject Classification - https://mathscinet.ams.org/msc

[5] Journal of Economic Literature - https://www.aeaweb.org/econlit/jelCodes.php

[6] STW Thesaurus for Economics - http://zbw.eu/stw/

The ontologies mentioned above can comprehensively represent specific areas of science. However, in literature we can also find more broad ontologies covering a multitude of fields. The most popular ones are the Library of Congress Classification and the Dewey Decimal Classification which encompass many areas of science. Indeed, these two schemes are employed to classify books within large academic libraries, globally.

Another ontology of research areas in the Fields of Research (FoR) which is developed by the Australian Bureau of Statistics, New Zealand Ministry of Business, Innovation, and Employment, and other partners. This scheme covers many areas of science and indeed it is currently adopted by Dimensions.ai[7], a company that provides commercial solutions to support users in exploring the research landscape.

A common issue of these ontologies is that, being manually crafted and maintained by domain experts, they tend to evolve relatively slow and became quickly outdated. To keep-up with the pace of the constant evolution of the research landscape, some institutions (e.g., the American Physical Society) are crowd-sourcing their classification scheme. However, crowd-sourcing strategies suffer from limitations, such as trust and reliability [5]. Indeed, those institutions need to entrust a committee to moderate such amendments to the classification scheme.

2.2 Automatic extraction of ontologies from scholarly data

A new strategy which is becoming increasingly popular is the automatic or semi-automatic generation of ontologies of research topics using data-driven methodologies. In the state of the art we can find a variety of approaches that allow us to learn ontologies using clustering techniques, natural language processing, statistical methods and others.

TaxGen [14] is an approach for the automatic generation of taxonomies from a corpus using both hierarchical agglomerative clustering algorithm and text mining techniques. The algorithm, in a bottom-up fashion, first identifies the bottom clusters by observing the linguistic features in the documents, such as co-occurrences of words, domain terms, names of people, organisations and other significant words from text. Then the clusters are merged creating higher-level clusters, which form the hierarchy.

Text2Onto [4] is another approach for learning ontologies from a collection of documents. This approach identifies sub-/superclass hierarchies, synonyms, and other linguistic features through the application of natural language processing techniques on the sentence structure, where phrases like "such as..." and "and other..." imply a hierarchy between terms. This method presents some similarities with the Klink-2 algorithm [17], which we will describe later, but requires the full text of documents.

Sanderson et al. [28] developed an approach for automatically deriving a hierarchical organisation of concepts from a set of documents without use of

[7] Dimensions.ai - https://www.dimensions.ai

training data. Their approach computes the conditional probability for a keyword to be associated with another based on their co-occurrence. Given a pair of keywords, this system tries to understand whether there is a subsumption relationship between them, according to certain heuristics.

In literature, we can also find semi-automatic approaches that take advantage of external knowledge, either from pre-existing taxonomies or sourced by the community. For instance, Shen et al. [30] developed the Fields of Study (FoS) taxonomy, currently in use within Microsoft Academic, in which the first two levels are hand-crafted. Their approach is based on a variation of [28] and automatically infers topics derived from Wikipedia. However, considering only Wikipedia is a limitation as many research topics are not described there. Conversely, Klink-2 considers both academic publications and external sources. Another approach from Wohlgenannt et al. [32], combines human effort and machine computation by crowd-sourcing the evaluation of an automatically generated ontology with the aim of dynamically validating the extracted relations. Klink-2 can benefit these systems by generating an accurate, large-scale and up-to-date topic network.

3 Ontology Generation

An essential characteristic for ontologies of research areas is the hierarchical description of branches and sub-branches for a given discipline. Indeed, almost all ontologies mentioned above, e.g. MeSH, PhySH, ACM-CSS, are structured like a taxonomy: from the most generic to the most specific topics. In cases when there are different labels associated to the same topic, for instance for acronyms or synonyms, the ontology needs to be modelled in a way that can contain information about equivalences.

To this end, we designed the Klink-2 algorithm [17] that from a large corpus of scholarly publications is able to automatically generate ontologies of research areas, describing both hierarchical and equivalence relationships between research topics.

The algorithm starts by comparing each keyword in input to all the other keywords with which it shares at least n co-occurrences. In particular, it infers the semantic relationship between topics x and y using three metrics: i) $H_R(x, y)$, which uses a semantic variation of the subsumption method [28] for measuring the intensity of a hierarchical relationship; ii) $T_R(x, y)$, which uses temporal information to estimate the existence of a hierarchical relationship; and iii) $S_R(x, y)$, which estimates the similarity between two topics. The first two metrics are used to indicate whether a topic is super-area of another one (*superTopicOf*) or that the research outputs of one topic contributes to research of the other (*contributesTo*). The last metric is used to infer equivalence relationships between topics (*relatedEquivalent*).

$H_R(x,y)$ quantifies the hierarchical relationship between x and y according to the following formula:

$$H_R(x,y) = \left(\frac{I_R(x,y)}{I_R(x,x)} - \frac{I_R(y,x)}{I_R(y,y)} \right) \cdot c_R(x,y) \cdot n(x,y) \tag{1}$$

where $I_R(x,y)$ is the number of elements associated with both x and y according to relation R (e.g., number of co-occurrences in research papers), $(\frac{I_R(x,y)}{I_R(x,x)})$ is the conditional probability that an element associated with keyword x will be associated also with keyword y, $n(x,y)$ defines the string similarity between the two topics using the normalised Levenshtein distance, and finally $c_R(x,y)$ measures how similar are the distributions of topics with which both topic x and y are co-occurring, using cosine similarity.

$T_R(x,y)$ is a temporal version of $H_R(x,y)$, which weighs more the information associated with the first years of x. It is useful to detect the cases in which the relationship between two terms fades because their association has become implicit (e.g., Artificial Intelligence and Machine Learning). $T_R(x,y)$ is calculated using a variation of Eq. 1 in which $I_R(x,y)$ is computed by weighting the intensity of the relationships in each year according to the distance from the debut of x. The weight is computed as $w(year,x) = (year - debut(x) + 1) - \gamma$, with $\gamma > 0$ ($\gamma = 2$ in the prototype). Finally, $S_R(x,y)$ is used to assess the similarity of two terms and is computed according to the following formula:

$$S_R(x,y) = \frac{c_R(x,y)}{\max\left(c_R^{super}(x,y), c_R^{sib}(x,y)\right) + 1} \tag{2}$$

where $c_R^{super}(x,y)$ is the cosine similarity of the super topics of the two terms in the taxonomy produced by previous iteration, and $c_R^{sib}(x,y)$ is the cosine similarity of their siblings. A hierarchical relationship between two topics is inferred when a sufficient number of hierarchical indicators are above a threshold. An analysis of the precision/recall trade-off associated with different thresholds is available in [17]. The nature of the inferred relationship is assessed by Klink-2 using a rule-based approach. In brief, if x is older, associated with more entities, and the $T_R(x,y)$ indicators score higher, Klink-2 will infer a *superTopicOf* relationship, otherwise a *contributesTo* one. Then, Klink-2 removes loops in the topic network, merges keywords linked by a *relatedEquivalent* relationship, and splits ambiguous keywords associated to multiple meanings (e.g., “Java”). The keywords produced in this step are added to the initial set of keywords to be further analysed in the next iteration and the while-loop is re-executed until there are no more keywords to be processed.

Klink-2 filters the keywords considered too generic or not academic according to a set of heuristics that take in consideration the frequency of a keyword in various online sources and distribution of its co-occurrences [16]. Finally it generates the RDF triples describing the ontology. For a more comprehensive explanation of Klink-2, we refer the reader to Osborne at al. [17].

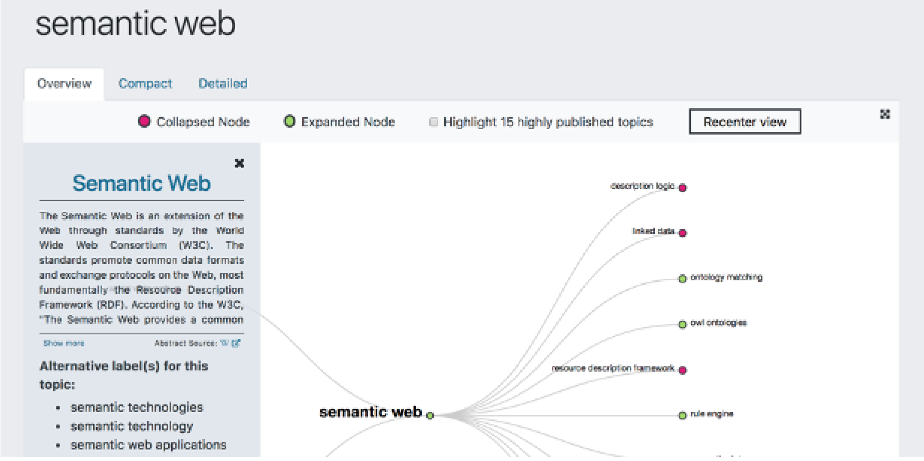
4 Computer Science Ontology

The Computer Science Ontology (CSO) is a large-scale, granular, and automatically generated ontology of research areas. It was generated by running the Klink-2 algorithm on the Rexplore dataset [18] containing 16 million publications in the field of Computer Science.

Currently CSO includes about 14K topics and 163K semantic relationships. The main root is Computer Science; however, the ontology includes also a few secondary roots, such as Linguistics, Geometry, Semantics, and so on. The CSO data model[8] is an extension of SKOS[9] and it includes eight semantic relations:

– *relatedEquivalent*, which is a subproperty of skos:related, indicates that two topics can be treated as equivalent for the purpose of exploring research data (e.g., Ontology Matching and Ontology Mapping).
– *superTopicOf*, which is a subproperty of skos:narrower, indicates that a topic is a super-area of another one (e.g., Semantic Web is a super-area of Linked Data). The inverse of this relationship is subTopicOf.
– *contributesTo*, which indicates that the research output of one topic contributes to another. For instance, research in Ontology Engineering contributes to Semantic Web, but arguably Ontology Engineering is not a subarea of Semantic Web, since there is plenty of research in Ontology Engineering outside the Semantic Web area.
– *owl:sameAs*, which is used for mapping CSO topics to equivaled entities in other knowledge graphs (DBpedia[10], Freebase[11], Wikidata[12], YAGO[13], and Cyc[14]).
– *schema:relatedLink*, which links CSO concepts to relevant web pages that either describe the research topics (Wikipedia articles) or provide additional information about the research domains (Microsoft Academic).
– *preferentialEquivalent*, which is used to state the main label for topics belonging to a cluster of relatedEquivalent. For instance, the topics Ontology Matching and Ontology Alignment both have their preferentialEquivalent set to Ontology Matching. Similarly to relatedEquivalent, in our data model we defined preferentialEquivalent as a subproperty of skos:related.
– *rdf:type*, this relation is used to state that a resource is an instance of a class. For example, a resource in our ontology is an instance of Topic, which is a subclass of skos:Concept.
– *rdfs:label*, this relation is used to provide a human-readable version of a resource's name.

[8] CSO data model - https://cso.kmi.open.ac.uk/schema/cso
[9] SKOS Simple Knowledge Organization System - http://www.w3.org/2004/02/skos.
[10] DBpedia - https://wiki.dbpedia.org
[11] Freebase - https://en.wikipedia.org/wiki/Freebase
[12] Wikidata - https://www.wikidata.org
[13] Yago - https://github.com/yago-naga/yago3
[14] Cyc - https://www.cyc.com



The Semantic Web is an extension of the Web through standards by the World Wide Web Consortium (W3C). The standards promote common data formats and exchange protocols on the Web, most fundamentally the Resource Description Framework (RDF). According to the W3C, "The Semantic Web provides a common

Alternative label(s) for this topic:

- semantic technologies
- semantic technology
- semantic web applications

Fig. 2. Overview of the resource page related to the topic in a new "semantic web".

In the previous section, we discussed how the Klink algorithm infers the first three of these relationships: *relatedEquivalent*, *superTopicOf*, and *contributesTo*. The rest of the relationships are also automatically generated. In particular, the *rdf:type* and *rdfs:label* relations respectively identify all topic entities and their labels. We generated the *preferentialEquivalent* choosing, within a cluster of topics linked by a *relatedEquivalent*, the label associated with most articles in the source corpus [18]. To generate the *owl:sameAs* relationships, linking CSO concepts to equivalent entities in the other KBs, we identified the DBpedia entities corresponding to CSO topics by exploiting the DBpedia Spotlight API [12]. Then we extracted the links from DBpedia to other KBs in the Linked Open Data (LOD) cloud by using the DBpedia SPARQL endpoint.

We also generated *schema:relatedLink* relations toward external web pages containing further information about the research topic. In particular, the links between CSO concepts and Wikipedia articles were extracted from the DBpedia. We also mapped the CSO topics to the Fields of Study (FoS) concepts reseased by Microsoft Academic. More details regarding the alignment between CSO and other knowledge bases are avaliable in Salatino et al. [26].

To facilitate the uptake of CSO we have developed the CSO Portal[15] (see Fig. 2), a web application that enables users to browse, download – in various formats[16], e.g. N-Triples, OWL, TTL and CSV – and provide granular feedback on CSO at different levels.

[15] CSO Portal - https://cso.kmi.open.ac.uk

[16] This ontology is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0) - https://creativecommons.org/licenses/by/4.0..

5 CSO Classifier

In order to facilitate users in integrating CSO in their pipelines, we developed the CSO Classifier [25], a tool that automatically annotates documents according to CSO. This application takes in input the metadata associated with a research paper (title, abstract, and keywords) and returns a selection of research concepts drawn from CSO. Figure 3 displays its workflow.

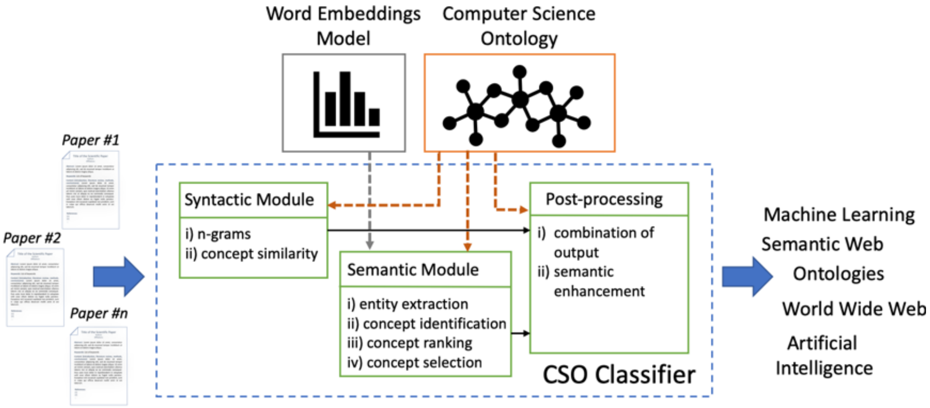


Fig. 3. Workflow of the CSO Classifier.

The CSO Classifier works in three steps. First, it finds all topics in the ontology that are explicitly mentioned in the paper (syntactic module). Then it identifies further semantically related topics by means of part-of-speech tagging and word embeddings (semantic module). Finally, it enriches this set by including the super-areas of these topics according to CSO.

In particular, the *syntactic module* removes English stop words and collects unigrams, bigrams, and trigrams. Then, for each n-gram, it computes the Levenshtein similarity with the labels of the topics in CSO. Finally, it returns all research topics whose labels have similarity to one of the n-grams, which is equal to or higher than a threshold.

The *semantic module* takes advantage of a pre-trained word embedding model which captures semantic properties of words [13]. We trained this Word2Vec model using titles and abstracts of 4,654,062 English publications in the field of Computer Science from Microsoft Academic Graph, which is an heterogeneous graph containing scientific publication records, citation relationships, authors, institutions, journals, conferences, and fields of study. We pre-processed this data by replacing spaces with underscores in all n-grams matching the CSO topic labels (e.g., “semantic web” became “semantic_web”) as well as for frequent bigrams and trigrams (e.g., “highest_accuracies”, “highly_cited_journals”). The latter were identified by analysing collocations, i.e. combinations of words that co-occur together [13]. This solution allows the CSO Classifier to better disambiguate concepts and treat terms such as "deep_learning" and "e-learning" as completely different words.

Specifically, to compute the semantic similarity between the terms in the document and the CSO concepts, the semantic module uses part-of-speech tagging to identify candidate terms composed by a combination of nouns and adjectives and decomposes them into unigrams, bigrams, and trigrams. For each n-gram, it retrieves its most similar words from the Word2Vec model. For this task, the n-gram tokens are initially glued with an underscore, creating one single word, e.g., "semantic_web". If this word is not available within the model vocabulary, the classifier uses the average of the embedding vectors of all its tokens. Then, it computes the relevance score for each topic in the ontology as the product between the number of times it was identified in those n-grams (frequency) and the number of unique n-grams that led to it (diversity). Finally, it uses the elbow method [29] for selecting the set of most relevant topics. The CSO Classifier aggregates the topics returned by the two modules and enriches them by inferring the list of all their super topics, exploiting the *superTopicOf* relationship within CSO. For instance, given the topic "Neural Networks", it will infer "Machine Learning", "Artificial Intelligence", and "Computer Science". This feature allows us to capture both high-level fields and very granular research areas, in order to generate a comprehensive representation of the classified papers.

The latest release of the CSO Classifier can be installed via *pip* from PyPI: pip install cso-classifier; or it simply downloading it from https://github.com/angelosalatino/cso-classifier.

6 High-level applications

This section describes some high-level applications that take advantage of CSO for supporting users in exploring, analysing, and making sense of large corpora of research publications.

6.1 Exploring and making sense of scholarly data

Rexplore is a system to support users in exploring and making sense of scholarly data [18]. It uses CSO to characterise research papers, authors, and organisations according to research topics. An interesting feature available in Rexplore is that it can plot a graph of researchers based on their topic similarity, reflecting how similar two authors are with respect to their research areas. Rexplore allows users to detect and make sense of important trends in research, such as significant migrations of researchers from one area to another, the emergence of new topics, the evolution of communities within a particular area. It also provides powerful query/search facilities, supporting complex multidimensional queries that can include logical connectives, such as retrieving career-young authors who have worked in both *Semantic Web* and *Social Networks*, and have published at the International Semantic Web Conference.

The Rexplore system was shown to be able to support users in performing specific tasks more effectively than Microsoft Academic Search (MAS), thanks to its organic representation of research topics [18].

6.2 Automatic classification of conference proceedings

The Smart Topic Miner (STM) [22] is a web application that supports the Springer Nature editorial team in classifying editorial products according to a taxonomy of research topics drawn both from CSO and the Product Market Codes (PMC), Springer Nature's own editorial classification system. STM takes as input the metadata associated with the proceedings of a conference (titles, abstracts and author-provided keywords for each paper in the proceedings) and returns the set of relevant CSO topics and PMCs as output.

STM uses the CSO Classifier to annotate each paper with the topics from CSO. Then it groups and ranks the topics according to the number of papers addressing them. Finally, it infers the relevant PMCs, using the mapping between the CSO ontology and PMC. The editors then review the CSO topics and the PMC categories and submit these annotations to the Springer Nature production system. This outcome is displayed in the Springer Nature's digital library: SpringerLink; and included in the ONIX[17] metadata feeds, delivered to various libraries and bookshops.

We released STM back in 2016, and since then it has been routinely used by the editorial team to annotate all book series covering conference proceedings in Computer Science, including LNCS, LNBIP, CCIS, IFIP-AICT and LNICST, for an amount of 800 volumes per year. STM has the advantage of halving the time required for classifying proceedings and it also reduced the complexity of the task, which was traditionally carried out by senior editors but now performed by junior editors. Its adoption produced a significant increment of the discoverability of relevant publications on SpringerLink, resulting in about 9 million additional downloads over the last three years. A demo of STM is available at http://stm-demo.kmi.open.ac.uk/.

6.3 Book recommendation

The Smart Book Recommender (SBR) [31] is an ontology-based recommender system that supports the Springer Nature editorial team in promoting their publications at Computer Science venues. It takes as input the proceedings of a conference and returns books, journals and other proceedings that are likely be of interest for its attendees.

SBR uses the CSO Classifier to represent more than 27K books and 320 journals according to their distribution of topics. Then, it identifies the most relevant editorial products, computing the similarity between the topical representation of the input conferences and other products available in the system. SBR also exploits the CSO topic taxonomy to graphically represent and compare conferences

[17] ONIX for Books - https://bisg.org/page/ONIXforBooks

and books, allowing users to understand the rationale behind its recommendations. A demo of SBR is available at http://rexplore.kmi.open.ac.uk/SBR-demo.

6.4 Forecasting research topics

Understanding and reacting timely to new developments in the research landscape is critical for a variety of stakeholders, such as funding bodies, academic publishers, companies and others. Augur [24] is a novel approach which uses CSO for anticipate the emergence of new research topics. Specifically, Augur analyses *topic networks*, i.e. collaboration networks between research communities associated with specific research areas, and identifies clusters associated with a significant increase in the pace of collaboration. Over these networks, Augur applies a novel clustering algorithm called the Advanced Clique Percolation Method (ACPM). The resulting clusters of topics indicate the areas of the network that are nurturing new research areas. Augur uses CSO for creating semantically-enhanced topic networks describing the collaboration between research topics over time. The evaluation of Augur proved that semantically enriching topics networks with CSO yields more than 30% increase of f-measure on the task of predicting the emergence of new research areas. Further details of Augur and its evaluation are available in Salatino et al. [24].

6.5 Systematic literature reviews

The aim of systematic reviews (SRs) is to find all the evidence relevant to a particular research question, and identify what can be said based on those findings. Typically, systematic reviews require domain experts to collect, annotate and synthesise hundreds of papers manually, using a well-defined methodology meant to mitigate the risks of biases and ensure repeatability for later updates. This task becomes extremely hard when investigating large numbers of papers (e.g. hundreds of thousands). The Expert-Driven Automatic Methodology (EDAM) [19] was developed for reducing the amount of tedious manual tasks involved in SRs while taking advantage of the value provided by human expertise. In particular, EDAM is able to i) characterise the area of interest using an ontology of topics, ii) ask domain experts to refine such an ontology, and iii) take advantage of this knowledge base for classifying relevant papers and producing useful analytics.

This approach uses the CSO Classifier to classify all research papers using title, abstract, and keywords. For a given topic, it then categorises all papers that were annotated with that specific topic, as well as all its *relatedEquivalent* and all its sub-branches, using the *superTopicOf* relation within CSO.

We evaluated the ability of EDAM to correctly discriminate between different topics in the field of Software Architecture by classifying a set of randomly-selected papers both with EDAM and with six human experts. We compared the annotation produced by both human experts and EDAM, considering the latter as an additional annotator. EDAM performance was not statistically significantly different from that of six senior researchers in the field (p=0.77). The approach adopting CSO yielded the highest average agreement and also obtained the highest agreement with three out of six domain experts. Further details about this evaluation and our results are available in Osborne et al. [19].

6.6 Forecasting technology adoption

Typically, the spreading of a technology from a one research area (e.g., Semantic Web) to a different and possibly conceptually distant area (e.g., Digital Humanities) may take several years, potentially delaying the research process.

The Technology-Topic Framework (TTF) [15] is an approach that suggests promising technologies to scholars in order to accelerate the pace of technology propagation. It characterises technologies according to their propagation through research topics drawn from CSO, and uses this representation to forecast the propagation of novel technologies across research fields.

TTF was evaluated on a set of 1,118 technologies in the fields of Semantic Web and Artificial Intelligence, yielding a precision of 74.4% and a recall of 47.7% for the first 20 research areas. More details about Technology-Topic Framework are available in Osborne et al. [15].

6.7 Scientific knowledge graphs generation

Scientific Knowledge Graphs (SKGs) are semantic graph databases that model scholarly knowledge in a structured, interlinked, and semantically rich manner. SKGs describe the actors (e.g., authors, organisations), the documents (e.g., publications, patents), and the research knowledge (e.g., research topics, tasks, technologies) in this space, as well as their reciprocal relationships.

CSO currently support two of these resources. The first is the Academia and Industry DynAmics (AIDA) knowledge graph [1] whicb describes 14M papers and 8M patents in the field of Computer Science. It was generated by by automatically integrating data from Microsoft Academic Graph, Dimensions, English DBpedia, the Computer Science Ontology, and the Global Research Identifier Database. The CSO Classifier was used to annotate both papers and patents according to their relevant topics in CSO. 4M papers and 5M patents are also categorised according to the type of the author's affiliations (academy, industry, or collaborative) and 66 industrial sectors (e.g., automotive, financial, energy, electronics) obtained from DBpedia. More details are available in Angioni et al. [1]. AIDA can be browsed and downloaded from http://w3id.org/aida.

The second knowledge base integrating CSO is the Artificial Intelligence Knowledge Graph (AI-KG) [6], which is a large-scale knowledge graph that describes about 850K research entities. AI-KG includes 1,2M statements extracted from 333K research publications in the field of AI and describes 5 types of entities (e.g., tasks, methods, metrics, materials, others) linked by 27 relations. It was designed to support a large variety of intelligent services for analyzing and making sense of research dynamics, supporting researchers in their daily job, and informing decision of founding bodies and governments.

AI-KG was generated by applying an automatic pipeline that extracts entities and relationships using three tools: DyGIE++, Stanford CoreNLP, and the CSO Classifier. It is available under CC BY 4.0 can be browsed and downloaded from http://w3id.org/aikg.

7 Conclusions and Future Work

In this paper, we presented the Computer Science Ontology Framework, a conceptual framework that characterises the design, extraction and use of the Computer Science Ontology, which is currently the largest taxonomy of research topics in Computer Science. This framework includes the CSO classifier, a tool for annotating research papers according to a domain ontology. We described several high-level applications that take advantage of CSO for supporting the exploration of the research landscape and forecasting research dynamics.

We are now working on a new version of Klink-2, in order to produce larger and more accurate ontologies of research topics. We also plan to explore the application of this framework in other research fields, such as Engineering and Life Science.

References

1. Angioni, S., Salatino, A., Osborne, F., Reforgiato Recupero, D., Motta, E.: Integrating knowledge graphs for analysing academia and industry dynamics. In: ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium. Springer International Publishing, Cham (2020)
2. Bettencourt, L.M., Kaiser, D.I., Kaur, J.: Scientific discovery and topological transitions in collaboration networks. Journal of Informetrics **3**(3), 210–221 (2009)
3. Boyack, K.W., Newman, D., Duhon, R.J., Klavans, R., Patek, M., Biberstine, J.R., Schijvenaars, B., Skupin, A., Ma, N., Börner, K.: Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. PLOS ONE **6**(3), 1–11 (03 2011). https://doi.org/10.1371/journal.pone.0018029, https://doi.org/10.1371/journal.pone.0018029
4. Cimiano, P., Völker, J.: Text2onto. In: Montoyo, A., Muńoz, R., Métais, E. (eds.) Natural Language Processing and Information Systems. pp. 227–238. Springer Berlin Heidelberg, Berlin, Heidelberg (2005)
5. Clough, P., Sanderson, M., Tang, J., Gollins, T., Warner, A.: Examining the limits of crowdsourcing for relevance assessment. IEEE Internet Computing **17**(4), 32–38 (July 2013). https://doi.org/10.1109/MIC.2012.95
6. Dessì, D., Osborne, F., Reforgiato Recupero, D., Buscaldi, D., Motta, E., Sack, H.: Ai-kg: an automatically generated knowledge graph of artificial intelligence. In: International Semantic Web Conference. Springer International Publishing, Cham (2020)
7. Di Noia, T., Magarelli, C., Maurino, A., Palmonari, M., Rula, A.: Using ontology-based data summarization to develop semantics-aware recommender systems. In: European Semantic Web Conference. pp. 128–144. Springer (2018)

8. Fathalla, S., Vahdati, S., Auer, S., Lange, C.: Towards a knowledge graph representing research findings by semantifying survey articles. In: International Conference on Theory and Practice of Digital Libraries. pp. 315–327. Springer (2017)
9. Kirrane, S., Sabou, M., Fernández, J.D., Osborne, F., Robin, C., Buitelaar, P., Motta, E., Polleres, A.: A decade of semantic web research through the lenses of a mixed methods approach. Semantic Web Journal (2019)
10. Kishore, R., Ramesh, R.: Ontologies: a handbook of principles, concepts and applications in information systems, vol. 14. Springer Science & Business Media (2007)
11. Livingston, K.M., Bada, M., Baumgartner, W.A., Hunter, L.E.: Kabob: ontology-based semantic integration of biomedical databases. BMC bioinformatics **16**(1), 126 (2015)
12. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: Shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems. pp. 1–8. I-Semantics '11, ACM, New York, NY, USA (2011). https://doi.org/10.1145/2063518.2063519, http://doi.acm.org/10.1145/2063518.2063519
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. pp. 3111–3119. NIPS'13, Curran Associates Inc., USA (2013), http://dl.acm.org/citation.cfm?id=2999792.2999959
14. Muller, A., Dorre, J., Gerstl, P., Seiffert, R.: The taxgen framework: automating the generation of a taxonomy for a large document collection. In: Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers. vol. Track2, pp. 9 pp.– (Jan 1999). https://doi.org/10.1109/HICSS.1999.772687
15. Osborne, F., Mannocci, A., Motta, E.: Forecasting the spreading of technologies in research communities. In: Proceedings of the Knowledge Capture Conference. pp. 1:1–1:8. K-CAP 2017, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3148011.3148030, http://doi.acm.org/10.1145/3148011.3148030
16. Osborne, F., Motta, E.: Mining semantic relations between research areas. In: International Semantic Web Conference. pp. 410–426. Springer (2012)
17. Osborne, F., Motta, E.: Klink-2: Integrating multiple web sources to generate semantic topic networks. In: Arenas, M., Corcho, O., Simperl, E., Strohmaier, M., d'Aquin, M., Srinivas, K., Groth, P., Dumontier, M., Heflin, J., Thirunarayan, K., Thirunarayan, K., Staab, S. (eds.) The Semantic Web - ISWC 2015. pp. 408–424. Springer International Publishing, Cham (2015)
18. Osborne, F., Motta, E., Mulholland, P.: Exploring scholarly data with rexplore. In: International semantic web conference. pp. 460–477. Springer (2013)
19. Osborne, F., Muccini, H., Lago, P., Motta, E.: Reducing the effort for systematic reviews in software engineering. Data Science pp. 1–29 (08 2019). https://doi.org/10.3233/DS-190019
20. Osborne, F., Salatino, A., Birukou, A., Motta, E.: Automatic classification of springer nature proceedings with smart topic miner. In: Groth, P., Simperl, E., Gray, A., Sabou, M., Krötzsch, M., Lecue, F., Flöck, F., Gil, Y. (eds.) The Semantic Web – ISWC 2016. pp. 383–399. Springer International Publishing, Cham (2016)
21. Osborne, F., Scavo, G., Motta, E.: Identifying diachronic topic-based research communities by clustering shared research trajectories. In: European Semantic Web Conference. pp. 114–129. Springer (2014)

22. Salatino, A.A., Osborne, F., Birukou, A., Motta, E.: Improving editorial workflow and metadata quality at springer nature. In: Ghidini, C., Hartig, O., Maleshkova, M., Svátek, V., Cruz, I., Hogan, A., Song, J., Lefrançois, M., Gandon, F. (eds.) The Semantic Web – ISWC 2019. pp. 507–525. Springer International Publishing, Cham (2019)
23. Salatino, A.A., Osborne, F., Motta, E.: How are topics born? understanding the research dynamics preceding the emergence of new areas. PeerJ Computer Science **3**, e119 (2017)
24. Salatino, A.A., Osborne, F., Motta, E.: Augur: Forecasting the emergence of new research topics. In: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries. pp. 303–312. JCDL '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3197026.3197052
25. Salatino, A.A., Osborne, F., Thanapalasingam, T., Motta, E.: The cso classifier: Ontology-driven detection of research topics in scholarly articles. In: Doucet, A., Isaac, A., Golub, K., Aalberg, T., Jatowt, A. (eds.) Digital Libraries for Open Knowledge. pp. 296–311. Springer International Publishing, Cham (2019)
26. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Birukou, A., Osborne, F., Motta, E.: The computer science ontology: A comprehensive automatically-generated taxonomy of research areas. Data Intelligence **0**(0), 1–38 (0). https://doi.org/10.1162/dint_a_00055, https://doi.org/10.1162/dint_a_00055
27. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Osborne, F., Motta, E.: The computer science ontology: a large-scale taxonomy of research areas. In: International Semantic Web Conference. pp. 187–205. Springer (2018)
28. Sanderson, M., Croft, B.: Deriving concept hierarchies from text. In: Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 206–213. SIGIR '99, ACM, New York, NY, USA (1999). https://doi.org/10.1145/312624.312679
29. Satopaa, V., Albrecht, J., Irwin, D., Raghavan, B.: Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In: 2011 31st International Conference on Distributed Computing Systems Workshops. pp. 166–171 (June 2011). https://doi.org/10.1109/ICDCSW.2011.20
30. Shen, Z., Ma, H., Wang, K.: A web-scale system for scientific knowledge exploration. In: Proceedings of ACL 2018, System Demonstrations. pp. 87–92. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-4015
31. Thanapalasingam, T., Osborne, F., Birukou, A., Motta, E.: Ontology-based recommendation of editorial products. In: Vrandečić, D., Bontcheva, K., Suárez-Figueroa, M.C., Presutti, V., Celino, I., Sabou, M., Kaffee, L.A., Simperl, E. (eds.) The Semantic Web – ISWC 2018. pp. 341–358. Springer International Publishing, Cham (2018)
32. Wohlgenannt, G., Weichselbraun, A., Scharl, A., Sabou, M.: Dynamic integration of multiple evidence sources for ontology learning. JIDM **3**, 243–254 (2012)