Detection, Analysis, and Prediction of Research Topics with Scientific Knowledge Graphs

Angelo A. Salatino[1 [0000−0002−4763−3943]], Andrea Mannocci[2 [0000−0002−5193−7851]], and Francesco Osborne[1 [0000−0001−6557−3131]]

[1] Knowledge Media Institute - The Open University,
Milton Keynes, United Kingdom
{angelo.salatino, francesco.osborne}@open.ac.uk
[2] Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo",
Italian National Research Council, Pisa, Italy
andrea.mannocci@isti.cnr.it

Abstract. Analysing research trends and predicting their impact on academia and industry is crucial to gain a deeper understanding of the advances in a research field and to inform critical decisions about research funding and technology adoption. In the last years, we saw the emergence of several publicly-available and large-scale Scientific Knowledge Graphs fostering the development of many data-driven approaches for performing quantitative analyses of research trends. This chapter presents an innovative framework for detecting, analysing, and forecasting research topics based on a large-scale knowledge graph characterising research articles according to the research topics from the Computer Science Ontology. We discuss the advantages of a solution based on a formal representation of topics and describe how it was applied to produce bibliometric studies and innovative tools for analysing and predicting research dynamics.

Keywords: Scholarly Communication · Scientific Knowledge Graphs · Research Trends · Research Topics · Research Trends

1 Introduction

Analysing research trends and predicting their impact on academia and industry is key to gain a deeper understanding of the research advancements in a field and to inform critical decisions about research funding and technology adoption. These analyses were initially performed through qualitative approaches in a top-down fashion, relying on experts' knowledge and manual inspection of the state of the art [36,85]. However, the ever-increasing number of research publications [11] has made this manual approach less feasible [56].

In the last years, we witnessed the emergence of several publicly-available, large-scale Scientific Knowledge Graphs (SKGs) [30], describing research articles and their metadata, e.g., authors, organisations, keywords, and citations. These resources fostered the development of many bottom-up, data-driven approaches to perform quantitative analyses of research trends. These methods usually cover one or more of three tasks: *i)* detection of research topics, *ii)* scientometric analyses, and *iii)* prediction of research

trends. In the first task, the articles are classified according to a set of research topics, typically using probabilistic topics models or different types of classifiers [8,9,33,12,87]. In the second task, the topics are analysed according to different bibliometrics over time [32,29,22]. For instance, each topic may be associated with the number of relevant publications or citations across the years in order to determine if their popularity in the research community is growing or decreasing. In the third task, regression techniques or supervised forecasters are typically used for predicting the future performance of research topics [84,37,70].

This chapter presents an innovative framework for detecting, analysing, and forecasting research topics based on a large-scale knowledge graph characterising research articles according to the research topics from the Computer Science Ontology (CSO) [72,73]. In describing this framework we leverage and reorganise our previous research works. Specifically, we discuss the advantage of a solution based on a formal representation of research topics and describe how it was applied to produce bibliometrics studies [34,47] and novel approaches for analysing [56,66,2] and predicting [66,70,52] research dynamics.

The chapter is articulated in six sections. In Section 2, we review state-of-the-art approaches for topic detection, analysis, and forecasting. In Section 3, we formally define a scientific knowledge graph, then we describe CSO [72] and the CSO Classifier [71], which is a tool for automatically annotating research papers with topics drawn from CSO. We then discuss how we generate the Academia/Industry DynAmics (AIDA) Knowledge Graph [3], a new SKG that describes research articles and patents according to the relevant research topics.

Section 4 describes three recent methods for analysing research trends that adopt this framework:

- *the EDAM methodology* [56] for supporting systematic reviews (Section 4.1);
- *the ResearchFlow framework* [66] for analysing trends across academia and industry (Section 4.2);
- *the AIDA Dashboard* [2] for supporting trend analysis in scientific conferences (Section 4.3).

In Section 5, we discuss three approaches that use the same framework for forecasting trends:

- *Augur* [70] (Section 5.1), a method for predicting the emergence of novel research topics;
- *the ResearchFlow forecaster* [66] (Section 5.2), a supervised deep learning classifier for predicting the impact of research topics on the industrial sector;
- *the Technology-Topic Framework (TTF)* [52] (Section 5.3), an approach for predicting the technologies that will be adopted by a research field.

Finally, in Section 6, we outline the conclusions and future work.

2 Literature Review

Detecting, analysing and predicting topics in the scientific literature has attracted considerable attention in the last two decades [9,78,19,24,44,67]. In this section, we review

the state of the art according to the main themes of this chapter. In particular, we initially review existing approaches for detecting topics in scholarly corpora. Then, we show approaches performing the analysis of research trends and finally, we describe the approaches for predicting research trends.

2.1 Topic detection

Topic detection in the scholarly communication field aims at identifying relevant subjects within a set of scientific documents. This operation can support several subsequent tasks, such as suggesting relevant publications [77], assisting the annotation of articles [60] and video lessons [10], assessing academic impact [16], recommending items [15,40], and building knowledge graphs of research concepts (e.g., AI-KG [21], ORKG [31], TKG [65]).

In the body of literature, we can find several approaches that can be characterised according to four main categories: *i)* topic modelling [8,29,22], *ii)* supervised machine learning approaches [33,45,27], *iii)* approaches based on citation networks [12], and *iv)* approaches based on natural language processing [19,32].

For what concerns the topic modelling category, we can find the Latent Dirichlet Analysis (LDA) developed by Blei et al. [8]. LDA is a three-level hierarchical Bayesian model to retrieve latent – or hidden – patterns in texts. The basic idea is that each document is modelled as a mixture of topics, where a topic is a multinomial distribution over words, which is a discrete probability distribution defining the likelihood that each word will appear in a given topic. In brief, LDA aims to discover the latent structure, which connects words to topics and topics to documents. Over the years, LDA influenced many other approaches, such as Griffiths et al. [28] and Bolelli et al. [9] in which they designed generative models for document collections. Their models simultaneously modelled the content of documents and the interests of authors. Besides, Bolelli et al. [9] exploit citations information to evaluate the main terms in documents. However, since topic modelling aims at representing topics as a distribution of words, it is often tricky to map them to research subjects.

For supervised machine learning approaches, Kandimalla et al. [33] propose a deep attentive neural network for classifying papers according to 104 Web of Science[3] (WoS) subject categories. Their classifier was trained on 9 million abstracts from WoS, and it can be directly applied to abstract without the need for additional information, e.g., citations. However, the fact that they can map research papers to a limited set of only 104 subject categories is due to the intensive human labelling effort behind the generation of a gold standard that foresees all possible research topics, and that is also balanced with regard to the number of papers labelled per topic. Indeed, broad areas tend to have many published papers and thus are highly represented, while very specific areas tend to have fewer papers.

Among the citation network approaches, there is Boyack and Klavans [12] who developed a map of science using 20 million research articles over 16 years using co-citation techniques. Through this map, it is possible to observe the disciplinary structure

[3]Web of Science - `https://clarivate.com/webofsciencegroup/solutions/web-of-science`

of science in which papers of the same area tend to group together. The main drawback of citation-based approaches is that they are able to assign each document to one topic only, while a document is seldom monothematic.

For the category of natural language processing approaches, we find Decker [19] who introduced an unsupervised approach that generates paper-topic relations by exploiting keywords and words extracted from the abstracts in order to analyse the trends of topics on different timescales. Additionally, Duvvuru et al. [23] relied on keywords to build their co-occurring network and subsequently perform statistical analysis by calculating degree, strength, clustering coefficient, and end-point degree to identify clusters and associate them to research topics. Some recent approaches use word embeddings aiming to quantify semantic similarities between words. For instance, Zhang et al. [87] applied clustering techniques on a set of words represented as embeddings. However, all these approaches to topic detection need to generate the topics from scratch rather than exploiting a domain vocabulary, taxonomy, or ontology, resulting in noisier and less interpretable results [53].

In brief, state-of-the-art approaches for detecting topics either use keywords as proxies for topics, or match terms to manually curated taxonomies, or apply statistical techniques to associate topics to bags of words.

Unfortunately, most of these solutions are suboptimal as they *i)* fail to manage polysemies and synonyms, respectively, when a keyword may denote different topics depending on the context and when multiple labels exist for the same research area (i.e., productive ambiguity vs maximal determinacy); and *ii)* fail to model and take advantage of the semantic relations that may hold between research areas, treating them as lists of unstructured keywords.

2.2 Research trends analysis

Research trends analysis deals with the dynamic component of topics, as it aims at observing their development over time. In the literature, we can find a variety of approaches based on citation patterns between documents [29,32], or co-word analysis [22]. Jo et al. [32] developed an approach that combines the distribution of terms (i.e., n-grams) with the distribution of the citation graph related to publications containing that term. They assume that if a term is relevant for a topic, documents containing that term will have a stronger connection (i.e., citations) than randomly selected ones. Then, the algorithm identifies the set of terms having citation patterns that exhibit synergy. Similarly, He et al. [29] combined the citation network with Latent Dirichlet Allocation (LDA) [8]. They detect topics in independent subsets of a corpus and leverage citations to connect topics in different time frames. However, these approaches suffer from time lag, as newly published papers usually require some time, if not several years, before getting noticed and being cited [79].

Di Caro et al. [22] designed an approach for observing how topics evolve over time. After splitting the collection of documents according to different time windows, their approach selects two consecutive slices of the corpus, extracts topics using LDA, and analyses how such topics change from one time window to the other. The central assumption is that by comparing the topics generated in two adjacent time windows, it is possible to observe how topics evolve as well as capture their birth and death.

Other methods for analysing trends in research include overlay mapping techniques, which create maps of science and enable users to assess trends [13,39]. Although these approaches provide a global perspective, the interpretation of such maps heavily relies on visual inspection by human experts.

2.3 Research trends prediction

Predicting research trends deals with anticipating the emergence of new topics. This task is particularly complex, mostly because of the limited availability of gold standards, i.e., extensive datasets of publication records manually and reliably labelled with research topics, that can be used to train a machine learning model to forecast trends. Indeed, regarding this aspect, we can find a very limited number of approaches in the literature.

Widodo et al. [84] performed a time series analysis to make predictions on PubMed data, and their experiments show that such time series are more suitable for machine learning techniques rather than statistical approaches. Instead, Krampen et al. [37] also rely on time series computing the Single Moving Average method to predict research trends in psychology. Watatani et al. [83] developed a bibliometric approach based on co-citation analysis, which provides insights for technology forecasting and strategic research, as well as development planning and policy-making.

In brief, in the literature, we can find many approaches capable of tracking topics over time. However, they focus on recognised topics, which are already associated with a good number of publications. Forecasting new research topics is yet an open challenge.

3 Topic Detection with Scientific Knowledge Graphs

In this section, we define the Scholarly Knowledge Graph (Section 3.1), and in particular, we introduce the main one used to fuel our research: Microsoft Academic Graph (MAG). In Section 3.2, we describe the Computer Science Ontology (CSO) and the CSO Classifier, which are at the foundation of our ontology-based topic detection method. Finally, in Section 3.3, we showcase the Academia/Industry DynAmics (AIDA) Knowledge Graph, an SKG derived by enriching MAG with the topics from CSO.

3.1 KG of scholarly data

Knowledge graphs are large collections of entities and relations describing real-world objects and events, or abstract concepts. Scientific Knowledge Graphs focus on the scholarly domain and describe the actors (e.g., authors, organisations), the documents (e.g., publications, patents), the research knowledge (e.g., research topics, tasks, technologies), and any other contextual information (e.g., project, funding) in an interlinked manner. Such descriptions have formal semantics allowing both computers and people to process them efficiently and unambiguously. Being interconnected within a network, each entity contributes to the description of the entities related to it, providing a wider context for their interpretation. SKGs provide substantial benefits to researchers, companies, and policymakers by powering several data-driven services for navigating, analysing, and making sense of research dynamics. Some examples include Microsoft Academic

Graph (MAG)[4] [82], AMiner [76], ScholarlyData[5] [50], PID Graph[6] [25], SciGraph[7], Open Research Knowledge Graph[8] [31], OpenCitations[9] [61], and OpenAIRE research graph[10] [46].

More formally, given a set of entities E, and a set of relations R, a scientific knowledge graph is a directed multi-relational graph G that comprises triples (subject, predicate, object) and is a subset of the cross product $G \subseteq E \times R \times E$.

Figure 1 depicts an excerpt of a scientific knowledge graph representing some metadata about this current chapter.

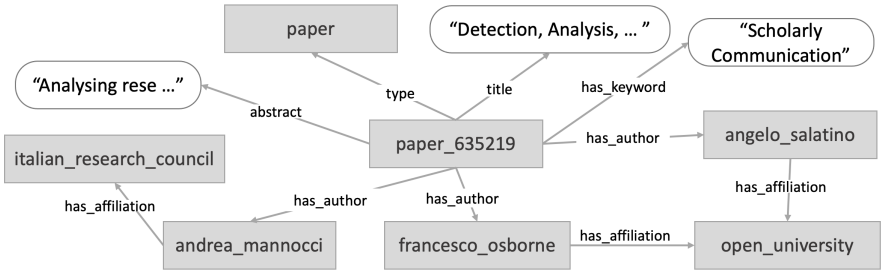

Fig. 1: Visualisation of a scientific knowledge graph. Entities are represented as rectangles and relations as arrows.

In this chapter, we will focus on *Microsoft Academic Graph* (MAG) [82,75], a pan-publisher, longitudinal scholarly knowledge graph produced and actively maintained by Microsoft, which contains scientific publication records, citation relations, authors, institutions, journals, conferences, and fields of study.

We decided to adopt MAG because it is the most extensive datasets of scholarly data publicly available [81], containing more than 250 million publications as of January 2021. In addition, it is widely adopted by the scientometrics community, making our results easier to replicate.

At the time of writing, MAG is released at no cost (*per se*) for research purposes, while a small fee is required to cover the costs and maintenance of the provisioning infrastructure, which is based on Azure. MAG can be either processed on the cloud, leveraging Azure cloud computing solutions (whose usage and costs fall outside the scope of the present chapter) or downloaded as a whole. In the latter case, the current

[4]Microsoft Academic Graph - `https://www.microsoft.com/en-us/research/project/academic`

[5]ScholarlyData - `http://www.scholarlydata.org`

[6]PID Graph - `https://www.project-freya.eu/en/pid-graph/the-pid-graph`

[7]SciGraph datasets - `https://sn-scigraph.figshare.com`

[8]Open Research Knowledge Graph - `https://www.orkg.org/orkg`

[9]OpenCitations - `https://opencitations.net`

[10]OpenAIRE research graph - `https://graph.openaire.eu`

due costs are around 100$ a go for bandwidth (likely to increase slightly over time as the dataset grows and is regenerated fortnightly), plus a few dollars a month for data parking in case the dataset is kept on the cloud as a backup, rather than deleted right away. In this form, MAG is provided as a collection of TSV files via Microsoft Azure Storage and features an ODC-BY[11] licence, an aspect that is essential to ensure the transparency and reproducibility of any resulting analysis in compliance with Open Science best practices. The information contained in this dataset consists of six types of main entities in the scholarly domain — publications, authors, affiliations, venues (journals and conferences), fields of study, and events (specific conference instances). These entities are connected through relations like citations, authorship, and others. The relations between such entities are described in details in [75]. The dataset contains publication metadata, such as DOI, title, year of publication, and venue of publication; moreover, MAG provides, whenever possible, key data often not available in other publicly accessible datasets (e.g., Crossref[12], DBLP[13]), including authors' and affiliations' identifiers, which are often required to address some compelling research questions (e.g., collaboration networks). Although paper abstracts are available for many records in MAG, this, however, does not redistribute the publications' full-texts, somehow limiting full-scale NLP applications.

For the sake of completeness, in the case current costs and technical skills associated with MAG provisioning are not viable, other alternative SKGs repackaging MAG are available for free. *The Open Academic Graph*[14] (OAG) is a large SKG unifying two billion-scale academic graphs: MAG and AMiner. In mid-2017, the first version of OAG, which contains 166,192,182 papers from MAG and 154,771,162 papers from AMiner, and 64,639,608 linking (matching) relations between the two graphs, was released. The current version of OAG combines the MAG snapshot as of November 2018 and AMiner snapshots of July 2018 and January 2019. In this new release, authors, venues, newer publication data, and the corresponding matchings are available as well. Another feasible alternative is the *OpenAIRE dataset DOIboost* [38] which provides an "enhanced" version of Crossref that integrates information from Unpaywall, ORCID and MAG, such as author identifiers, affiliations, organisation identifiers, and abstracts. DOIboost is periodically released on Zenodo[15].

3.2 Ontology-based topic detection

In the following section, we describe the two main technologies enabling our methodology for ontology-based topic detection: the Computer Science Ontology (Section 3.2.1) and the CSO Classifier (Section 3.2.2).

[11]Open Data Commons Attribution Licence (ODC-BY) v1.0 - `https://opendatacommons.org/licenses/by/1.0`

[12]Crossref API - `https://github.com/CrossRef/rest-api-doc`

[13]DBLP - `https://dblp.uni-trier.de`

[14]Open Academic Graph - `https://www.microsoft.com/en-us/research/project/open-academic-graph`

[15]DOIboost laster release - `https://zenodo.org/record/3559699`

3.2.1 The Computer Science Ontology

The Computer Science Ontology is a large-scale ontology of research areas in the field of Computer Science. It was automatically generated using the Klink-2 algorithm [53] on a dataset of 16 million publications, mainly in the field of Computer Science [55]. Compared to other solutions available in the state of the art (e.g., the ACM Computing Classification System), the Computer Science Ontology includes a much higher number of research topics, which can support a more granular representation of the content of research papers, and it can be easily updated by rerunning Klink-2 on more recent corpora of publications.

The current version of CSO[16] includes 14 thousand semantic topics and 159 thousand relations. The main root is Computer Science; however, the ontology includes also a few additional roots, such as Linguistics, Geometry, Semantics, and others. The CSO data model[17] is an extension of SKOS[18] and it includes four main semantic relations:

- `superTopicOf`, which indicates that a topic is a super-area of another one (e.g., Semantic Web is a super-area of Linked Data).
- `relatedEquivalent`, which indicates that two given topics can be treated as equivalent for exploring research data (e.g., Ontology Matching and Ontology Mapping).
- `contributesTo`, which indicates that the research output of one topic contributes to another.
- `owl:sameAs`, this relation indicates that a research concept is identical to an external resource, i.e., DBpedia.

The Computer Science Ontology is available through the CSO Portal[19], a web application that enables users to download, explore, and visualise sections of the ontology. Moreover, users can use the portal to provide granular feedback at different levels, such as rating topics and relations, and suggesting missing relations.

CSO has been adopted by Springer Nature, one of the two main academic publishers, that used it to support several innovative applications, including the Smart Topic Miner [57], a tool designed to assist the Springer Nature editorial team in classifying proceedings[68], and Smart Book Recommender [77], an ontology-based recommend system for scientific volumes.

In the last few years CSO supported the creation of many innovative approaches and applications, including ontology-driven topic models (e.g., CoCoNoW [7]), recommender systems for articles (e.g., SBR [77]) and video lessons [10], tools for detecting research communities (e.g., TST [59], RCMB [58]), visualisation frameworks (e.g., ScholarLensViz [42], ConceptScope [86]), temporal knowledge graphs (e.g., TGK [65]), NLP frameworks for entity extraction [20], knowledge graph embeddings (e.g., Trans4E [49]), tools for identifying domain experts (e.g., VeTo [80]), and systems for predicting academic impact (e.g., ArtSim [16]), research topics (e.g., Augur [70]), ontology concepts (e.g., SIM [14], POE [54]), and technologies (e.g., TTF [52], TechMiner [51]). It was

[16]CSO is available for download at `https://w3id.org/cso/downloads`
[17]CSO Data Model - `https://cso.kmi.open.ac.uk/schema/cso`
[18]SKOS Simple Knowledge Organisation System - `http://www.w3.org/2004/02/skos`
[19]Computer Science Ontology Portal - `https://cso.kmi.open.ac.uk`

also used for several large-scale analyses of the literature (e.g., Cloud Computing [43], Software Engineering [17], Ecuadorian publications [17]).

3.2.2 The CSO Classifier

The CSO Classifier [71] is a tool for automatically classifying research papers according to the Computer Science Ontology. This application takes in input the metadata associated with a research document (title, abstract, and keywords) and returns a selection of research concepts drawn from CSO, as shown in Figure 2.

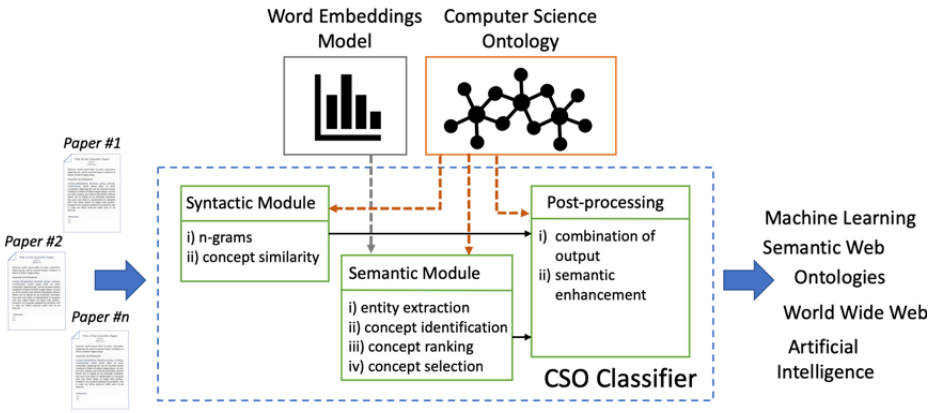

Fig. 2: Workflow of the CSO Classifier.

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 (*post-processing module*).

In particular, the syntactic module splits the text into unigrams, bigrams, and trigrams. Each n-gram is then compared with concept labels in CSO using the Levenshtein similarity. As a result, this module returns all matched topics having similarities greater than or equal to the pre-defined threshold.

The semantic module instead leverages a pre-trained Word2Vec word embedding model that captures the semantic properties of words [48]. We trained this model[20] using titles and abstracts of over 4.6 million English publications in the field of Computer Science from MAG. We preprocessed this data by replacing spaces with underscores in all n-grams matching the CSO topic labels (e.g., "semantic web" turned into "semantic_web"). We also performed a collocation analysis to identify frequent bigrams and trigrams (e.g., "highest_accuracies", "highly_cited_journals") [48]. This solution allows

[20]The model parameters are: *method* = skipgram, *embedding-size* = 128, *window-size* = 10, *min-count-cutoff* = 10, *max-iterations* = 5.

the classifier to disambiguate concepts better and treat terms such as "deep_learning" and "e-learning" as entirely different words.

More precisely, 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. 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 [74] for selecting the set of most relevant topics.

The resulting set of topics returned by the two modules is enriched in the post-processing module by including all their super-topics in CSO. For instance, a paper tagged as *neural network* is also tagged with *machine learning* and *artificial intelligence*. This solution yields an improved characterisation of high-level topics that are not directly referred to in the documents. More details about the CSO Classifier are available in [71].

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

3.3 The Academia/Industry DynAmics Knowledge Graph

In order to support new approaches for monitoring and predicting research trends across academia and industry, we build on CSO and the CSO Classifier to generate the Academia/Industry DynAmics (AIDA) Knowledge Graph [3]. This novel resource describes 21 million publications and 8 million patents according to the research topics drawn from CSO. 5.1 million publications and 5.6 million patents are further characterised according to the type of the author's affiliation (e.g., academia, industry) and 66 industrial sectors (e.g., automotive, financial, energy, electronics) organised in a two-level taxonomy.

AIDA was generated using an automatic pipeline that integrates several knowledge graphs and bibliographic corpora, including Microsoft Academic Graph, Dimensions, English DBpedia, the Computer Science Ontology [73], and the Global Research Identifier Database (GRID). Its pipeline consists of three phases: *i)* topics extraction, *ii)* extraction of affiliation types, and *iii)* industrial sector classification.

To extract topics from publications and patents, we run the CSO Classifier. Instead, to extract the affiliation types of a given document, we queried GRID[21], an open database identifying over 90 thousand organisations involved in research. GRID describes research institutions with a multitude of attributes, including their type, e.g., education, company, government, and others. A document is classified as "academic" if all its authors have an educational affiliation and as "industrial" if all its authors have an in-

[21]Global Research Identifier Database (GRID) - `https://grid.ac`

dustrial affiliation. Documents whose authors are from both academia and industry are classified as "collaborative".

Finally, we characterised documents from industry according to the Industrial Sectors Ontology (INDUSO)[22]. In particular, from GRID, we retrieved the Wikipedia entry of companies, and we queried DBpedia, the knowledge graph of Wikipedia. Then, we extracted the objects of the properties "`About:Purpose`" and "`About:Industry`", and mapped them to INDUSO sectors. All industrial documents were associated with the industrial sectors based on their affiliations.

Table 1: Distribution of publications and patents classified as Academia, Industry and Collaboration.

	Publications	**Patents**
Total documents	20,850,710	7,940,034
Documents with GRID IDs	5,133,171	5,639,252
Academia	3,585,060	133,604
Industry	787,151	4,741,695
Collaborative	133,781	16,335
Additional categories with GRID ID	627,179	747,618

Table 1 reports the number of publications and patents from academy, industry, and collaborative efforts. Most scientific publications (69.8%) are written by academic institutions, but industry is also a strong contributor (15.3%). Conversely, 84% of the patents are from industry and only 2.3% from academia. Collaborative efforts appear limited, including only 2.6% of the publications and 0.2% of patents.

The data model of AIDA is available at `http://aida.kmi.open.ac.uk/ontology` and it builds on SKOS[23] and CSO[24]. It focuses on four types of entities: *publications, patents, topics,* and *industrial sectors*.

The main information about publications and patents are given by means of the following semantic relations:

- `hasTopic`, which associates to the documents all their relevant topics drawn from CSO;
- `hasAffiliationType` and `hasAssigneeType`, which associates to the documents the three categories (academia, industry, or collaborative) describing the affiliations of their authors (for publications) or assignees (for patents);
- `hasIndustrialSector`, which associates to documents and affiliations the relevant industrial sectors drawn from the Industrial Sectors Ontology (INDUSO) we describe in the next sub-section.

A dump of AIDA in Terse RDF Triple Language (Turtle) is available at `http://aida.kmi.open.ac.uk/downloads`.

[22]INDUSO - `http://aida.kmi.open.ac.uk/downloads/induso.ttl`
[23]SKOS Simple Knowledge Organisation System - `http://www.w3.org/2004/02/skos`
[24]CSO Data Model - `https://cso.kmi.open.ac.uk/schema/cso`

4 Research Trends Analysis

In this section, we discuss different methodologies that use the framework described in the previous section for analysing the evolution of research topics in time. In particular, we first introduce the EDAM methodology for supporting systematic reviews (Section 4.1) and present two exemplary bibliometric analyses, then we describe the ResearchFlow approach for analysing research trends across academia and industry (Section 4.2), and finally, we introduce the AIDA Dashboard, a web application for examining research trends in scientific conferences (Section 4.3).

4.1 The EDAM Methodology for Systematic Reviews

A systematic review is *"a means of evaluating and interpreting all available research relevant to a particular research or topic area or phenomenon of interest"* [35]. Systematic reviews are used by researchers to create a picture of the state of the art in a research field, often including the most important topics and their trends. Given a set of research questions, and by following a systematically defined and reproducible process, a systematic review can assist the selection of a set of research articles (primary studies) that contribute to answering them [62].

In this section, we describe the EDAM (Expert-Driven Automatic Methodology) [56], which is a recent methodology for creating systematic reviews, thus limiting the number of tedious tasks that have to be performed by human experts. Specifically, EDAM can produce a comprehensive analysis of topic trends in a field by *i)* characterising the area of interest using an ontology of topics, *ii)* asking domain experts to refine this ontology, and *iii)* exploiting this knowledge base for classifying relevant papers and producing useful analytics. The main advantage of EDAM is that the authors of the systematic review are still in full control of the ontological schema used to represent the topics.

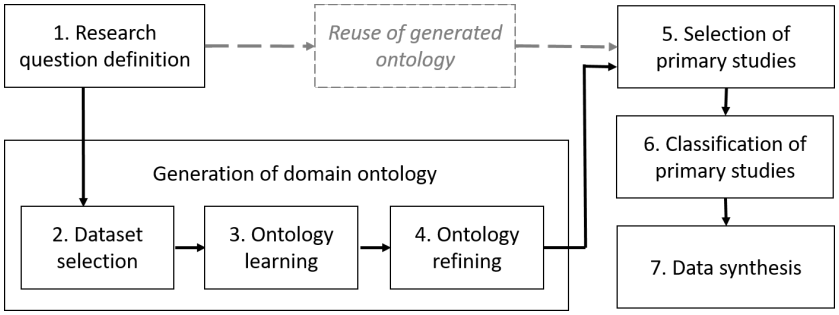

Fig. 3: Steps of a systematic mappings adopting the EDAM methodology. The gray-shaded elements refer to the alternative step of reusing the previously generated ontology.

Figure 3 illustrates the workflow of EDAM. First, the researchers define the research questions that will be answered by the produced analytics and select the data sources.

Then, they reuse or create the ontology that will support the classification process. The creation can also be supported by automatic methodologies for ontology learning such as *i)* statistical methods for deriving taxonomies from keywords [41]; *ii)* natural language processing approaches (e.g., FRED [26], LODifier [6], Text2Onto [18]); *iii)* approaches based on deep learning (e.g., recurrent neural networks [63]); *iv)* specific approaches for generating research topic ontologies (e.g., Klink-2 [53]). The resulting ontology is corrected and refined by domain experts. When the ontology is ready, the authors define the primary studies' inclusion criteria according to the domain ontology and other metadata of the papers (e.g., year, venue, language). The inclusion criteria are typically expressed as a search string, which uses simple logic constructs, such as AND, OR, and NOT [4]. The simplest way for mapping categories to papers is to associate to each category each paper that contains the label of the category or any of its subcategories. However, this step can also take advantage of a more advanced classification approach, such as the CSO Classifier described in Section 3.2.2. The last step is the computation of analytics to answer the research questions.

EDAM has also been used in a recent review that analyses the main research trends in the field of Semantic Web [34]. This study compares two datasets covering respectively the main Semantic Web venues (e.g., ISWC, ESWC, SEMANTiCS, SWJ, and JWS) and 32,431 publications associated with the Semantic Web from a dump of Scopus. The authors annotated the articles with CSO and produced several analyses based on the number of papers and citations in each research topic over the years. Specifically, articles were associated with a given topic if their title, abstract, or keywords field contained: *i)* the label of the topic (e.g., "semantic web"), *ii)* a `relevantEquivalent` of the topic (e.g., "semantic web tecnologies"), *iii)* a `skos:broaderGeneric` of the topic (e.g., "ontology matching"), or *iv)* a `relevantEquivalent` of any `skos:broaderGeneric` of the topic (e.g., "ontology mapping"). Figure 4 shows, for example, the popularity of the main topics in the main venue of the Semantic Web community. The analysis highlights two main dynamics confirmed by domain experts: the fading of Semantic Web Services and the rapid growth of Linked Data after the release of DBpedia in 2007 [5].

A further instantiation of the EDAM methodology can be found in [47], where Mannocci et al. performed a circumstantial, narrow-aimed study for the 50th anniversary of the International Journal of Human-Computer Studies (IJHCS) and delivered an analysis on the evolution throughout the last 50 years of the journal making an extensive comparison with another top-tier venue in the HCI research field: the CHI conference.

Among several other analyses (e.g., bibliometrics, spatial scientometrics), one aspect of the study focuses on topic analysis in both venues. To this end, it instantiates EDAM using CSO and the CSO classifier as reference ontology and classifier, respectively.

Figure 5 shows a comparison of the most growing topics in the two venues according to their number of articles in 2009-2013 and 2014-2018. These plots highlight very clearly which topics are gaining momentum in a community.

Figure 6 shows the geographical distribution of some of the most important topics in the two venues under analysis. This kind of analytics allows to easily assess the contribution of different countries for each topic. For instance, Japan is ranked seventh when considering the full CHI dataset, but in Robotics is third. Similarly, South Korea,

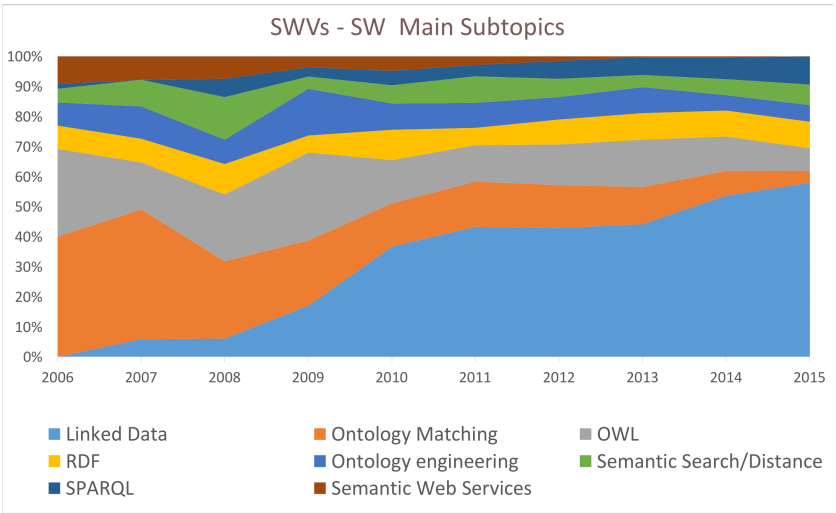

Fig. 4: Exemplary analysis in Semantic Web: Percentage of articles associated with the eight main subtopics of Semantic Web.

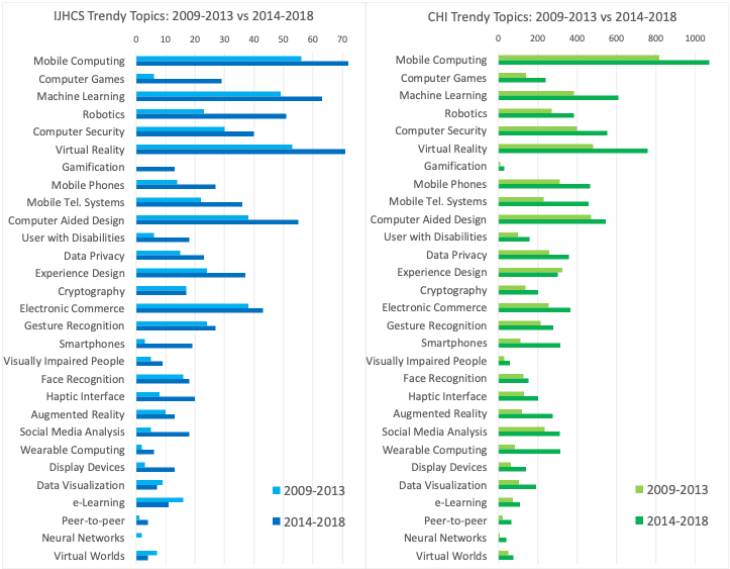

Fig. 5: Exemplary analysis in Human Computer Interaction: Growing Topics in IJHCS and CHI.

ranked 12th and 10th in IJHCS and CHI, is particularly active in Computer Security, ranking in fourth and sixth place when considering this topic.

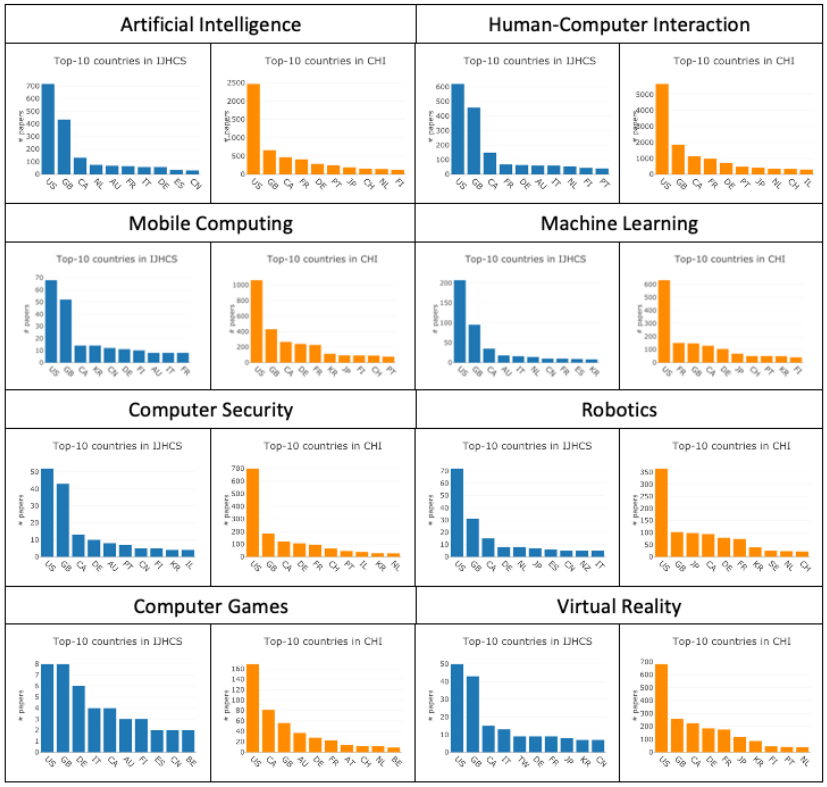

Fig. 6: Overall country contributions in IJHCS and CHI across the top-8 topics.

4.2 ResearchFlow: analysing trends across academia and industry

ResearchFlow [66] is a recent methodology for quantifying the diachronic behaviour of research topics in academia and industry that builds on the Academia/ Industry DynAmics (AIDA) Knowledge Graph [3], described in Section 3.3.

Specifically, ResearchFlow represents the evolution of research topics in CSO according to four time series reporting the frequency of *i)* papers from academia, *ii)* papers from industry, *iii)* patents from academia, and *iv)* patents from industry. Such characterisation allows us to *i)* study the diachronic behaviour of topics in order to characterise their trajectory across academia and industry, *ii)* compare each pair of signals to understand which one typically precedes the other and in which order they usually tackle a research topic, and *iii)* assess how signals influence each other by identifying pairs of signals that are highly correlated, after compensating for a time delay.

Figure 7 shows the distribution of the most frequent 5,000 topics from CSO in a bi-dimensional diagram according to two indexes: academia-industry (x-axis) and papers-patents (y-axis). The academia-industry index for a given topic t is the difference between the documents in academia and industry related to t, over the whole set of

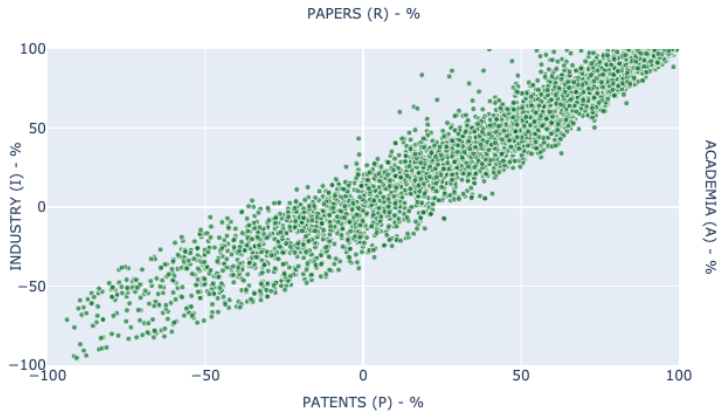

Fig. 7: Distribution of the most frequent 5,000 topics according to their academia-industry and publications-papers indexes.

documents. A topic is predominantly academic if this index is positive; otherwise, it is mostly populated by industry. On the other hand, the papers-patents index of a given topic t is the difference between the number of research papers and patents related to t, over the whole set of documents. If this index is positive, a topic tends to be associated with a higher number of publications; otherwise, it has a higher number of patents.

From the figure, we can observe that topics are tightly distributed around the bisector: the ones which attract more interest from academia are prevalently associated with publications (top-right quadrant). In contrast, the ones in industry are mostly associated with patents (bottom-left quadrant).

We also analysed which time series typically precedes another in first addressing a research topics. In particular, we determined when a topic emerges in all associated signals, and then we compared the time elapsed between each couple of signals. To avoid false positives, we considered a topic as ”emerged” when associated with at least ten documents. Our results showed that 89.8% of the topics first emerged in academic publications, 3.0% in industrial publications, 7.2% in industrial patents, and none in academic patents. On average, publications from academia preceded publications from industry by 5.6±5.6 years, and in turn, the latter preceded patents from industry by 1.0±5.8 years. Publications from academia also preceded patents from industry by 6.7±7.4 years.

4.3 The AIDA dashboard

The framework discussed in this chapter can also be used to support tools that allow users to analyse research dynamics. This is the case of the AIDA Dashboard [2], a novel web application for analysing scientific conferences which integrates statistical analysis, semantic technologies, and visual analytics. This application has been developed to support the editorial team at Springer Nature in assessing and evaluating the evolution of conferences through a number of bibliometric analytics. The AIDA Dashboard builds

on AIDA SKG and, compared to other state-of-the-art solutions, it introduces three novel features. First, it associates conferences with a very granular representation of their topics from the CSO [73] and uses it to produce several analyses about its research trends over time. Second, it enables an easy comparation and raking of conferences according to several metrics within specific fields (e.g., Digital Libraries) and timeframes (e.g., last five years). Finally, it offers several parameters for assessing the involvement of industry in a conference. This includes the ability to focus on commercial organisations and their performance, to report the ratio of publications and citations from academia, industry, and collaborative efforts, and to distinguish industrial contributions according to the 66 industrial sectors from INDUSO.

AIDA Dashboard is highly scalable and allows users to browse the different facets of a conference according to eight tabs: *Overview*, *Citation Analysis*, *Organisations*, *Countries*, *Authors*, *Topics*, *Related Conferences*, and *Industry*.

Figure 8 shows the *Overview* tab of the International Semantic Web Conference (ISWC). This is the main view of a conference that provides introductory information about its performance, the main authors and organisation, and the conference rank in its main fields in terms of average citations for papers during the last five years. More detailed and refined analytics are available in the other tabs.

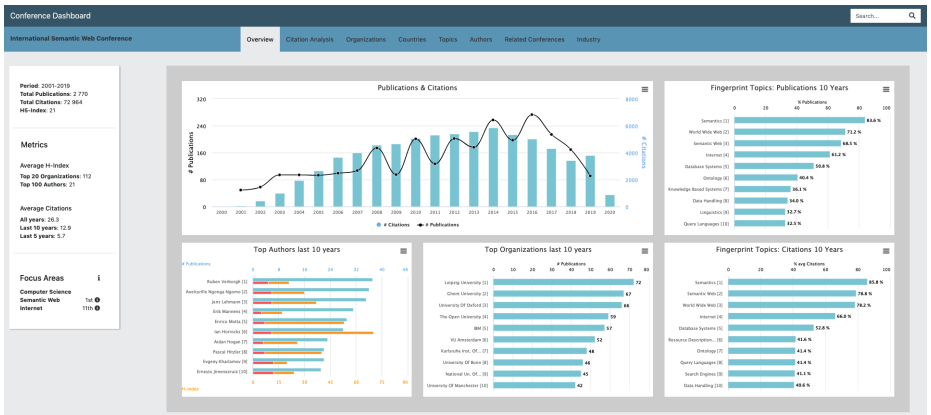

Fig. 8: The Overview of the International Semantic Web Conference (ISWC) according to the AIDA Dashboard.

5 Research Trends Prediction

In this section, we explore two approaches that use our framework for predicting research trends. In Section 5.1, we present Augur, a method for forecasting the emergence of new research trends which takes advantage of the collaborative networks of topics. In Section 5.2, we present an approach that was developed within the ResearchFlow framework and aims at predicting the impact of research topics on the industrial landscape.

In Section 5.3, we introduce the Technology-Topic Framework (TTF) [52], a novel methodology that takes advantage of a semantically enhanced technology-topic model for predicting the technologies that will be adopted by a research community.

5.1 Predicting the emergence of research topics

Augur [70] is an approach that aims to effectively detect the emergence of new research topics by analysing topic networks and identifying research areas exhibiting an overall increase in the pace of collaboration between already existing topics. This dynamic has been shown to be strongly associated with the emergence of new topics in a previous study by Salatino et al. [69].

A topic network is a fully weighted graph in which nodes represent topics, and the edges identify whether the connected pair of topics co-occurred together in at least one research paper. Additionally, the node weights count the number of papers in which a topic appears, and edge weights count the number of papers the two topics co-appear together. The original implementation of Augur extracted these networks using CSO on the Rexplore dataset [55].

Augur takes as input five topic networks characterising the collaborations between topics over five subsequent years (e.g., 2017-2021), as shown in the left side of Fig. 9. Over these networks, it then applies the Advanced Clique Percolation Method, a clustering algorithm for identifying groups of topics that over time intensify their pace of collaboration (centre of Fig. 9). The resulting clusters outline the portions of the network that may be nurturing research topics that will emerge in the following years (right side of Fig. 9).

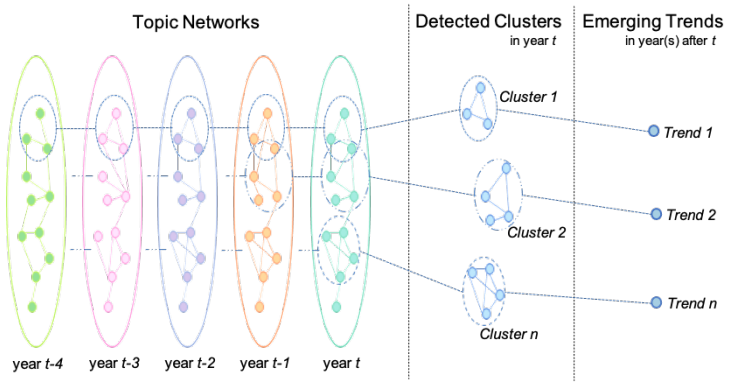

Fig. 9: Augur approach. On the left, the topic networks, whereas on the right there are emerging trends. The dashed circles represent groups of research topics nurturing new research trends.

In order to evaluate this approach, we built a gold standard of 1408 emerging trends in the 2000-2011 timeframe. Each emerging topic has been associated with a list of

related topics considered as "ancestors". The latter is crucial because Augur returns clusters of topics, which can also be considered ancestors of the future emerging trends. To this end, the evaluation consisted of checking whether the clusters of topics returned by Augur could be mapped with the ancestors of the emerging trends in the gold standard.

Table 2: Values of Precision and Recall for the five approaches along time. In bold the best results. The table and the experiments were previously reported in [70].

	FG		LE		FCM		CPM		ACPM	
Years	Precis.	Recall	Precis.	Recall	Precis.	Recall	Precis.	Recall	Precis.	Recall
1999	0.27	0.11	0.00	0.00	0.00	0.00	0.06	0.01	**0.86**	**0.76**
2000	0.21	0.07	0.14	0.02	0.96	0.01	0.05	0.00	**0.78**	**0.70**
2001	0.13	0.04	0.11	0.01	0.00	0.00	0.17	0.00	**0.77**	**0.72**
2002	0.14	0.04	0.11	0.01	0.00	0.00	0.29	0.01	**0.82**	**0.80**
2003	0.09	0.02	0.20	0.02	0.00	0.00	0.08	0.02	**0.83**	**0.79**
2004	0.11	0.05	0.06	0.00	0.00	0.00	0.00	0.00	**0.84**	**0.68**
2005	0.07	0.11	0.06	0.01	0.00	0.00	0.00	0.00	**0.71**	**0.66**
2006	0.01	0.01	0.07	0.01	0.00	0.00	0.00	0.00	**0.43**	**0.51**
2007	0.01	0.08	0.00	0.00	0.00	0.00	0.00	0.00	**0.28**	**0.44**
2008	0.01	0.04	0.00	0.00	0.00	0.00	0.00	0.00	**0.15**	**0.33**
2009	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.09**	**0.76**

We evaluated Augur against other four state-of-the-art algorithms: Fast Greedy (FG), Leading Eigenvector (LE), Fuzzy C-Means (FCM), and Clique Percolation Method (CPM). The evaluation has been performed across time, and as shown in Table 2, Augur outperformed the alternative approaches in terms of both precision and recall in all years. Further details of Augur, the gold standard and the evaluation are available in Salatino et al. [70].

5.2 The ResearchFlow Forecaster

In this section, we show how the ResearchFlow approach introduced in Section 4.2 can also support a very accurate forecaster to predict the impact of research topics in industry.

A good measure to assess the impact of research trends on commercial organisations is the number of relevant patents granted to companies. The literature proposes a range of approaches for this task [1,64]. However, most of these methods focus only on patents since they are limited by current datasets that do not typically integrate research articles nor can they distinguish between documents produced by academia or industry. We thus hypothesised that a scientific knowledge graph like AIDA, which integrates all the information about publications and patents and their origin, should offer a richer set of features, ultimately yielding a better performance in comparison to approaches that rely solely on the number of publications or patents.

Table 3: Performance of the five classifiers on 17 combinations of time series. In bold the best F1 (F) for each combination. The table and the experiments were previously reported in [66].

	LR			RF			AB			CNN			LSTM		
	P%.	R%	F%	P%	R%	F%	P%	R%	F%	P%	R%	F%	P%	R%	F%
RA	70.8	45.2	55.2	63.3	55.8	59.2	66.0	58.4	61.9	64.1	66.3	**65.0**	65.2	64.2	64.6
RI	83.5	67.1	74.4	78.9	69.8	74.0	80.0	73.1	76.4	79.2	75.1	**77.0**	79.1	74.8	76.9
PA	58.3	15.3	24.2	60.4	15.4	24.5	59.3	16.0	**25.2**	60.5	15.7	24.9	60.8	15.6	24.8
PI	76.5	69.0	72.5	73.9	68.4	71.0	75.6	71.8	73.6	73.7	76.6	75.0	74.1	76.6	**75.2**
R	73.7	48.8	58.7	65.5	59.7	62.5	68.6	63.1	65.6	67.6	69.2	**68.3**	67.2	69.4	68.2
P	76.5	68.6	72.3	72.8	67.6	70.0	74.4	71.6	73.0	73.2	76.1	74.6	73.1	76.6	**74.8**
RA, RI	85.7	70.9	77.6	80.5	76.0	78.2	82.6	76.6	79.5	78.9	75.1	76.8	82.2	79.3	**80.7**
RA, PA	70.3	47.0	56.3	63.1	55.5	59.0	66.5	59.3	62.6	64.5	65.1	64.5	65.4	64.2	**64.6**
RA, PI	79.6	73.7	76.5	77.2	74.3	75.7	79.1	76.5	77.7	75.2	76.3	75.7	77.4	81.9	**79.5**
RI, PA	83.3	67.0	74.3	77.9	70.8	74.1	79.6	73.0	76.1	78.6	75.6	77.0	79.1	75.2	**77.1**
RI, PI	83.4	77.3	80.2	81.0	77.3	79.1	82.7	78.6	80.6	82.0	78.6	80.2	81.7	81.2	**81.4**
PA, PI	76.7	68.6	72.4	74.2	69.0	71.5	75.9	71.5	73.6	71.1	70.8	70.9	73.8	76.7	**75.2**
RA, RI, PA	85.2	71.4	77.7	80.8	75.4	78.0	82.5	77.0	79.6	82.6	78.1	**80.3**	82.6	78.2	**80.3**
RA, RI, PI	85.4	79.8	82.5	84.5	80.5	82.4	84.6	81.2	82.9	83.8	84.7	84.2	84.1	85.4	**84.7**
RA, PA, PI	79.6	73.9	76.6	77.5	74.4	75.9	79.2	76.5	77.8	78.9	78.6	78.6	77.4	81.4	**79.2**
RI, PA, PI	83.6	77.5	80.4	81.1	78.0	79.5	82.7	78.6	80.6	82.2	80.9	**81.5**	81.1	81.0	81.1
RA, RI, PA, PI	85.4	79.8	82.5	83.8	80.0	81.8	84.6	81.2	82.9	84.7	81.3	82.9	83.2	86.1	**84.6**

In order to train a forecaster, we created a gold standard, in which, for each topic in CSO, we selected all the timeframes of five years in which the topic had not yet emerged (less than ten patents). We then labelled each of these samples as *True* if the topic produced more than 50 industrial patents in the following ten years and *False* otherwise. The resulting dataset includes 9,776 labelled samples, each composed of four time series: *i)* papers from academia **RA**, *ii)* papers from industry **RI**, *iii)* patents from academia **PA**, *iv)* patents from industry **PI**. We then trained five machine learning classifiers on the gold standard: Logistic Regression (**LR**), Random Forest (**RF**), AdaBoost (**AB**), Convoluted Neural Network (**CNN**), and Long Short-term Memory Neural Network (**LSTM**). We ran each of them on 17 possible combinations of the four time series in order to assess which set of features would yield the best results. **RA-RI** concatenates the time series of research papers in academia (**RA**) and industry (**RI**) together. Instead, R and P sum the contribution of all papers and patents, respectively. Finally, **RA-RI-PA-PI** concatenates the four time series. performed a 10-fold cross-validation of the data and measured the performance of the classifiers by computing the average precision (**P**), recall (**R**), and F1 (**F**).

Table 3 reports the results of our experiment. LSTM outperforms all the other solutions, yielding the highest F1 for 12 of the 17 feature combinations and the highest average F1 (73.7%). CNN (72.8%) and AB (72.3%) also produce competitive results.

As hypothesised, taking advantage of the full set of features available in AIDA – broken into its times series – significantly outperforms (with F1 = 84.6%) the version which uses only the number of patents by companies (75.2%). Considering the origin (academia and industry) of the publications and the patents also increases performance: RA-RI (80.7%) significantly outperforms R (68.2%) and PA-PI (75.2%) is marginally better than P (74.8%).

In conclusion, the experiments confirm that the granular representation of research topics in AIDA can effectively support deep learning approaches for forecasting the impact of research topics on the industrial sector. It also validates the intuition that including features from research articles can be very useful when predicting industrial trends.

5.3 Predicting the spreading of technologies

Every new piece of research, no matter how ground-breaking, adopts previous knowledge and reuses tools and methodologies. Typically, a technology is expected to originate in the context of a research area and then spread and contribute to several other fields. Unfortunately, given the tremendous amount of knowledge produced yearly, it is difficult to digest it and pick up on an interesting piece of research from a different field. Thus, the transfer of a technology from one research area (e.g., Semantic Web) to a different, and possibly conceptually distant, one (e.g., Digital Humanities) may take several years, slowing down the whole research progress. It is, therefore, crucial to be able to anticipate technology spreading across the different research community.

The Technology-Topic Framework (TTF) [52] is a recent approach that uses a semantically enhanced technology-topic model to predict the technologies that will be adopted by a research field. It aims at suggesting promising technologies to scholars from a field, thus helping to accelerate the knowledge flow and the pace of technology propagation. It is based on the hypothesis that *technologies that exhibit similar spreading patterns across multiple research topics will tend to be adopted by similar topics*. It thus characterises technologies in terms of a set of topics drawn from a large-scale ontology of research areas over a given time period and applies machine learning on these data so to forecast technology spreading.

The prototype of TTF takes as input three knowledge bases: *i)* a dataset of 16 million research papers from 1990 to 2013 mainly in the field of Computer Science, described by means of their titles, abstracts, and keywords; *ii)* a list of 1,118 input technologies extracted automatically from paper abstracts and from generic KBs (e.g., DBpedia [5]) and then manually curated, which are associated to at least 10 relevant publications in the research paper dataset; *iii)* 173 relevant research topics and their relationships drawn from the Computer Science Ontology [72,73]. TTF generates a 3-dimensional co-occurrence matrix that tracks the number of papers in which a given technology is associated with a given topic in a given year. It then uses this matrix to train a classifier on a technology recent history in order to predict its future course. The Random Forest classifier performed best on this data, obtaining a precision of 74.4% and a recall of 47.7%, outperforming alternative solutions such as Logistic Regression, Decision Tree, Gradient Boosting, SVM, and a simple Neural Network (see [52] for the full details).

Table 4: Performance of Random Forest on the first 24 topics, with at least 50 positive labels, ordered by F1 score. The table and the experiments were previously reported in [52].

Topics	**Prec. %**	**Rec. %**	**F1 %**	**Topics**	**Prec. %**	**Rec. %**	**F1 %**
information retrieval	92.6	66.8	77.6	wireless networks	64.7	47.8	55.0
database systems	82.6	65.9	73.3	sensor networks	71.9	43.6	54.3
world wide web	88.6	56.1	68.7	software engineering	70.6	44.0	54.2
artificial intelligence	83.6	55.2	66.5	distributed comp. sys.	67.5	45.0	54.0
computer architecture	68.3	63.3	65.7	quality of service	59.6	48.6	53.5
computer networks	82.1	54.0	65.2	imaging systems	100.0	35.8	52.8
image coding	96.8	46.9	63.2	data mining	60.8	45.3	52.0
P2P networks	78.9	50.8	61.9	computer vision	92.3	36.0	51.8
Telecom. traffic	70.8	48.1	57.3	programming languages	65.3	42.0	51.2
wireless telecom. sys.	74.4	46.4	57.1	problem solving	69.0	39.7	50.4
sensors	78.8	43.7	56.2	semantic web	77.8	37.1	50.2
web services	83.3	42.2	56.0	image quality	74.2	37.7	50.0

Table 4 reports the results of the version of TTF using Random Forest on the most frequent 24 topics in the experiment. We can see that TTF obtains the best performance on topics associated with a large set of publications and technologies, such as Information Retrieval and Databases. However, it is also able to obtain satisfactory results in many other fields, especially the ones that are usually associated with a coherent set of technologies, such as Computer Networks, Sensors, and Peer-to-peer Systems. This confirms the hypothesis that is indeed possible to learn from historical spreading patterns and forecast technology propagation, at least for the set of topics that are more involved in technology propagation events.

6 Conclusion

In this chapter, we presented a framework for detecting, analysing, and predicting research topics based on a large-scale knowledge graph characterising research articles according to the research topics from the Computer Science Ontology (CSO) [72,73]. We first illustrated how to annotate a scientific knowledge graph describing research articles and their metadata with a set of research topics from a domain ontology. We then discussed several methods that build on this knowledge graph for analysing research from different perspectives. Finally, we presented two approaches for predicting research trends based on this framework.

References

1. Altuntas, S., Dereli, T., Kusiak, A.: Analysis of patent documents with weighted association rules. Technological Forecasting and Social Change **92**, 249–262 (2015). https://doi.org/10.1016/j.techfore.2014.09.012

2. Angioni, S., Salatino, A., Osborne, F., Recupero, D.R., Motta, E.: The aida dashboard: Analysing conferences with semantic technologies. In: 19th International Semantic Web Conference (ISWC 2020) (2020), http://oro.open.ac.uk/72293/
3. Angioni, S., Salatino, A.A., Osborne, F., Recupero, D.R., Motta, E.: Integrating knowledge graphs for analysing academia and industry dynamics. In: Bellatreche, L., Bieliková, M., Boussaïd, O., Catania, B., Darmont, J., Demidova, E., Duchateau, F., Hall, M., Merčun, T., Novikov, B., Papatheodorou, C., Risse, T., Romero, O., Sautot, L., Talens, G., Wrembel, R., Žumer, M. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium. pp. 219–225. Springer International Publishing, Cham (2020)
4. Aromataris, E., Riitano, D.: Constructing a search strategy and searching for evidence. American Journal of Nursing **114**(5), 49–56 (2014). https://doi.org/10.1097/01.NAJ.0000446779.99522.f6
5. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The semantic web, pp. 722–735. Springer (2007). https://doi.org/10.1007/978-3-540-76298-0_52
6. Augenstein, I., Padó, S., Rudolph, S.: Lodifier: Generating linked data from unstructured text. In: Extended Semantic Web Conference. pp. 210–224. Springer (2012). https://doi.org/10.1007/978-3-642-30284-8_21
7. Beck, M., Rizvi, S.T.R., Dengel, A., Ahmed, S.: From automatic keyword detection to ontology-based topic modeling. In: International Workshop on Document Analysis Systems. pp. 451–465. Springer (2020). https://doi.org/10.1007/978-3-030-57058-3_32
8. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(null), 993–1022 (Mar 2003)
9. Bolelli, L., Ertekin, Ş., Giles, C.L.: Topic and trend detection in text collections using latent dirichlet allocation. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) Advances in Information Retrieval. pp. 776–780. Springer Berlin Heidelberg, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-642-00958-7_84
10. Borges, M.V.M., dos Reis, J.C.: Semantic-enhanced recommendation of video lectures. In: 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT). vol. 2161, pp. 42–46. IEEE (2019). https://doi.org/10.1109/ICALT.2019.00013
11. Bornmann, L., Mutz, R.: Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology **66**(11), 2215–2222 (2015). https://doi.org/10.1002/asi.23329
12. Boyack, K.W., Klavans, R.: Creation of a highly detailed, dynamic, global model and map of science. Journal of the Association for Information Science and Technology **65**(4), 670–685 (2014). https://doi.org/10.1002/asi.22990
13. Boyack, K.W., Klavans, R., Börner, K.: Mapping the backbone of science. Scientometrics **64**(3), 351–374 (2005). https://doi.org/10.1007/s11192-005-0255-6
14. Cano-Basave, A.E., Osborne, F., Salatino, A.A.: Ontology forecasting in scientific literature: Semantic concepts prediction based on innovation-adoption priors. In: Blomqvist, E., Ciancarini, P., Poggi, F., Vitali, F. (eds.) Knowledge Engineering and Knowledge Management. pp. 51–67. Springer International Publishing, Cham (2016)
15. Cena, F., Likavec, S., Osborne, F.: Anisotropic propagation of user interests in ontology-based user models. Information Sciences **250**, 40–60 (2013). https://doi.org/10.1016/j.ins.2013.07.006
16. Chatzopoulos, S., Vergoulis, T., Kanellos, I., Dalamagas, T., Tryfonopoulos, C.: Artsim: improved estimation of current impact for recent articles. In: ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium. pp. 323–334. Springer (2020). https://doi.org/10.1007/978-3-030-55814-7_27

17. Chicaiza, J., Reátegui, R.: Using domain ontologies for text classification. a use case to classify computer science papers. In: Iberoamerican Knowledge Graphs and Semantic Web Conference. pp. 166–180. Springer (2020). https://doi.org/10.1007/978-3-030-65384-2_13
18. Cimiano, P., Völker, J.: Text2onto. In: International Conference on Application of Natural Language to Information Systems. pp. 227–238. Springer (2005). https://doi.org/10.1007/11428817_21
19. Decker, S.L.: Detection of bursty and emerging trends towards identification of researchers at the early stage of trends (2007)
20. Dessì, D., Osborne, F., Recupero, D.R., Buscaldi, D., Motta, E.: Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain. Future Generation Computer Systems **116**, 253–264 (2021)
21. Dessì, D., Osborne, F., Recupero, D.R., Buscaldi, D., Motta, E., Sack, H.: Ai-kg: an automatically generated knowledge graph of artificial intelligence. In: International Semantic Web Conference. pp. 127–143. Springer (2020). https://doi.org/10.1007/978-3-030-62466-8_9
22. Di Caro, L., Guerzoni, M., Nuccio, M., Siragusa, G.: A Bimodal Network Approach to Model Topic Dynamics (2017), `https://arxiv.org/pdf/1709.09373.pdf`
23. Duvvuru, A., Radhakrishnan, S., More, D., Kamarthi, S., Sultornsanee, S.: Analyzing structural & temporal characteristics of keyword system in academic research articles. Procedia Computer Science **20**, 439–445 (2013). https://doi.org/10.1016/j.procs.2013.09.300, complex Adaptive Systems
24. Erten, C., Harding, P.J., Kobourov, S.G., Wampler, K., Yee, G.: Exploring the computing literature using temporal graph visualization. Visualization and Data Analysis 2004 **5295**, 45–56 (2004). https://doi.org/10.1117/12.539245
25. Fenner, M., Aryani, A.: Introducing the pid graph (2019). https://doi.org/10.5438/jwvf-8a66
26. Gangemi, A., Presutti, V., Reforgiato Recupero, D., Nuzzolese, A.G., Draicchio, F., Mongiovì, M.: Semantic web machine reading with fred. Semantic Web **8**(6), 873–893 (2017). https://doi.org/10.3233/SW-160240
27. Garcia-Silva, A., Gomez-Perez, J.M.: Classifying scientific publications with bert – is self-attention a feature selection method? (2021)
28. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences **101**(suppl 1), 5228–5235 (2004). https://doi.org/10.1073/pnas.0307752101
29. He, Q., Chen, B., Giles, C.L.: Detecting Topic Evolution in Scientific Literature : How Can Citations Help ? Cikm pp. 957–966 (2009). https://doi.org/10.1145/1645953.1646076
30. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Gayo, J.E.L., Kirrane, S., Neumaier, S., Polleres, A., et al.: Knowledge graphs. arXiv preprint arXiv:2003.02320 (2020)
31. Jaradeh, M.Y., Oelen, A., Farfar, K.E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., Auer, S.: Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In: Proceedings of the 10th International Conference on Knowledge Capture. p. 243–246. K-CAP '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3360901.3364435
32. Jo, Y., Lagoze, C., Giles, C.L.: Detecting research topics via the correlation between graphs and texts. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 370–379. KDD '07, Association for Computing Machinery, New York, NY, USA (2007). https://doi.org/10.1145/1281192.1281234
33. Kandimalla, B., Rohatgi, S., Wu, J., Giles, C.L.: Large scale subject category classification of scholarly papers with deep attentive neural networks. Frontiers in Research Metrics and Analytics **5**, 31 (2021). https://doi.org/10.3389/frma.2020.600382
34. 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 **11**(6), 979–1005 (2020). https://doi.org/10.3233/SW-200371

35. Kitchenham, B.: Procedures for performing systematic reviews. Keele, UK, Keele University **33**(2004), 1–26 (2004)
36. Kitchenham, B.A., Charters, S.: Guidelines for performing systematic literature reviews in software engineering (2007)
37. Krampen, G., von Eye, A., Schui, G.: Forecasting trends of development of psychology from a bibliometric perspective. Scientometrics **87**, 687–694 (2011). https://doi.org/10.1007/s11192-011-0357-2
38. La Bruzzo, S., Manghi, P., Mannocci, A.: OpenAIRE's DOIBoost - Boosting Crossref for Research. In: Manghi, P., Candela, L., Silvello, G. (eds.) Digital Libraries: Supporting Open Science. pp. 133–143. Springer International Publishing, Cham (jan 2019). https://doi.org/10.1007/978-3-030-11226-4_11
39. Leydesdorff, L., Rafols, I., Chen, C.: Interactive overlays of journals and the measurement of interdisciplinarity on the basis of aggregated journal–journal citations. Journal of the American Society for Information Science and Technology **64**(12), 2573–2586 (2013). https://doi.org/10.1002/asi.22946
40. Likavec, S., Osborne, F., Cena, F.: Property-based semantic similarity and relatedness for improving recommendation accuracy and diversity. International Journal on Semantic Web and Information Systems (IJSWIS) **11**(4), 1–40 (2015). https://doi.org/10.4018/IJSWIS.2015100101
41. Liu, X., Song, Y., Liu, S., Wang, H.: Automatic taxonomy construction from keywords. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1433–1441. ACM (2012). https://doi.org/10.1145/2339530.2339754
42. Löffler, F., Wesp, V., Babalou, S., Kahn, P., Lachmann, R., Sateli, B., Witte, R., König-Ries, B.: Scholarlensviz: A visualization framework for transparency in semantic user profiles. In: Taylor, K., Gonçalves, R., Lecue, F., Yan, J. (eds.) Proceedings of the ISWC 2020 Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 19th International Semantic Web Conference (ISWC 2020), Globally online, November 1-6, 2020 (UTC). (2020)
43. Lula, P., Dospinescu, O., Homocianu, D., Sireteanu, N.A.: An advanced analysis of cloud computing concepts based on the computer science ontology. Computers, Materials & Continua **66**(3), 2425–2443 (2021). https://doi.org/10.32604/cmc.2021.013771
44. Lv, P.H., Wang, G.F., Wan, Y., Liu, J., Liu, Q., Ma, F.C.: Bibliometric trend analysis on global graphene research. Scientometrics **88**, 399–419 (2011). https://doi.org/10.1007/s11192-011-0386-x
45. Mai, F., Galke, L., Scherp, A.: Using deep learning for title-based semantic subject indexing to reach competitive performance to full-text. In: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries. p. 169–178. JCDL '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3197026.3197039
46. Manghi, P., Atzori, C., Bardi, A., Baglioni, M., Schirrwagen, J., Dimitropoulos, H., La Bruzzo, S., Foufoulas, I., Löhden, A., Bäcker, A., Mannocci, A., Horst, M., Jacewicz, P., Czerniak, A., Kiatropoulou, K., Kokogiannaki, A., De Bonis, M., Artini, M., Ottonello, E., Lempesis, A., Ioannidis, A., Manola, N., Principe, P.: Openaire research graph dump (Nov 2020). https://doi.org/10.5281/zenodo.4279381
47. Mannocci, A., Osborne, F., Motta, E.: The evolution of ijhcs and chi: A quantitative analysis. International Journal of Human-Computer Studies **131**, 23–40 (2019). https://doi.org/10.1016/j.ijhcs.2019.05.009
48. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013)
49. Nayyeri, M., Cil, G.M., Vahdati, S., Osborne, F., Rahman, M., Angioni, S., Salatino, A., Recupero, D.R., Vassilyeva, N., Motta, E., Lehmann, J.: Trans4e: Link prediction on scholarly knowledge graphs. Neurocomputing (2021)

50. Nuzzolese, A.G., Gentile, A.L., Presutti, V., Gangemi, A.: Conference linked data: The scholarlydata project. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 9982 LNCS, pp. 150–158 (2016). https://doi.org/10.1007/978-3-319-46547-0_16
51. Osborne, F., De Ribaupierre, H., Motta, E.: Techminer: extracting technologies from academic publications. In: European Knowledge Acquisition Workshop. pp. 463–479. Springer (2016)
52. Osborne, F., Mannocci, A., Motta, E.: Forecasting the spreading of technologies in research communities. In: Proceedings of the Knowledge Capture Conference. K-CAP 2017, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3148011.3148030
53. 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). https://doi.org/10.1007/978-3-319-25007-6_24
54. Osborne, F., Motta, E.: Pragmatic ontology evolution: reconciling user requirements and application performance. In: International Semantic Web Conference. pp. 495–512. Springer (2018)
55. Osborne, F., Motta, E., Mulholland, P.: Exploring scholarly data with rexplore. In: Alani, H., Kagal, L., Fokoue, A., Groth, P., Biemann, C., Parreira, J.X., Aroyo, L., Noy, N., Welty, C., Janowicz, K. (eds.) The Semantic Web – ISWC 2013. pp. 460–477. Springer Berlin Heidelberg, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41335-3_29
56. Osborne, F., Muccini, H., Lago, P., Motta, E.: Reducing the effort for systematic reviews in software engineering. Data Science **2**(1-2), 311–340 (2019). https://doi.org/10.3233/DS-190019
57. 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)
58. Osborne, F., Scavo, G., Motta, E.: A hybrid semantic approach to building dynamic maps of research communities. In: International Conference on Knowledge Engineering and Knowledge Management. pp. 356–372. Springer (2014)
59. Osborne, F., Scavo, G., Motta, E.: Identifying diachronic topic-based research communities by clustering shared research trajectories. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) The Semantic Web: Trends and Challenges. pp. 114–129. Springer International Publishing, Cham (2014)
60. Peroni, S., Osborne, F., Di Iorio, A., Nuzzolese, A.G., Poggi, F., Vitali, F., Motta, E.: Research articles in simplified html: a web-first format for html-based scholarly articles. PeerJ Computer Science **3**, e132 (2017). https://doi.org/10.7717/peerj-cs.132
61. Peroni, S., Shotton, D.: OpenCitations, an infrastructure organization for open scholarship. Quantitative Science Studies **1**(1), 428–444 (2020). https://doi.org/10.1162/qss_a_00023
62. Petersen, K., Vakkalanka, S., Kuzniarz, L.: Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology **64**, 1–18 (2015). https://doi.org/10.1016/j.infsof.2015.03.007
63. Petrucci, G., Ghidini, C., Rospocher, M.: Ontology learning in the deep. In: European Knowledge Acquisition Workshop. pp. 480–495. Springer (2016). https://doi.org/10.1007/978-3-319-49004-5_31
64. Ramadhan, M.H., Malik, V.I., Sjafrizal, T.: Artificial neural network approach for technology life cycle construction on patent data. In: 2018 5th International Conference on Industrial Engineering and Applications (ICIEA). pp. 499–503 (2018). https://doi.org/10.1109/IEA.2018.8387152

65. Rossanez, A., dos Reis, J.C., da Silva Torres, R.: Representing scientific literature evolution via temporal knowledge graphs (2020)
66. Salatino, A., Osborne, F., Motta, E.: Researchflow: Understanding the knowledge flow between academia and industry. In: Keet, C.M., Dumontier, M. (eds.) Knowledge Engineering and Knowledge Management. pp. 219–236. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-61244-3_16
67. Salatino, A., Thanapalasingam, T., Mannocci, A., Osborne, F., Motta, E.: Classifying research papers with the computer science ontology (2018)
68. 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)
69. 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). https://doi.org/10.7717/peerj-cs.119
70. 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. p. 303–312. JCDL '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3197026.3197052
71. 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). https://doi.org/10.1007/978-3-030-30760-8_26
72. 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 **2**(3), 379–416 (2020). https://doi.org/10.1162/dint_a_00055
73. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Osborne, F., Motta, E.: The computer science ontology: A large-scale taxonomy of research areas. 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. 187–205. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-00668-6_12
74. 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 (2011). https://doi.org/10.1109/ICDCSW.2011.20
75. Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.J.P., Wang, K.: An overview of microsoft academic service (mas) and applications. In: Proceedings of the 24th International Conference on World Wide Web. p. 243–246. WWW '15 Companion, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2740908.2742839
76. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: Extraction and mining of academic social networks. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 990–998 (2008). https://doi.org/10.1145/1401890.1402008
77. 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)
78. Tseng, Y.H., Lin, Y.I., Lee, Y.Y., Hung, W.C., Lee, C.H.: A comparison of methods for detecting hot topics. Scientometrics **81**(1), 73–90 (2009). https://doi.org/10.1007/s11192-009-1885-x
79. Van Raan, A.F.: Sleeping Beauties in science. Scientometrics **59**(3), 467–472 (2004). https://doi.org/10.1023/B:SCIE.0000018543.82441.f1

80. Vergoulis, T., Chatzopoulos, S., Dalamagas, T., Tryfonopoulos, C.: Veto: Expert set expansion in academia. In: Hall, M., Merčun, T., Risse, T., Duchateau, F. (eds.) Digital Libraries for Open Knowledge. pp. 48–61. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-54956-5_4
81. Visser, M., van Eck, N.J., Waltman, L.: Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic (04 2021). https://doi.org/10.1162/qss_a_00112
82. Wang, K., Shen, Z., Huang, C., Wu, C.H., Dong, Y., Kanakia, A.: Microsoft Academic Graph: When experts are not enough. Quantitative Science Studies **1**(1), 396–413 (2020). https://doi.org/10.1162/qss_a_00021
83. Watatani, K., Xie, Z., Nakatsuji, N., Sengoku, S.: Global competencies of regional stem cell research: bibliometrics for investigating and forecasting research trends. Regenerative medicine **8**(5), 659–68 (2013). https://doi.org/10.2217/rme.13.51
84. Widodo, A., Fanany, M.I., Budi, I.: Technology forecasting in the field of apnea from online publications: Time series analysis on latent semantic. In: 2011 Sixth International Conference on Digital Information Management. pp. 127–132 (2011). https://doi.org/10.1109/ICDIM.2011.6093365
85. Wohlin, C., Prikladnicki, R.: Systematic literature reviews in software engineering. Information and Software Technology **55**(6), 919–920 (2013). https://doi.org/10.1016/j.infsof.2013.02.002
86. Zhang, X., Chandrasegaran, S., Ma, K.L.: Conceptscope: Organizing and visualizing knowledge in documents based on domain ontology. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–13 (2021)
87. Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., Zhang, G.: Does deep learning help topic extraction? a kernel k-means clustering method with word embedding. Journal of Informetrics **12**(4), 1099–1117 (2018). https://doi.org/10.1016/j.joi.2018.09.004