AIDA: a Knowledge Graph about Research Dynamics in Academia and Industry

Angioni, S., Salatino, A., Osborne, F., Recupero, D., and Motta, E. (2021) AIDA: a Knowledge Graph about Research Dynamics in Academia and Industry. Quantitative Science Studies. Advance Publication. https://doi.org/10.1162/qss_a_00162

RESEARCH

AIDA: a Knowledge Graph about Research Dynamics in Academia and Industry

Simone Angioni[1] and Angelo Salatino[2] and Francesco Osborne[2] and Diego Reforgiato Recupero[1] and Enrico Motta[2]

[1]Department of Mathematics and Computer Science, University of Cagliari (Italy)

[2]Knowledge Media Institute, The Open University, Milton Keynes (UK)

an open access journal

Citation:

DOI:
Supporting Information:
Received:
Accepted:
Published:

Competing Interests: The authors have declared that no competing interests exist.

Corresponding Author:
Simone Angioni
simone.angioni@unica.it

Handling Editor:

ABSTRACT

Academia and industry share a complex, multifaceted, and symbiotic relationship. Analysing the knowledge flow between them, understanding which directions have the biggest potential, and discovering the best strategies to harmonise their efforts is a critical task for several stakeholders. Research publications and patents are an ideal medium to analyze this space, but current datasets of scholarly data cannot be used for such a purpose since they lack a high-quality characterization of the relevant research topics and industrial sectors. In this paper, we introduce the Academia/Industry DynAmics (AIDA) Knowledge Graph, which describes 21M publications and 8M patents according to the research topics drawn from the Computer Science Ontology. 5.1M publications and 5.6M patents are further characterized according to the type of the author's affiliations and 66 industrial sectors from the proposed Industrial Sectors Ontology (INDUSO). AIDA was generated by an automatic pipeline that integrates data from Microsoft Academic Graph, Dimensions, DBpedia, the Computer Science Ontology, and the Global Research Identifier Database. It is publicly available under CC BY 4.0 and can be downloaded as a dump or queried via a triplestore. We evaluated the different parts of the generation pipeline on a manually crafted gold standard yielding competitive results.

Keywords: Scholarly Data, Knowledge Graph, Topic Detection, Bibliographic Data, Scholarly Ontologies, Research Dynamics

1 INTRODUCTION

Academia and industry share a complex, multifaceted, and symbiotic relationship. Their collaboration and exchange of ideas, resources, and persons (Anderson, 2001a) are conducive to the production of new knowledge that will ultimately shape the society of the future. Analyzing the knowledge flow between academia and industry, understanding which directions have the biggest potential, and discovering the best strategies to harmonize their efforts is thus a critical task for several stakeholders (A. Salatino, Osborne, & Motta, 2020). Governments and funding agencies need to regularly assess the potential impact of research areas and technologies to inform funding decisions. Commercial organizations have to monitor research developments and adapt to technological advancements. Researchers must keep up with the latest trends and be aware of complementary research efforts from the industrial sector.

The relationship between academia and industry has been analyzed from several perspectives in the literature, focusing for instance on the characteristics of direct collaborations (S. Ankrah & Omar, 2015), the influence of industrial trends on curricula (Weinstein, Kellar, & Hall, 2016), and the quality of the knowledge transfer (S. N. Ankrah, Burgess, Grimshaw, & Shaw, 2013). However, most of the quantitative studies on this relationship were limited to small-scale datasets or focused on very specific research questions (Anderson, 2001a; Bikard, Vakili, & Teodoridis, 2019).

Research articles and patents are an ideal medium to analyze the knowledge generated and developed by academia and industry (S. Ankrah & Omar, 2015; S. N. Ankrah et al., 2013). Today, we have several large-scale knowledge graphs which describe research papers according to their titles, abstracts, authors, organizations, and other metadata. Examples include Microsoft Academic Graph[1] (K. Wang et al., 2020), Scopus[2], Semantic Scholar[3], Aminer (Y. Zhang, Zhang, Yao, & Tang, 2018), CORE (Knoth & Zdrahal, 2012), OpenCitations (Peroni & Shotton, 2020), and others. Other resources, such as Dimensions[4], the United States Patent and Trademark Office (USPTO)[5], the Espacenet dataset[6], and the PatentScope corpus[7], offer a similar description of patents. However, these datasets cannot be directly used to analyze the research dynamics of academia and industry since they lack a high-quality characterization of the relevant research topics and industrial sectors.

In particular, they suffer from three main limitations. First, current solutions do not allow us to easily discriminate if a document (research paper or patent) is from academia or industry. Second, they typically offer a coarse-grained characterization of research topics, which are usually represented only as a list of terms chosen by the authors or extracted from the abstract. This purely syntactic solution is unsatisfactory (Osborne & Motta, 2015), as it fails: i) to distinguish research topics from other generic keywords; ii) to deal with situations where multiple labels exist for the same research area; and iii) to model and take advantage of the semantic relationships that hold between research areas. For instance, we want to be able to infer that all documents tagged with the topic Neural Network are also about Machine Learning and Artificial Intelligence. This richer representation would allow us to retrieve all the publications which address the concept Artificial Intelligence, even if the metadata does not contain the specific string "artificial intelligence". A third issue is that current scholarly datasets do not characterize companies according to their sectors. Therefore, it is not possible to measure the impact of a topic (e.g., sentiment analysis, deep learning, semantic web) on different types of industry (e.g., automotive, financial, energy).

These limitations affect also the performance of machine learning systems, typically based on neural networks, for predicting the impact of research trends and forecasting patents (Choi & Jun, 2014; Marinakis, 2012; Ramadhan, Malik, & Sjafrizal, 2018; Zang & Niu, 2011). These solutions typically work with limited features, such as the number of patents associated with a topic for each year, since current datasets do not integrate articles and patents, lack a granular representation of research topics, and cannot distinguish whether a document was produced by academia or industry. We hypothesize that consid-

[1] Microsoft Academic Graph - http://aka.ms/microsoft-academic
[2] Scopus - https://www.scopus.com/
[3] Semantic Scholar - https://www.semanticscholar.org/
[4] Dimensions - https://www.dimensions.ai/
[5] USPTO - https://www.uspto.gov/
[6] Espacenet dataset - https://worldwide.espacenet.com/
[7] PatentScope - https://patentscope.wipo.int/

ering a richer characterization of this space would ultimately yield better performance in comparison to state-of-the-art approaches.

In this paper, we introduce the Academia/Industry DynAmics (AIDA) Knowledge Graph, which describes 21M publications and 8M patents in the field of *Computer Science*. Papers and patents are associated to the research topics in the Computer Science Ontology (CSO). In addition, 5.1M publications and 5.6M patents are also characterized according to the type of the author's affiliations (e.g., academia, industry) and 66 industrial sectors (e.g., automotive, financial, energy, electronics) from the Industrial Sectors Ontology (INDUSO). AIDA is also linked to several other knowledge bases, including MAKG, Dimensions, Google Patents, GRID, DBpedia, and Wikidata.

AIDA is available at `http://w3id.org/aida/`. It can be downloaded as a dump or queried via a Virtuoso triplestore at `http://w3id.org/aida/sparql/`. We plan to release a new version of AIDA every six months, to regularly update the publications, the topics, and the industrial sectors.

AIDA was generated using an automatic pipeline that integrates data from Microsoft Academic Graph (MAG)[8], Dimensions, English DBpedia, the Computer Science Ontology (CSO), and the Global Research Identifier Database (GRID), respectively containing information about 242M research papers, 38M patents, 4.58M entities, 14K research topics, and 97K organizations.

The resulting knowledge base enables analyzing the evolution of research topics across academia and industry and studying the characteristics of several industrial sectors. For instance, it enables detecting the research trends most interesting for the automotive sector or which prevalent industrial topics were recently adopted by academia. It can thus be utilized by a variety of deep learning methods for predicting the impact of research trends on industry and academia (Chung & Sohn, 2020; Ramadhan et al., 2018; Zang & Niu, 2011). It can also be used to characterize authors, citations, countries, and several other entities in MAG according to their topics and industrial sectors. This makes it possible to study further dynamics such as the migration of researchers and the citation flow between academia and the industry.

We evaluated the different parts of the pipeline for generating AIDA on manually crafted gold standards yielding competitive results. We also report an evaluation of the impact of AIDA on forecasting systems for predicting the impact of research topics on the industry. Specifically, we tested five classifiers on 17 combinations of features and found that the forecaster based on Long Short-Term Memory neural networks and exploiting the full set of features from AIDA obtain significantly better performance ($p<0.0001$) than alternative methods.

A preliminary version of AIDA which included a smaller data set and a limited number of semantic relations was previously discussed in a short workshop paper (Angioni, Salatino, Osborne, Recupero, & Motta, 2020). The current paper greatly expands on that work by presenting a novel and up-to-date version of AIDA (including about 5M additional articles), an improved version of the pipeline for generating AIDA, a more extensive ontological schema, and a comprehensive evaluation of AIDA.

[8] We used the dump released in April 2020.

In summary, our main contributions include:

- the first official release of AIDA, a knowledge graph for studying the research dynamics of academia and industry;
- a pipeline for automatically generating AIDA based on a robust semantic model and a state-of-the-art topic detection approach;
- a detailed discussion of AIDA schema, content, and links to other knowledge graphs;
- an evaluation of the AIDA pipeline and its ability to classify documents in terms of research topics and industrial sectors;
- an illustrative overview of the Computer Science domain according to the data in AIDA.
- a discussion of AIDA possible usage that summarizes some research efforts that adopted preliminary versions of AIDA;
- an analysis of the current limitations of the AIDA pipeline and a sustainability plan developed in collaboration with Springer Nature for replacing MAG with a combination of Dimensions and DBLP, after MAG will be decommissioned at the end of 2021;
- an appendix detailing several exemplary SPARQL queries in order to support the reuse of AIDA.

The rest of the paper is organized as follows. In Section 2, we review the literature on methods and datasets for studying and quantifying the relationship between academia and industry. In Section 3, we describe the pipeline to generate AIDA, give an overview of the resulting knowledge graph, and discuss our strategy for releasing new versions. Section 4 presents the evaluation of the different parts of the AIDA pipeline and the experiments showing that AIDA can support effectively deep learning approaches for predicting the impact of research topics. In Section 5 we focus on the usage of AIDA and report three exemplary research efforts that adopted preliminary versions of AIDA: i) a bibliometric analysis of the research dynamics across academia and industry, ii) a study of the main research trends in two main venues of Human-Computer Interaction, and iii) a new web application that we developed to support Springer Nature editors in assessing the quality of scientific conferences. Section 6 describes the main limitations of the proposed pipeline and how we will address them going forward. Finally, in Section 7 we summarise the main conclusions and outline future directions of research.

2 LITERATURE REVIEW

In this section, we review the current state of the art regarding knowledge graphs describing research papers and patents (Section 2.1) and approaches for analyzing the relationships between industry and academia (Section 2.2).

2.1 Knowledge Graphs of Research Articles and Patents

Knowledge graphs are graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities (Hogan et al., 2021). Such descriptions have formal semantics allowing both computers and people to process them efficiently and unambiguously. Knowledge Graphs about research articles and patents typically describe the relevant actors (e.g., authors, organisations), entities (e.g., topics, tasks, technologies), as well as any other contextual information (e.g., project, funding) in an interlinked manner.

In the last years, we saw the emergence of several knowledge graphs describing research publications and their metadata.

Microsoft Academic Graph (MAG) (K. Wang et al., 2020) is a heterogeneous knowledge graph, that contains the metadata of more than 248M scientific publications, including citations, authors, institutions, journals, conferences, and fields of study. Microsoft Academic Knowledge Graph (MAKG)[9] (Färber, 2019) is a large RDF dataset based on MAG that also provides entity embeddings for the research papers.

The Semantic Scholar Open Research Corpus[10] (Ammar et al., 2018) is a dataset of about 185M publications released by Semantic Scholar, an academic search engine provided by the Allen Institute for Artificial Intelligence (AI2). The OpenCitations Corpus (Peroni & Shotton, 2020) is released by OpenCitations, an independent infrastructure organization for open scholarship dedicated to the publication of open bibliographic and citation data with semantic technologies. The current version includes 55M publications and 655M citations. Scopus is a well-known dataset curated by Elsevier, which includes about 70M publications and is often used by governments and funding bodies to compute performance metrics. The AMiner Graph (Y. Zhang et al., 2018) is the corpus of more than 200M publications generated and used by the AMiner system[11]. AMiner is a free online academic search and mining system that also extracts researchers' profiles from the Web and integrates them into the metadata. The Open Academic Graph (OAG)[12] is a large knowledge graph integrating Microsoft Academic Graph and AMiner Graph. The current version contains 208M papers from MAG and 172M from AMiner. CORE (Knoth & Zdrahal, 2011)[13] is a repository that integrates 24M open access research outputs from repositories and journals worldwide. The Dimensions corpus is a dataset produced by Digital Science that integrates and interlinks 109M research publications, 5.3M grants, and 40M patents. Publications and citations are freely available for personal, non-commercial use.

DBLP (Ley, 2009) is a very well-curated bibliographic database of conferences, workshops, and journals in Computer Science. It currently covers 5.7M articles, 5,443 conferences, and 1,773 journals. The ACL Anthology Reference Corpus (Bird et al., 2008) is a digital archive of conference and journal papers in natural language processing and computational linguistics, which aims to serve as a reference repository of research results. UnarXive (Saier & Färber, 2020) is a dataset including over one million publications from arXiv.org for which it provides the full text and in-text citations annotated via global identifiers. AceKG (R. Wang et al., 2018) is a large-scale KG which provides 3 billion triples of academic facts about papers, authors, fields of study, venues and institutes, as well as the relations among them. It was designed as benchmark dataset for challenging data mining tasks, including link prediction, community detection, and scholar classification. DOIboost (La Bruzzo, Manghi, & Mannocci, 2019) provides an enhanced version of Crossref[14] that integrates information from Unpaywall, ORCID and MAG, such as author identifiers, affiliations, organisation identifiers, and abstracts. It is periodically released on Zenodo[15].

[9] MAKG - https://makg.org/
[10] ORC - http://s2-public-api-prod.us-west-2.elasticbeanstalk.com/corpus/
[11] AMiner - https://www.aminer.cn/
[12] Open Academic Graph - https://www.openacademic.ai/oag/
[13] CORE - https://core.ac.uk/
[14] Crossref - https://www.crossref.org/
[15] DOIboost laster release - https://zenodo.org/record/3559699

Several other knowledge graph and resources focus specifically on patents (Schwartz & Sichelman, 2019). For instance, the European Patent Office (EPO) curates the Espacenet dataset, which currently covers about 110 million patents from all over the world. Similarly, the United States Patent and Trademark Office produces a corpus that includes more than 14M US patents. The World Intellectual Property Organization (WIPO) offers the PatentScope dataset, which contains 84M patent documents, including 4M international patent applications

Deng at al. (Deng, Huang, & Zhu, 2019) propose a method based on conditional random field for automatically generating KGs describing technologies extracted from a set of patents. However, the approach was only tested on about 5,000 patents and the resulting knowledge base was not made available. TechNet (Saricaa, Luoab, & Woodab, 2019) [16] is a semantic networks which includes 4M terms extracted from 5.8M patents in the U.S. patents database. Specifically, the authors created an NLP approach to mine generic engineering terms and used their word embeddings to assess their semantic similarity.

Another category of knowledge graphs offer a semantic representation of the content of scientific articles. The Semantic Web community has been working for a while on this direction, fostering the Semantic Publishing paradigm (Shotton, 2009), creating bibliographic repositories in the Linked Data Cloud (Nuzzolese, Gentile, Presutti, & Gangemi, 2016), generating knowledge bases of biological data (Belleau, Nolin, Tourigny, Rigault, & Morissette, 2008), formalising research workflows (Wolstencroft et al., 2013), implementing systems for managing nano-publications (Groth, Gibson, & Velterop, 2010; T. Kuhn et al., 2016) and micropublications (Schneider, Ciccarese, Clark, & Boyce, 2014), and developing a variety of ontologies to describe scholarly data, e.g., SWRC[17], BIBO[18], BiDO[19], FABIO[20], SPAR[21] (Peroni & Shotton, 2018), and SKGO[22] (Fathalla, Auer, & Lange, 2020).

A recent example is the Open Research Knowledge Graph (ORKG) (Jaradeh et al., 2019)[23], which aims to describe research papers in a structured manner to make them easier to find and compare.

Several of these knowledge bases focus on describing the research areas of scientific publications. These include the Medical Subject Heading (MeSH)[24] in Biology, Mathematics Subject Classification (MSC)[25] in Mathematics, Physics Subject Headings (PhySH)[26] in Physics, and many others.

In the field of Computer Science, the best-known taxonomies of research areas are the ACM Computing Classification System[27] and the Computer Science Ontology (CSO) (A. A. Salatino, Thanapalasingam, Mannocci, Osborne, & Motta, 2018b). The first one is developed and maintained by the Association for Computing Machinery (ACM). It

[16] TechNet - http://www.tech-net.org/
[17] SWRC - http://ontoware.org/swrc
[18] BIBO - http://bibliontology.com
[19] BiDO - http://purl.org/spar/bido
[20] FABIO - http://purl.org/spar/fabio
[21] SPAR - http://www.sparontologies.net/
[22] SKGO - https://github.com/saidfathalla/Science-knowledge-graph-ontologies
[23] ORKG - https://www.orkg.org/orkg/
[24] Medical Subject Heading - https://www.ncbi.nlm.nih.gov/mesh
[25] Mathematics Subject Classification - https://mathscinet.ams.org/msc
[26] Physics Subject Headings - https://physh.aps.org/
[27] ACM Classification System - https://www.acm.org/publications/class-2012

contains around 2K concepts and it is manually curated. Conversely, CSO is automatically generated from a large collection of publications by the Open University and includes about 14K research areas. We adopted CSO for AIDA because it is one order of magnitude larger than the alternatives and it comes with the CSO Classifier (A. A. Salatino, Osborne, Thanapalasingam, & Motta, 2019; A. A. Salatino, Thanapalasingam, & Mannocci, 2019), which is a tool for automatically annotating documents with CSO topics. Hence, it allows us to easily generate a granular representation of all the documents integrated from MAG and Dimensions.

Currently, there are no datasets that enable the study of fine-grained research topics and their relation with industrial sectors across research papers and patents.

For this reason, we decided to undertake this new endeavor and develop AIDA.

We decided to adopt MAG over the alternatives knowledge graphs of articles for two main reasons. First, it appears to be the most comprehensive among the publicly available datasets of publications (Visser, van Eck, & Waltman, 2020). Second, it associates articles with DOIs and organizations with GRID identifiers and therefore can be easily integrated with other knowledge graphs.

For patents, we chose Dimensions because of its comprehensiveness and also because it identifies organizations with GRID IDs, allowing us to easily integrate them with MAG affiliations.

After the first version of this manuscript, Microsoft announced that MAG will be decommissioned in 2022. For this reason, we formulated a plan in collaboration with Springer Nature for using a combination of Dimensions and DBLP as source for research publications in the following versions of AIDA. This plan is presented in Section 6.

2.2 Relationship between Academia and Industry

Academia and industry typically tend to influence each other by exchanging ideas, resources, and researchers (Powell & Snellman, 2004). Analyzing their relationship allows us to understand their role within the whole knowledge economy (Anderson, 2001b): from production, towards adoption, enrichment, and ultimately deployment as a new commercial product or service. In some cases, academia and industry engage in collaborations as an opportunity for a more productive division of tasks: academia focusing on scientific insights, and industry on commercialization (Bikard et al., 2019). Stilgoe (2020) discusses the main drivers of scientific innovation and focuses on the central role of the industry sector in pushing innovation by constantly deploying new technologies. However, it can be argued that innovation advances also through a more complex route, which involves the birth of a new scientific area, the development of its theoretical framework, and the creation of innovative products that capitalize on the new knowledge (T. S. Kuhn, 1962).

The knowledge transfer between academia and industry has been studied according to both qualitative (Grimpe & Hussinger, 2013; Michaudel, Ishihara, & Baran, 2015) and quantitative methods (Huang, Yang, & Chen, 2015; Larivière, Macaluso, Mongeon, Siler, & Sugimoto, 2018). A good example of the first category is Michaudel et al. (2015) who share their personal experience on how the collaboration between industry and academia impacted their research program. Similarly, Grimpe and Hussinger (2013) perform a survey-based analysis to understand the innovation performance associated with collaborations between universities and German manufacturers. In the category of quantitative approaches, Larivière et al. (2018) employ both research papers and patents to understand the primary interests of both sides in this symbiosis. Huang et al. (2015) also take a quantitative approach and analyse 20K research papers and 8K patents in the area of *fuel cells* to assess the direct benefits of collaborations between academia and industry.

Hanieh, AbdElall, Krajnik, and Hasan (2015) argue that a partnership agreement between industry and academia aims at enhancing economic prosperity, social equity, and environmental protection. This partnership includes also carrying out scientific research activities and solving industrial problems. In their paper, the authors analyse the state of affairs in Palestine, showing that such a cooperation is weak, and hence they advocate to improve this partnership. Also, they suggest to develop curricula by including sustainability concepts and improving teaching methods.

However, these approaches focus on relatively narrow areas of science and do not use a granular characterization of research areas. Conversely, AIDA allows researchers to analyse the interaction of research topics and industrial sectors across millions of documents. The resulting data can support a variety of studies that are not feasible with current knowledge bases. For instance, AIDA makes it possible to analyse how industrial sectors (e.g., automotive) contribute to specific research fields (e.g., AI, Robotics) and how certain research lines lead to the development of concrete commercial services. It also enables to quantify the impact of a field on industry across the years, in order to better assess the concrete fallback of scientific research.

3 AIDA: ACADEMIA INDUSTRY DYNAMICS KNOWLEDGE GRAPH

The Academia/Industry DynAmics (AIDA) Knowledge Graph includes about 1.3B triples that describe a large collection of publications and patents in *Computer Science* according to their research topics, industrial sectors, and author's affiliations (academia, industry, or collaborative). Specifically, 21M publications from MAG and 8M patents from Dimensions are classified according to the research topics drawn from the Computer Science Ontology (CSO). On average, each publication is associated with 27 ± 19 topics and each patent with 33 ± 14[28].

The 5.1M publications and 5.6M patents that were associated with GRID IDs in the original data are also classified according to the type of the author's affiliations (e.g., academia, industry) and 66 industrial sectors (e.g., automotive, financial, energy, electronics) drawn from the Industrial Sectors ontology (INDUSO)[29], which was specifically designed to support AIDA.

Since these annotations require at least an affiliation of the authors of the document to be associated with a GRID ID (as detailed in Section 3.1), they are currently restricted only to the document linked to GRID by Microsoft Academics Graph and Dimensions.

About 4.5M articles and 4.9M patents were also typed with the three main categories of our schema: academia, industry, and collaboration (between academia and industry). We also included additional affiliation categories from GRID, such as ”Government”, ”Facility”, ”Healthcare”, and ”Nonprofit”.

[28] With $x \pm y$ we refer to x being the average and y the standard deviation.

[29] INDUSO - `http://w3id.org/aida/downloads/induso.ttl`

AIDA was generated and it will be regularly updated by an automatic pipeline that integrates and enriches data from Microsoft Academic Graph (MAG), Dimensions, English DBpedia, the Global Research Identifier Database (GRID), CSO, and INDUSO.

Table 1: AIDA - Affiliation Types.

	Publications	**Patents**
Academia	3,906,131	122,390
Industry	834,443	4,760,614
Collaborative	133,781	16,806
Additional categories in GRID	627,179	747,618
Documents with GRID ID	5,133,171	5,639,252
Total documents	20,850,710	7,940,034

Table 1 shows the number of publications and patents from academia, industry, and collaborative efforts. Please note that only the documents associated with a GRID ID (about 5.1M publications and 5.6M patents) can be classified as academia, industry, collaborative or any other additional category from GRID.

When considering the affiliation types, most publications (69.8%) are written by academic institutions, however, the industry contributes to a good number of them (15.3%). The situation is reversed when considering patents: 84% of them are from industry and only 2.3% from academia. Another interesting finding is that the collaborative efforts are limited, involving only 2.6% of the publications and 0.2% of the patents. These numbers require further analysis but may suggest that we need to improve the mechanisms to support and fund collaborative works.

The data model of AIDA builds on AIDA Schema, Schema.org, FOAF, OWL, CSO and others. We created AIDA Schema to define all the specific relations that could not be reused from state-of-the-art ontologies. It is available at `http://w3id.org/aida/ontology`.

Figure 1 depicts the full data model of AIDA KG, including both relations that we defined within AIDA schema and the ones we imported from external schemas. It focuses on six types of entities (light-blue boxes in Figure 1): papers, patents, authors, affiliations, industrial sectors, and DBpedia categories. To be compatible with other knowledge graphs in this space (e.g., MAG, Scopus, DBLP, Semantic Scholar), papers are identified according to their Digital Object Identifier (DOI) and patents according to their World Intellectual Property Organization (WIPO) ID. We also retain the original MAG IDs for papers and authors as additional identifiers. These are used to link AIDA to MAKG and to identify articles that lack a DOI. In addition, affiliations are identified with GRID IDs. Industrial sectors and DBpedia categories are identified according to the instances available within INDUSO.

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

- *hasTopic*, which associates with the documents all their relevant topics drawn from CSO.

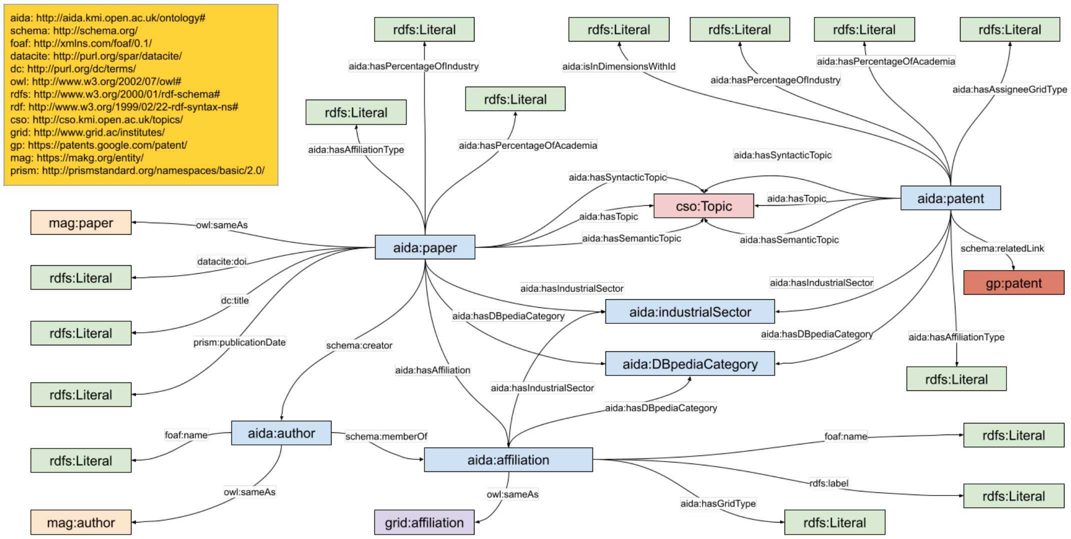


Figure 1: AIDA KG data model. For an enlarged version, please visit `http://w3id.org/aida#aidaschema`.

- *hasIndustrialSector*, which associates with documents and affiliations the relevant industrial sectors drawn from INDUSO.
- *hasAffiliationType*, which associates with the documents the three categories (academia, industry, or collaborative) describing the affiliations of their authors.

AIDA schema includes also some additional relationships which support more complex queries:

- *hasSyntacticTopic* and *hasSemanticTopic*, which indicate, respectively, all the topics extracted using the syntactic module and the semantic module of the CSO Classifier (A. A. Salatino, Osborne, Thanapalasingam, & Motta, 2019). The first set is composed by topics that are explicitly mentioned in the documents. It has high precision but low recall and may be used by applications for which precision is paramount. The second one consists of topics that do not directly appear in the text but were inferred using word embeddings.
- *hasAffiliation*, which identifies the affiliations of a paper.
- *hasPercentageOfAcademia* and *hasPercentageOfIndustry*, which link to articles and patents the percentage of authors from academia and industry. It may be used to generate analytics that need to further segment the *collaborative* category.
- *hasGridType*, *hasAssigneeGridType*, which associate the eight categories of organizations described in GRID (Education, Healthcare, Company, Archive, Nonprofit, Government, Facility, and Other) with affiliations and patents.
- *hasDBpediaCategory*, which associates with papers the industrial categories found in DBpedia (through the *About:Property* and *About:Industry*).
- *isInDimensionsWithId*, which identifies the patent id used within the Dimensions database.

Table 2: Number of triples for each relation in AIDA

Provenance	Relation	N. Triples
AIDA	http://aida.kmi.open.ac.uk/ontology#hasTopic	847,931,791
AIDA	http://aida.kmi.open.ac.uk/ontology#hasSemanticTopic	159,711,581
AIDA	http://aida.kmi.open.ac.uk/ontology#hasSyntacticTopic	70,349,962
AIDA	http://www.w3.org/2000/01/rdf-schema#type	54,839,960
AIDA	http://www.w3.org/2002/07/owl#sameAs	46,950,925
AIDA	http://aida.kmi.open.ac.uk/ontology#hasIndustrialSector	12,006,596
AIDA	http://aida.kmi.open.ac.uk/ontology#hasAffiliationType	9,774,165
AIDA	http://aida.kmi.open.ac.uk/ontology#hasDBpediaCategory	9,691,511
AIDA	http://aida.kmi.open.ac.uk/ontology#isInDimensionWithId	7,940,034
AIDA	http://schema.org/relatedLink	7,940,034
AIDA	http://aida.kmi.open.ac.uk/ontology#hasPercentageOfAcademia	4,179,108
AIDA	http://aida.kmi.open.ac.uk/ontology#hasPercentageOfIndustry	5,745,644
MAG	http://schema.org/creator	53,647,155
MAG	http://xmlns.com/foaf/0.1/name	26,048,450
MAG	http://purl.org/dc/terms/title	20,850,710
MAG	http://prismstandard.org/namespaces/basic/2.0/publicationDate	20,850,710
MAG	http://purl.org/spar/datacite/doi	5,636,401
MAG	http://schema.org/memberOf	4,828,260
MAG	http://aida.kmi.open.ac.uk/ontology#hasAffiliation	6,613,216
GRID	http://aida.kmi.open.ac.uk/ontology#hasAssigneeGridType	5,056,426
GRID	http://aida.kmi.open.ac.uk/ontology#hasGridType	13,171
GRID	http://www.w3.org/2000/01/rdf-schema#label	13,171

As already mentioned, the AIDA knowledge graph also adopts several relations from external sources. These are:

- *https://schema.org/creator*, which links documents to authors and authors to affiliations.
- *https://schema.org/memberOf*, which links authors to affiliations.
- *http://www.w3.org/1999/02/22-rdf-syntax-ns#type*, which defines the type of the entity.
- *http://www.w3.org/2000/01/rdf-schema#label*, which indicates the label of an affiliation.
- *http://purl.org/dc/terms/title*, which indicates the title of a paper.
- *http://purl.org/spar/datacite/doi*, which indicates the DOI of a paper.
- *http://xmlns.com/foaf/0.1/name*, which indicates the name of an author or an affiliation.
- *http://schema.org/relatedLink*, which states the related link of a patent (typically a Google Patent URL).
- *http://prismstandard.org/namespaces/basic/2.0/publicationDate*, which indicates the year of publication of a paper.
- *http://www.w3.org/2002/07/owl/sameAs*, which links papers, authors, or affiliations to their representations on external knowledge bases.

Table 2 reports the number of triples available in the current version of AIDA for each relation. AIDA includes a total of about 1.3B triples: 1.2B with object properties and 98M with datatype properties. Here, we distinguish the provenance of the triples to highlight which ones are directly generated by the AIDA pipeline (described in Section 3.1) and which ones are reused from other knowledge graphs. Overall, 1.18B triples (89,1 % of the total) were generated by our pipeline, while 185M were derived from MAG and 7M from GRID. We reused some relations from MAG, because they enable several kinds of useful queries involving, for instance, the years of publication of the articles and the names of the authors. In the set of triples generated by the AIDA pipeline, 1.08B (82,6%) regard the three main contributions of AIDA. Specifically, 1.07B triples regard the topics (*hasSyntacticTopic*, *hasSeman-*

Table 3: Links of AIDA with external Knowledge Bases.

Knowledge Base	Type	Distinct Entities	Total triples
CSO	Topic	11,091	1,077,993,334
MAKG	Author	26,035,279	26,035,279
MAKG	Paper	20,850,710	20,850,710
INDUSO	Industrial Sector	66	12,007,438
Dimensions	Patent	7,940,034	7,940,034
Google Patents	Patent	7,940,034	7,940,034
GRID	Affiliation	13,171	13,171
DBpedia	Organization	13,171	13,171
DBpedia	Concept	3,864	3,864
Wikidata	Concept	3,842	3,842

ticTopic, *hasTopic*), 19,6M the affiliation types (*hasAffiliationType*, *hasPercentageOfAcademia*, *hasPercentageOfIndustry*), and 12.0M the industrial sectors (*hasIndustrialSector*).

Table 3 reports the number of triples linking AIDA to external knowledge bases and the number of relevant distinct entities. For instance, AIDA includes more then 1B triples having as object a topic in CSO and overall links to 11K unique topics. AIDA is mostly linked to MAKG (the RDF version of MAG), including *own:sameAs* relationships for 21M papers and 25M authors. It also links to Dimensions (8M patents), Google Patents (8M patents), GRID (13K affiliations), and DBpedia (3,864 concepts and 13K affiliations), and Wikidata (3,842 concepts). It should be noted that we cannot link directly to MAG, since it is not available online. However, since we use MAG IDs for papers and authors, mapping MAG and AIDA is trivial.

AIDA includes also the most recent mappings between CSO and DBpedia and between CSO and Wikidata, which implicitly links the documents in AIDA to 3,864 DBpedia entities and 3,842 Wikidata entities. Currently, those statements are not materialized for reason of space. However, materializing these links would yield additional 460M triples linking papers and patents to DBpedia entities (e.g., `http://dbpedia.org/resource/Machine_learning`) and 450M triples linking them to Wikidata entities (e.g., `http://www.wikidata.org/entity/Q2539`). Alternatively, the user can explore these links by formulating SPARQL queries that take advantage of the *owl:sameAs* relationship between CSO, DBpedia, and Wikidata (see example in the Appendix).

The online documentation of AIDA schema is available at `https://w3id.org/aida#aidaschema`.

AIDA is accessible via a Virtuoso triplestore at `http://w3id.org/aida/sparql`. The user can click the "help" button in the upper right of the web page for instructions on how to use the endpoint and some exemplary queries. The full dump of the last versions of AIDA is available at `http://w3id.org/aida/`. The dumps of the previous versions are available at `http://w3id.org/aida/downloads.php#datasets`.

AIDA is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0), meaning that everyone is allowed to i) copy and redistribute the material in any medium or format; ii) remix, transform and build upon the material for any purpose, even commercially.

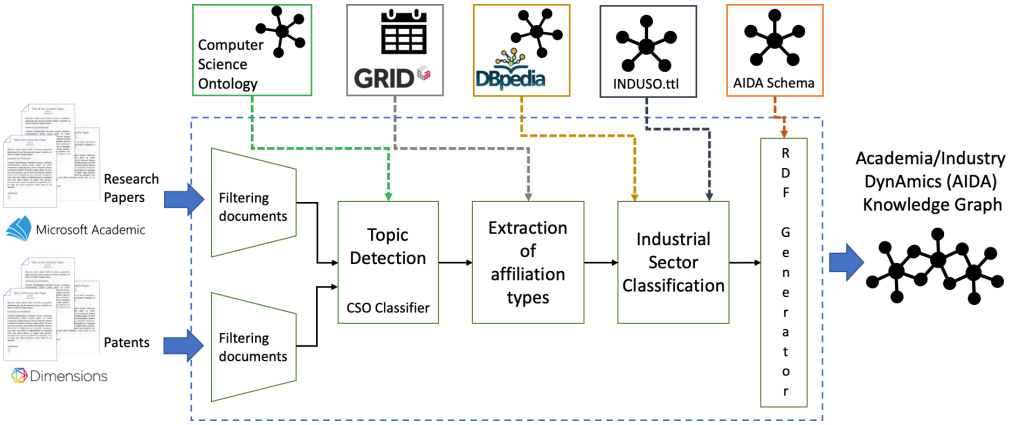


Figure 2: Workflow for the generation of AIDA.

In the following subsections, we will describe the pipeline for the automatic generation of AIDA (Section 3.1) and present an overview of the data (Section 3.2).

3.1 AIDA Generation

The automatic pipeline for generating AIDA works in three steps: topics detection, integration of affiliation types, and industrial sector classification, as shown in Figure 2.

In the following, we will describe each phase of the process (Sections 3.1 - 3.1), discuss the scalability (Section 3.1), and present our plan for producing new versions (Section 3.1).

Topic Detection We first collect all the publications and patents from MAG and Dimensions within the Computer Science domain. In particular, we extract the papers from MAG classified as "Computer Science" in their Field of Science (FoS) (Sinha et al., 2015), an in-house taxonomy of research domains developed by Microsoft. Similarly, the patents in Dimensions are classified according to the International Patent Classification (IPC) and the fields of research (FoR) taxonomy, which is part of the Australian and New Zealand Standard Research Classification (ANZSRC). To extract only the patents from the Computer Science domain, we select those with the following IPC classification: "Computing, Calculating or Counting" (G06), "Educating, Cryptography, Display, Advertising, Seals" (G09), "Information Storage" (G11), "Information and Communication Technology" (G16), and others (G99). We also select those having the following field of research: "Information and Computing Science" (08), and "Technology" (10).

In the current version, the resulting dataset includes 21M publications and 8M patents. The publications (21M) and authors (25M) extracted from MAG are also linked (*owl:sameAs*) to the relevant entities in MAKG. The patents obtained from Dimensions (8M) are linked (*schema:relatedLink*) to the relevant patents in Google Patents.

Since the fields of study in MAG and fields of research in Dimensions are not specific enough for a detailed analysis of the knowledge flow, we then annotate each document with the research topics from the Computer Science Ontology (CSO) (A. A. Salatino et al., 2018b). CSO is an automatically generated ontology of research topics in the field of *Com-*

puter Science. We used the current version (3.2), which includes 14K research topics and 159K semantic relationships. The CSO data model[30] is an extension of SKOS[31] and the main semantic relationships are *superTopicOf*, which is used to define the hierarchical relations within the field of Computer Science (e.g., <*artificial intelligence, superTopicOf, machine learning*>) and *relatedEquivalent*, which is used to define alternative labels for the same topic (e.g., <*ontology matching, relatedEquivalent, ontology alignment*>).

We adopted CSO since it offers a much more granular characterization of research topics than standard classification schemas (e.g., the ACM Classification) and generic knowledge graphs (e.g., DBpedia, Wikidata). For instance, a recent analysis (A. A. Salatino et al., 2020) reported that less than 37% of the topics in CSO are covered by DBpedia.

CSO was officially released in 2019 and has been already adopted by several major organizations, including Springer Nature. In the last two year, CSO supported the creation of many innovative applications and technologies, including ontology-driven topic models (e.g., CoCoNoW (Beck, Rizvi, Dengel, & Ahmed, 2020)), recommender systems for articles (e.g., SBR (Thanapalasingam, Osborne, Birukou, & Motta, 2018)) and video lessons (Borges & dos Reis, 2019), visualisation frameworks (e.g., ScholarLensViz (Löffler et al., 2020), ConceptScope (X. Zhang, Chandrasegaran, & Ma, 2021)), temporal knowledge graphs (e.g., TGK (Rossanez, dos Reis, & da Silva Torres, 2020)), NLP frameworks for entity extraction (Dessì, Osborne, Recupero, Buscaldi, & Motta, 2021), tools for identifying domain experts (e.g., VeTo (Vergoulis, Chatzopoulos, Dalamagas, & Tryfonopoulos, 2020a)), and systems for predicting academic impact (e.g., ArtSim (Chatzopoulos, Vergoulis, Kanellos, Dalamagas, & Tryfonopoulos, 2020a)). It was also used for several large-scale analyses of the literature (e.g., Cloud Computing (Lula, Dospinescu, Homocianu, & Sireteanu, 2021), Software Engineering (Chicaiza & Reátegui, 2020), Ecuadorian publications (Chicaiza & Reátegui, 2020)).

We annotated publications and patents using the CSO Classifier (A. A. Salatino, Osborne, Thanapalasingam, & Motta, 2019), an open-source Python tool[32] that we developed for annotating documents with research topics from CSO (A. A. Salatino, Thanapalasingam, & Mannocci, 2019).

The CSO Classifier was initially developed in the context of a collaboration with Springer Nature, with the aim of automatically classifying scientific volumes according to a granular set of research areas. In this context, it supported Smart Topic Miner (A. A. Salatino, Osborne, Birukou, & Motta, 2019), a web application for assisting the Springer Nature editorial team in annotating conference proceedings in Computer Science, such as LNCS, LNBIP, CCIS, IFIP-AICT and LNICST. This solution brought a 75% cost reduction and dramatically improved the quality of the annotations, resulting in 12M additional downloads over 3 years from the SpringerLink portal[33].

The CSO Classifier is an unsupervised method that operates in three phases. First the syntactic module finds all topics in the ontology that are explicitly mentioned in the paper. Secondly, a semantic module identifies further semantically related topics using part-of-

[30] CSO Schema - `https://cso.kmi.open.ac.uk/schema/cso`
[31] Simple Knowledge Organization System - `https://www.w3.org/2004/02/skos/`
[32] CSO Classifier - `https://pypi.org/project/cso-classifier/`
[33] SpringerLink - `https://link.springer.com/`

speech tagging and similarity over word embeddings. Finally, the CSO Classifier enriches the resulting set by including the super-areas of these topics according to CSO.

Specifically, in the *syntactic* module, the text is split into unigrams, bigrams, and trigrams. Each n-gram is then compared with concepts labels in CSO using the Levenshtein similarity. As result, it returns all matched topics having similarities greater than or equal to the pre-defined threshold.

The *semantic* module takes advantage of a pre-trained Word2Vec word embedding model which captures semantic properties of words (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). We trained this model using titles and abstracts of over 4.6M English publications in the field of Computer Science from MAG. 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"). We performed also a collocation analysis to identify frequent bigrams and trigrams (e.g., "highest_accuracies", "highly_cited_journals"). This solution allows the CSO Classifier to better disambiguate concepts and treat terms such as "deep_learning" and "e-learning" as completely different words. The model parameters are: *method* = skipgram, *embedding-size* = 128, *window-size* = 10, *min-count-cutoff* = 10, *max-iterations* = 5. The semantic module based on these embeddings identifies candidate terms composed of a combination of nouns and adjectives using a part-of-speech tagger. Then, it splits these candidate terms into unigrams, bigrams, and trigrams. For each n-gram we retrieve its most similar word from the Word2Vec model and we compute their cosine similarity with the topic labels in CSO. For bigrams and trigrams, we firstly check in the model their glued version, 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, for each identified topic, the CSO Classifier computes the relevance score as the product between the number of times it was identified (frequency) and the number of unique n-grams that helped it to be inferred (diversity). Finally, it uses the elbow method (Satopaa, Albrecht, Irwin, & Raghavan, 2011) for selecting the set of most relevant topics.

Finally, the resulting set of topics is enriched by including all their super-topics in CSO up to the root: *Computer Science*. For instance, a paper tagged as *neural network* is also tagged with *machine learning*, *artificial intelligence* and *computer science*. This solution yields an improved characterization of high-level topics that are not directly referred to in the documents.

The reader notices that the CSO ontology contains nine levels of topics. When we detect a specific topic (e.g., Neural Networks) we also infer all the super topics in the CSO taxonomy (Machine Learning, Artificial Intelligence, Computer Science). The user can choose to just use the topics directly mentioned in the paper (*hasSyntacticTopic*), the ones inferred by using word embeddings (*hasSemanticTopic*), or the full set of topics that also includes the super-topics (*hasTopic*).

More details about the CSO Classifier are available in A. A. Salatino, Osborne, Thanapalasingam, and Motta (2019).

We also import in AIDA the mapping between CSO and DBpedia, which is a set of 3,864 *owl:sameAs* relationships aligning the two knowledge bases and the mapping between CSO and Wikidata, which includes 3,842 *owl:sameAs* relationships. This allows us to establish several implicit links between documents in AIDA and concepts in DBpedia and Wikidata, which can be materialized with a reasoner or queried using SPARQL (see example in the Appendix).

Integration of Affiliation Types In the second step, we classify papers and patents according to the nature of the relevant organizations in the GRID database. Both MAG and Dimensions link organizations to their GRID IDs. In turn, GRID associates each ID with geographical location, date of establishment, alternative labels, external links, and type of institution (e.g. Education, Healthcare, Company, Archive, Nonprofit, Government, Facility, Other). In total 5.1M articles and 5.6M patents were associated with GRID IDs. We leverage this last field to tag 4.5M articles and 4.9M patents as 'academia', 'industry', or 'collaborative'. A document is assigned an 'academia' type if all the authors or original assignees have an academic affiliation ('Education' in GRID), an 'industry' type if they have an industrial affiliation ('Company' in GRID), and a 'collaborative' type if there is at least one creator from academia and one from industry. AIDA includes also the other categories from GRID through the relation *hasGridType*.

Industrial Sector Classification To characterize the industrial sectors addressed by each document we designed the Industrial Sector Ontology (INDUSO), which is a two-level taxonomy describing 66 sectors and their relationships. INDUSO was created using a bottom-up method that took into consideration the large collection of publications and patents from MAG and Dimensions. Specifically, for each affiliation described in the documents with a GRID ID, we extracted from DBpedia the objects of the properties *About:Purpose* and *About:Industry*. This resulted in a noisy and redundant set of 699 sectors. We then applied a bottom-up hierarchical clustering approach for merging similar sectors. For instance, the industrial sector "Computing and IT" was derived from categories such as "Networking hardware", "Cloud Computing", and "IT service management".

This structure was used as a starting point by a team of ontology engineers from the Open University and the University of Cagliari and domain experts from Springer Nature, who manually revised these categories and arranged the resulting sectors in a two-level taxonomy.

For example, the first level sector "energy" includes "nuclear power", "oil and gas industry", and "air conditioning". Specifically, the INDUSO ontology contains the following properties:

- the *skos:broader* property, which links the first level sectors to the second level sectors.
- the *prov:wasDerivedFrom* property, which associates each of the 66 industrial sectors to the original 699 sectors that were derived from DBpedia.
- the *rdf:type* property, which is used to define the 66 sectors as *:industrialSector* and the original 699 sectors as *:DBpediaCategory*

To tag a document with INDUSO, we identify its affiliations on DBpedia using the link between GRID and DBpedia and then retrieve the objects of the properties *About:Purpose* and *About:Industry*. We then use the previously defined mapping between DBpedia and INDUSO to obtain the industrial sectors.

For instance, a document with an author affiliation described in DBpedia as 'natural gas utility' is tagged with the second level sector 'Oil and Gas Industry' and the first level sector 'Energy'.

Scalability The pipeline currently runs on a server with 128GB of RAM, CPU Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz. Typically, one single paper requires 0.83 seconds to be processed and classified according to the CSO, Academia/Industry, and INDUSO classifications. Therefore, considering the 29M documents (21M papers and 8M patents), and using a multi-thread programming style (we used 10 threads), it takes about 27 days to perform the classification of the entire dataset.

For each following update, we only need to include new documents and update the citations of existing papers. This operation is much faster than processing the entire dataset and we plan to run it periodically. For instance, considering a typically amount of new papers for three months in 2020, equal to about 350k, the update will take around 8h.

Generation of updates We plan to periodically release new versions of AIDA, which will include the most recent publications and patents, as well as the latest versions of CSO and INDUSO. Specifically, we will run the pipeline described in this section – and depicted in Figure 2 – over a new dump of documents every six months. Besides, we also plan to release a new version whenever a significant new version of CSO or INDUSO is produced.

During the writing of this paper, Microsoft decided to decommission the MAG project after 2021. We formulated a plan to switch to other sources that is discussed in Section 6.

3.2 AIDA Overview

In this section, we present an overview of AIDA and discuss some exemplary analytics supported by this resource.

Figure 3 shows the 16 high-level topics (direct sub-topics of Computer Science in CSO) associated with most research articles in AIDA and reports the relevant percentage of academic publications, industrial publications, academic patents, and industrial patents.

These figures were computed by normalizing the number of documents associated with a topic in a category (e.g., academic publications) with the total number of documents in the same category. It should be noted that the percentages do not add to 100% since documents can be associated with multiples topics.

Some topics, such as Artificial Intelligence and Theoretical Computer Science, are mostly addressed by academic publications. Other ones, e.g., Computer Security, Computer Hardware, and Information Retrieval attract a stronger interest from the industry. The topics which are mostly associated with patents are Computer Networks, Internet, and Computer Hardware.

Figure 4 shows the percentage of publications from academia (A) and industry (I) for the same 16 topics across three windows of time (1991-2000, 2001-2010, and 2011-2020). The split in three intervals of ten years is useful to highlight the trend of each topic across the years. Some evident trends include the sharp growth of Computer Security, Information Retrieval, Computer Network, and Internet. Some other topics, such as Software Engineering and Computer Aided Design appear to become less prolific over the last years.

Figure 5 (Main Industrial Sectors I and Main Industrial Sectors II) shows the 16 industrial sectors associated with most research articles and reports their percentage of publications and patents in AIDA.

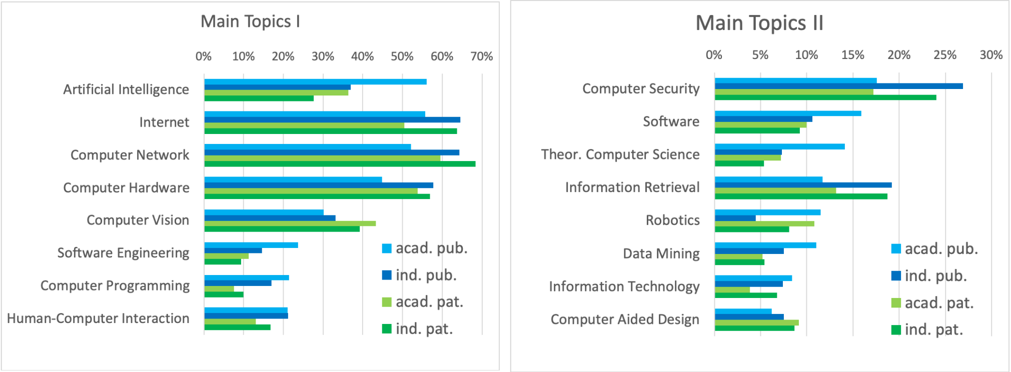


Figure 3: Distribution of the main topics.

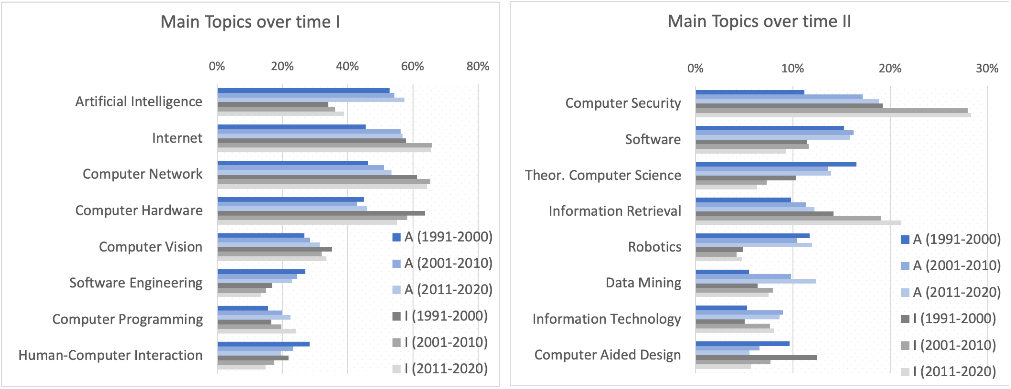


Figure 4: Distribution of the topics in publications across time.

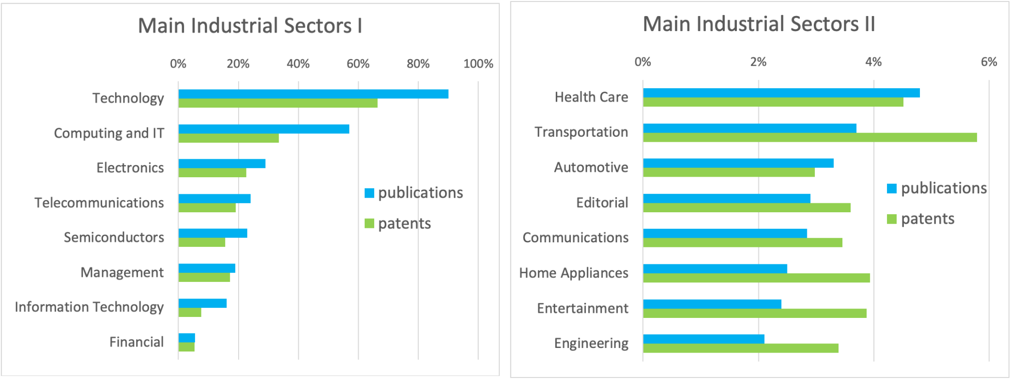


Figure 5: Distribution of the main industrial sectors.

Since AIDA mainly covers Computer Science, the most popular sectors (e.g., Technology, Computing and IT, Electronics, and Telecommunications, and Semiconductors) are linked to this field. However, we can also appreciate the solid presence of sectors such as Financial, Health Care, Transportation, Home Appliance, and Editorial.

AIDA also enables to analyze how these sectors have a different composition in regards to research topics. Table 4 highlights the key topics of a set of exemplary sectors by reporting the difference between the normalized number of publications in a sector and overall. The darker cells mark the main topics for each sector. For instance, the publications written by authors from the Semiconductor sector refer to the topics Computer Aided Design 90% more frequently than the average publication.

Table 4: Topic composition of some prominent industrial sectors. In bold the highest value for each row.

	Computing and IT	Telecommunications	Electronics	Semiconductor	Inf. Technoloogy	Photograpy	Automotive	Financial
Artificial Intelligence	9%	5%	9%	-17%	0%	22%	8%	-6%
Computer Aided Design	-21%	-27%	-2%	**90%**	1%	-5%	2%	-36%
Computer Hardware	-7%	7%	-7%	31%	-5%	-12%	-9%	-17%
Computer Network	-3%	**17%**	-9%	11%	-9%	-18%	-15%	-8%
Computer Programming	18%	-19%	-1%	12%	52%	-31%	-16%	-32%
Computer Security	6%	-1%	-2%	-27%	-1%	9%	-35%	21%
Computer Systems	1%	1%	-3%	1%	4%	-2%	-12%	-10%
Computer Vision	-7%	-1%	**21%**	-16%	-29%	44%	-7%	**52%**
Data Mining	**28%**	-25%	12%	-35%	49%	-18%	-34%	-17%
Human-computer Inter.	14%	-9%	8%	-41%	9%	-21%	-6%	32%
Information Retrieval	6%	-16%	14%	-55%	-6%	71%	-37%	29%
Information Technology	20%	-15%	-5%	-41%	55%	13%	-41%	-20%
Internet	4%	13%	-6%	-1%	1%	-19%	-24%	-4%
Operating Systems	14%	-40%	-8%	1%	**61%**	-24%	-55%	-30%
Robotics	3%	-1%	16%	-14%	-9%	-18%	**322%**	15%
Software Engineering	22%	16%	6%	2%	55%	-24%	20%	-31%

The industrial sectors have a very distinct composition, even when considering just the high-level topics in the table. For instance, the Automotive sector focuses mainly on Robotics, Software Engineering, and Artificial Intelligence; the Telecommunications sector mainly focuses on Computer Network, Internet, and Computer Hardware; and the Photography sector on Information Retrieval, Computer Vision, and Artificial Intelligence.

AIDA can also be queried via triplestore using SPARQL[34]. The ontological schema of AIDA allows users to formulate queries about topics, industrial sectors, and affiliation types associated with articles and patents. In the Appendix of this manuscript we report a selection of sample queries that can be run on our SPARQL endpoint.

4 EVALUATION

To show that AIDA is both correct and useful we performed two evaluations. In the first, reported in subsection 4.1, we measured precision and recall of the three components of the pipeline that produce the data about topics, the academia/industry classification, and the industrial sectors. In the second, presented in subsection 4.2, we evaluated the ability of AIDA to support the task of predicting the impact of a research topic on industry. Specifically, we ran several classifiers on different combination of features and found that the

[34] AIDA triplestore - `http://w3id.org/aida/sparql`

richer representation of topics in AIDA was conducive to significantly better performance than alternative solutions.

4.1 Evaluation of AIDA Generation

The following sub-sections describe the evaluations performed for assessing the topic classification, the academia/industry classification, and the industrial sector classification.

Topic Classification We compared the CSO Classifier, which we use to annotate documents according to their topics, against thirteen unsupervised approaches using a gold standard made of 70 most cited papers (A. A. Salatino, Osborne, Thanapalasingam, & Motta, 2019) within the fields of Natural Language Processing (23 papers), Semantic Web (23), and Data Mining (24). We chose the most cited papers since this solution offers a simple, deterministic, and not arbitrary selection criteria. The 70 papers were annotated by 21 human experts. Each human expert annotated 10 papers; each paper was annotated by 3 human experts resulting in 210 annotations overall. The 21 experts were researchers working in different areas of Computer Science with over 5 years of experience. They were asked to read title, abstract and keywords and assign all the relevant topics from the CSO ontology so as to emulate the classifier's task. Each paper was associated with 14 ± 7.0 topics using the majority voting strategy.

The inter-annotator agreement was 0.45 ± 0.18 according to Fleiss' Kappa, resulting in a moderate inter-rater agreement.

It should be noted that this range of agreement is normal when using a large number of granular categories, such as the 14K topics in CSO.

In Table 5 we report the values of precision, recall, and F1 of all tested classifiers.

The first eight classifiers are based on TF-IDF and Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), and their performance did not exceed a F1 of 30.1%. For each paper, TF-IDF returns a ranked list of words according to their TF-IDF score. The TF-IDF-M classifier, instead, returns the set of CSO topics having Levenshtein similarity higher than 0.8 with the words with the best TF-IDF score. This threshold was set empirically, because it yielded the best performance for the baselines.

LDA100, LDA500, LDA1000 are three LDA classifiers, respectively trained on 100, 500 and 1000 topics. These three classifiers select all LDA topics with a probability of at least j and return all their words with a probability of at least k. The best values of j and k were found performing a grid search. In a similar way, we trained LDA100-M, LDA500-M, and LDA1000-M, but the resulting keywords are then mapped to the CSO topics, as for TF-IDF-M.

W2V-W processes the input document with a ten-words sliding window, and uses the word2vec model to identify CSO topics that are semantically similar to the embedding of the window. The embedding of the window are obtained by averaging the embeddings of the single tokens.

STM is the classifier originally adopted by Smart Topic Miner (Osborne, Salatino, Birukou, & Motta, 2016), the application used by Springer Nature for classifying proceedings within the Computer Science domain. It detects exact matches between the terms extracted from the text and the CSO topics. SYN represents the syntactic module of the CSO

Table 5: Values of precision, recall, and f-measure. In bold the best results.

Classifier	Description	Prec.	Rec.	F1
TF-IDF	TF-IDF	16.7%	24.0%	19.7%
TF-IDF-M	TF-IDF mapped to CSO concepts	40.4%	24.1%	30.1%
LDA100	LDA with 100 topics	5.9%	11.9%	7.9%
LDA500	LDA with 500 topics	4.2%	12.5%	6.3%
LDA1000	LDA with 1,000 topics	3.8%	5.0%	4.3%
LDA100-M	LDA with 100 topics mapped to CSO	9.4%	19.3%	12.6%
LDA500-M	LDA with 500 topics mapped to CSO	9.6%	21.2%	13.2%
LDA1000-M	LDA with 1,000 topics mapped to CSO	12.0%	11.5%	11.7%
W2V-W	W2V on windows of words	41.2%	16.7%	23.8%
STM	Classifier used by STM	**80.8%**	58.2%	67.6%
SYN	Syntactic module	78.3%	63.8%	70.3%
SEM	Semantic module	70.8%	72.2%	71.5%
INT	Intersection of SYN and SEM	79.3%	59.1%	67.7%
CSO-C	The CSO Classifier	73.0%	**75.3%**	**74.1%**

classifier, introduced in (A. A. Salatino, Thanapalasingam, Mannocci, Osborne, & Motta, 2018a). SEM consists of the semantic module of the CSO classifier. INT represents a hybrid version that returns the intersection of the topics produced by the SYN and SEM modules. Finally, CSO-C is the default implementation of the CSO Classifier which produces the union of the topics returned by the two modules. The overall values of precision and recall for a given classifier are computed as the average of the values of precision and recall obtained over the papers.

The data produced in the evaluation, the Python implementation of the approaches, and the word embeddings are available at `http://w3id.org/cso/cso-classifier`.

To note that TF-IDF-M, LDA100-M, LDA500-M, LDA1000-M, W2V-W, STM, SYN, SEM, INT, and CSO-C are all general algorithms that classify a text according to the categories from an input taxonomy. Therefore, no method is specifically biased towards CSO.

The LDA500-M and TF-IDF-M approaches performed poorly with an f-measure of 30.1%. STM and SYN yielded a very good precision of, respectively, 80.8% and 78.3%. These methods were able to find topics explicitly mentioned in the text, which tend to be very relevant. However, they suffered from a low recall, 58.2%, and 63.8% respectively, as they failed to identify more subtle topics. SEM had lower precision than SYN but higher recall and f-measure, suggesting that it can identify further topics that do not directly appear in the paper. INT generated a higher precision (79.3%) compared to SYN and SEM (78.3% and 70.8%), but it did not yield a good recall dropping to 59.1%. Finally, CSO-C outperformed all the other methods in terms of both recall (75.3%) and f-measure (74.1%).

It should be noted that a F1 in the 70%-75% range is remarkably good, given the granularity of the topics in the benchmark, and consistent with the results of other studies that used large classification schemas (e.g., MeSH (Costa et al., 2021)).

Indeed, the agreement (computed with Fleiss' Kappa) among the three annotators which created the gold standard was 0.451 ± 0.177, indicating a moderate inter-rater agreement (Landis & Koch, 1977). When adding the CSO Classifier as fourth annotator the agreement lowers only slightly to 0.392 ± 0.144. The difference with human annotators may completely disappear when considering a simpler classification schema. A recent experiment using the CSO Classifier for assisting systematic reviews (Osborne, Muccini, Lago, & Motta, 2019) reported that its performance were not statistically significantly different from the ones of six senior researchers (p=0.77) when classifying 25 papers according to five main subtopics of Software Architecture. We report in Table 6 the degree of agreement between the annotator (including also CSO-C), computed as the ratio of papers which were tagged with the same category by both annotators.

Table 6: Agreement between annotators (including the CSO classifier) and average agreement of each annotator according to the evaluation in (Osborne et al., 2019). In bold the best agreements for each annotator.

	CSO-C	User1	User2	User3	User4	User5	User6
CSO-C	-	56%	68%	64%	64%	**76%**	64%
User1	56%	-	40%	**56%**	36%	48%	44%
User2	68%	40%	-	64%	52%	**76%**	64%
User3	64%	56%	64%	-	52%	64%	**68%**
User4	**64%**	36%	52%	52%	-	**64%**	52%
User5	**76%**	48%	76%	64%	64%	-	72%
User6	64%	44%	64%	68%	52%	**72%**	-
Av. Agreement	**66%**	45%	58%	59%	51%	63%	60%

Since its introduction, in 2019, the CSO Classifier was adopted by several applications and research efforts (Chatzopoulos, Vergoulis, Kanellos, Dalamagas, & Tryfonopoulos, 2020b; Dörpinghaus & Jacobs, 2020; Jose, Jagathy Raj, & George, 2021; Vergoulis, Chatzopoulos, Dalamagas, & Tryfonopoulos, 2020b). For instance, Dörpinghaus and Jacobs (2020) used it for annotating the articles from the DBLP computer science library. Chatzopoulos et al. (2020b) integrated it in ArtSim, an approach for predicting the popularity of new research papers. Vergoulis et al. (2020b) classified 1.5M papers and use such topical representation for identifying experts that share similar publishing habits. Finally, Jose et al. (2021) developed an ontology-based framework that integrates CSO and the CSO Classifier for retrieving journal articles from academic repositories and dynamically expanding the ontology with new research areas.

Academia/Industry and Industrial Sector Classifications In order to evaluate the quality of the academia/industry classification in AIDA we randomly selected 100 papers: (i) 33 academic papers meaning that all the authors of each paper are reported with academic affiliations only; (ii) 33 industry papers, whose authors are reported with affiliation in the industry only; (iii) 34 collaborative papers, meaning that each paper in this set includes authors with affiliations from academia and authors with affiliations from the industry.

We then asked three independent researchers to manually annotate each paper as 'academic', 'industrial', or 'collaborative' according to the classification above. They were allowed to check online whether a certain institution was academic or industrial. The average agreement score of the three experts was 92.6%. We generated a gold standard by using a majority voting strategy. That is, if a paper was considered an academic paper by at least two researchers, it was labeled as such. There were not cases where a paper was annotated with three different classes by the researchers.

The resulting gold standard perfectly matched the automatic classification.

Table 7: Performance of industrial sector classification task.

Industrial Sector	*Precision*	*Recall*	*F1-Score*
Automotive	1.000	1.000	1.000
Healthcare	0.894	0.894	0.894
Computing and it	0.850	0.809	0.829
Electronic	0.700	0.777	0.736
Telecommunication	0.944	0.894	0.918
Macro Average	0.877	0.875	0.875
Weighted Average	0.879	0.875	0.877

To evaluate the accuracy of our approach for identifying the industrial sectors of a document, we selected 100 organizations equally divided (20 per each industrial sector) among telecommunication companies, healthcare companies, automotive companies, computing and information technology companies, and electronic companies. We then asked three independent experts (three senior researchers working within ICT companies and with computer science background) to annotate each organization among the five classes above (or the *other* category if none of the previous categories was appropriate). The average agreement score of the experts was 84.0%.

We created a gold standard using a majority voting strategy. For instance, if a company was classified as healthcare by at least two experts, then its label was healthcare. To note that for each company at least two experts always gave the same label. We then performed a precision-recall analysis of the categories forecasted by our approach and, for each category, we obtained the performance shown in Table 7.

It is interesting to note that, while the performance of our approach is overall quite good, it can differ according to the category. For example it is quite easy to recognize organizations in the ’Automotive’ sector, but much less so to identify the ones in ’Electronic’. The same issues also affected human annotators. An analysis of the results seem to suggest that some categories (e.g., Electronic) are potentially more ambiguous according to both human annotators and the linked categories on DBpedia. Conversely, some other categories are more well defined and relatively easy to identify.

In conclusion, the evaluation substantiated that our approaches for classifying documents work remarkably well, performing similarly to human annotators.

4.2 Impact Forecasting

In this section, we present an evaluation of the ability of AIDA to support machine learning forecasters for predicting the impact of research topics on the industry, which is a typical task in the study of academia/industry relationship (Altuntas, Dereli, & Kusiak, 2015; Choi & Jun, 2014; Marinakis, 2012; Ramadhan et al., 2018; Zang & Niu, 2011). The impact of research topics on the industry has been traditionally quantified using the number of relevant patents. For instance, in AIDA the topic *wearable sensors* was granted only 2 patents during 2009. In the following years, a lot of commercial organizations started to invest in this area and submitted several patents, ultimately producing 135 patents in 2018. Predicting these dynamics is very advantageous for companies that need to stay at the forefront of innovation and anticipate new technologies.

The literature proposes a range of approaches to patent and technology prediction through patent data, using for instance weighted association rules (Altuntas et al., 2015), Bayesian clustering (Choi & Jun, 2014), and various statistical models (Marinakis, 2012) (e.g., Bass, Gompertz, Logistic, and Richards). In the last few years, we saw the emergence of several approaches based on Neural Networks (Ramadhan et al., 2018; Zang & Niu, 2011), which lately obtain the most competitive results. However, most of these tools 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 hypothesized that a knowledge graph like AIDA which integrates a lot of 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 (Choi & Jun, 2014; Marinakis, 2012; Ramadhan et al., 2018; Zang & Niu, 2011).

To test this hypothesis, we generated a gold standard that associates with each topic in AIDA all the time-frames of five years in which the topic had not yet emerged (less than 10 patents). These samples were labeled as *True* whenever the topic produced more than 50 industrial patents (PI) in the following 10 years and *False* otherwise. We then associated to each sample six-time series composed respectively of the number of research articles (R), the number of patents (P), the number of research articles from academia (RA), research articles from industry (RI), patents from academia (PA), patents from industry (PI). For instance, the sample involving the topic *wearable sensors* in 2005-2009 contains the six series (R,P,RA,RI,PA,PI) describing the number of documents in each category during those five years and was labeled as *True*, since *wearable sensors* produced more than 50 industrial patents (PI) in the following years. The resulting dataset includes 9,776 labeled samples.

We trained five machine learning classifiers on this gold standard: Logistic Regression (LR), Random Forest (RF), AdaBoost (AB), Convoluted Neural Network (CNN), and Long Short-term Memory Neural Network (LSTM). LR, RF, and AB use the standard implementation of scikit-learn 0.22. CNN and LSTM were implemented using Tensorflow and Keras. CNN was composed of two Convolution1D/MaxPooling1D layers and one output layer computing the softmax function. LSTM uses one LSTM hidden layer of 128 units and one output layer computing the softmax function. We used both binary cross-entropy as loss functions and trained them over 50 epochs. For the LSTM, we used 32, 64, 128, 256, and 512 units and 128 was the one performing the best. Moreover, after 50 epochs the accuracy started dropping.

We ran each of the classifiers on research papers (R), patents (P), and the 15 possible combinations of the other four-time series (RA, RI, PA, PI) to assess which set of features would yield the best results. We performed 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). The dataset, the results of experiments, the parameters, the implementation details, and the best models are available at `http://w3id.org/aida/downloads`.

Table 8 shows 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. The reader notices that our main goal was to show that the combination of the four time series (number of papers from academia, number of papers from industry, number of patents from academia, and number of patents from industry) improves the performance of all the predictors. This

Table 8: 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 A. Salatino et al. (2020).

	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**

proves that the granular representation of documents in AIDA yields significant advantages to these systems.

We can observe that using the combination (RA-RI-PI) significantly (p<0.0001) outperforms (F1: 84.7%) the version which uses only the number of patents by companies (74.8%). PA (academic patents) is the weakest of all the indicators, probably because there is a very small number of academic patents. Considering the origin (academia and industry) of the publications and the patents also increases performance: RA-RI (80.7%) significantly (p<0.0001) outperforms R (68.2%) and PA-PI (75.2%) is marginally better than P (74.8%). This confirms that the most granular representation of the document origin in AIDA can increase the forecaster performance.

Another interesting outcome is that, when considering only one of the time series, the number of publications from industry (RI) is a significant (p=0.004) better indicator than patents from industry (PI), yielding an F1 of 76.9%, followed by RA, and PA. The best combination of two-time series is RI-PI (81.4%), while the best combination of three-time series is RA-RI-PI (84.7%).

In conclusion, the experiments substantiate the hypothesis that the granular representation of publications and patents in AIDA can support effectively 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 AIDA USAGE

To test the AIDA's ability to generate advanced analytics, in the last year we generated preliminary versions of AIDA for analysing the research trends in Computer Science. The feedback collected during these studies was used to improve the semantic schema of AIDA and the scalability of its pipeline. We summarize here the main results of these research efforts. Specifically, in subsection 5.1 we report a study about topic dynamics across publications and patents from academia and industry (A. Salatino et al., 2020) that used an initial version of AIDA focused on the main 5K topics in Computer Science. In subsection 5.2 we present an analysis of the main research trends among papers published in two main venues of Human-Computer Interaction (HCI) (Mannocci, Osborne, & Motta, 2019).In order to further showcase AIDA ability to support tools for analysing the research landscape, in subsection 5.3 we describe the *AIDA Dashboard*, a new web application based on AIDA that we developed to support Springer Nature editors in assessing the quality of scientific conferences.

5.1 Analysing Academia Industry Relationship

Monitoring the research trends across articles and patents can lead to a deeper understanding of the knowledge flow between academia and industry. In our recent study (A. Salatino et al., 2020), we used an initial version of AIDA to represent a set of 5K topics in CSO according to four time series reporting the time frequency of i) papers from academia, ii) papers from industry, iii) patents from academia, and iv) patents from industry. We then analysed the resulting time series to identify insightful patterns.

Figure 6 shows the distribution of these topics in a bi-dimensional diagram according to two indexes: academia-industry (horizontal axis) and papers-patents (vertical axis). The papers-patents index of a certain topic t is the difference between the number of research papers R_t and patents P_t related to t, over the whole set of documents $(R_t + P_t)$: $(R_t - P_t)/(R_t + P_t)$. If this index is positive a topic tends to be associated with a higher number of publications, while if it is negative with a higher number of patents. On the other hand, the academia-industry index for a certain topic t is the difference between the documents in academia A_t and industry I_t, over the whole set of documents $(R_t + P_t)$: $(A_t - I_t)/(R_t + P_t)$. If this index is positive a topic tends to be mostly associated with academia, if it is negative with industry.

As we can observe from Figure 6, topics are tightly distributed around the bisector: the ones which attract more interest from academia are prevalently associated with publications (top-right quadrant), while the ones in industry are mostly associated with patents (bottom left quadrant).

We also performed an analysis of the emergence of topics across the four time series. In particular, we determined when a topic emerges in all time series, and compared the time elapsed between each couple of them. In order to avoid false positives, we considered a topic as 'emerged' when it was 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 the industry by 5.6±5.6 years, and in turn, the latter preceded patents from the industry by 1.0±5.8 years, as showed in Figure 7. Publications from academia also preceded by 6.7±7.4 years patents from the industry. This outcome is consistent with previous studies which identified academia as the main

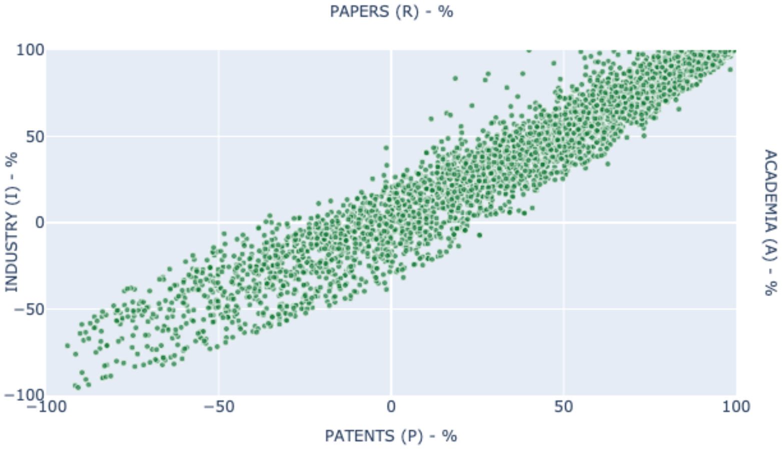


Figure 6: Distribution of the most frequent 5,000 topics according to their academia-industry and publication-papers indexes (A. Salatino et al., 2020).

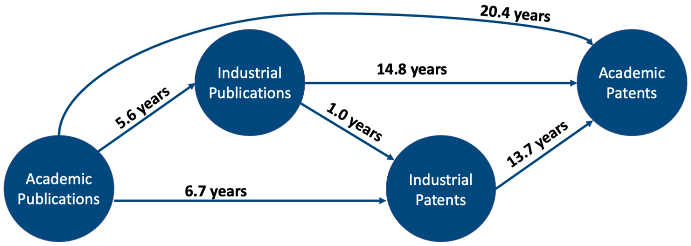


Figure 7: Average time lags when analysing the emergence of topics through their four time series.

creator of new knowledge (Larivière et al., 2018), but it is able to quantify much more accurately when specific research topics emerge. More details about this analysis are available in A. Salatino et al. (2020).

5.2 Detecting Research Trends

A preliminary version of AIDA focusing only on publications in Human-Computer Interaction (HCI) in 1969-2018 was used to perform an analysis of the field that was published on the special issue of the International Journal of Human-Computer Studies celebrating the 50 years of the journal (Mannocci et al., 2019). The analysis focuses on two main venues of HCI: the International Journal of Human-Computer Studies (IJHCS) and the Conference on Human Factors in Computing Systems (CHI). The resulting data reporting the evolution of topics were analyzed with the help of domain experts to detect the most prominent topics in various timeframes and the most significant trends in the last ten years. We briefly report the main results as they are an excellent example of the bibliographic analyses that AIDA can support.

Figure 8 compares the percentage of publications tagged with the main topics in IJHCS (blue) and CHI (orange). It was created by computing the percentage of publications associ-

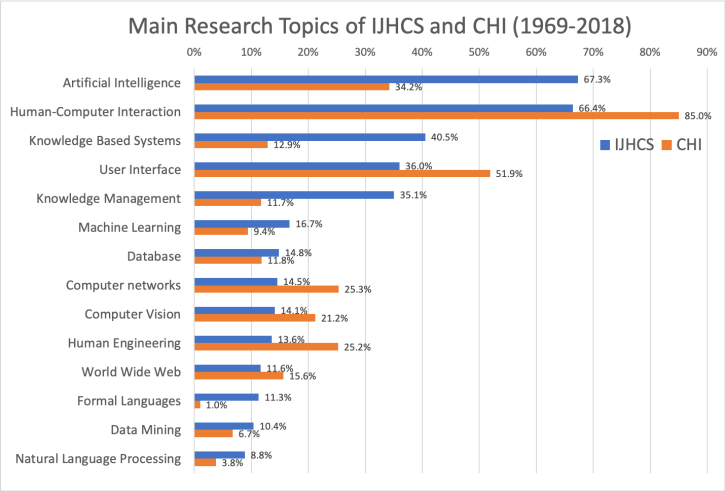


Figure 8: Comparison of the main research topics of IJHCS and CHI during 1960-2018

ated with the same research topics in the preliminary version of AIDA. The two top venues in HCI tend to address a similar set of topics but also present some intriguing differences. For instance, IJHCS has a more interdisciplinary focus, and in particular, it addresses several topics related to Artificial Intelligence such as Knowledge-Based Systems, Knowledge Management, Formal Languages, and Natural Language Processing. This outcome was also confirmed by the editors of IJHCS.

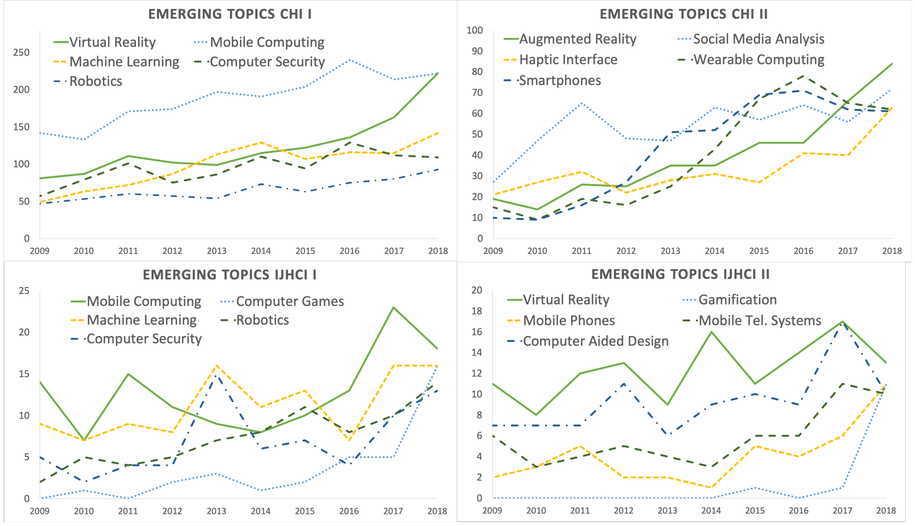


Figure 9: Emerging topics in IJHCS and CHI during 2009-2018.

Figure 9 shows the main emerging topics in the two venues under analysis. These were the topics that experienced the steepest improvement in terms of the number of associated articles in the decade 2009-2018. AIDA allows users to compute these analytics by simply querying and aggregating the relevant data. In this instance, we can easily detect that the emerging research trends of HCI in the last years include Virtual Reality, Mobile Computing, Robotics, Haptic Interfaces, Social Media Analysis, and Gamifications. A more comprehensive analysis of these trends is available in Mannocci et al. (2019).

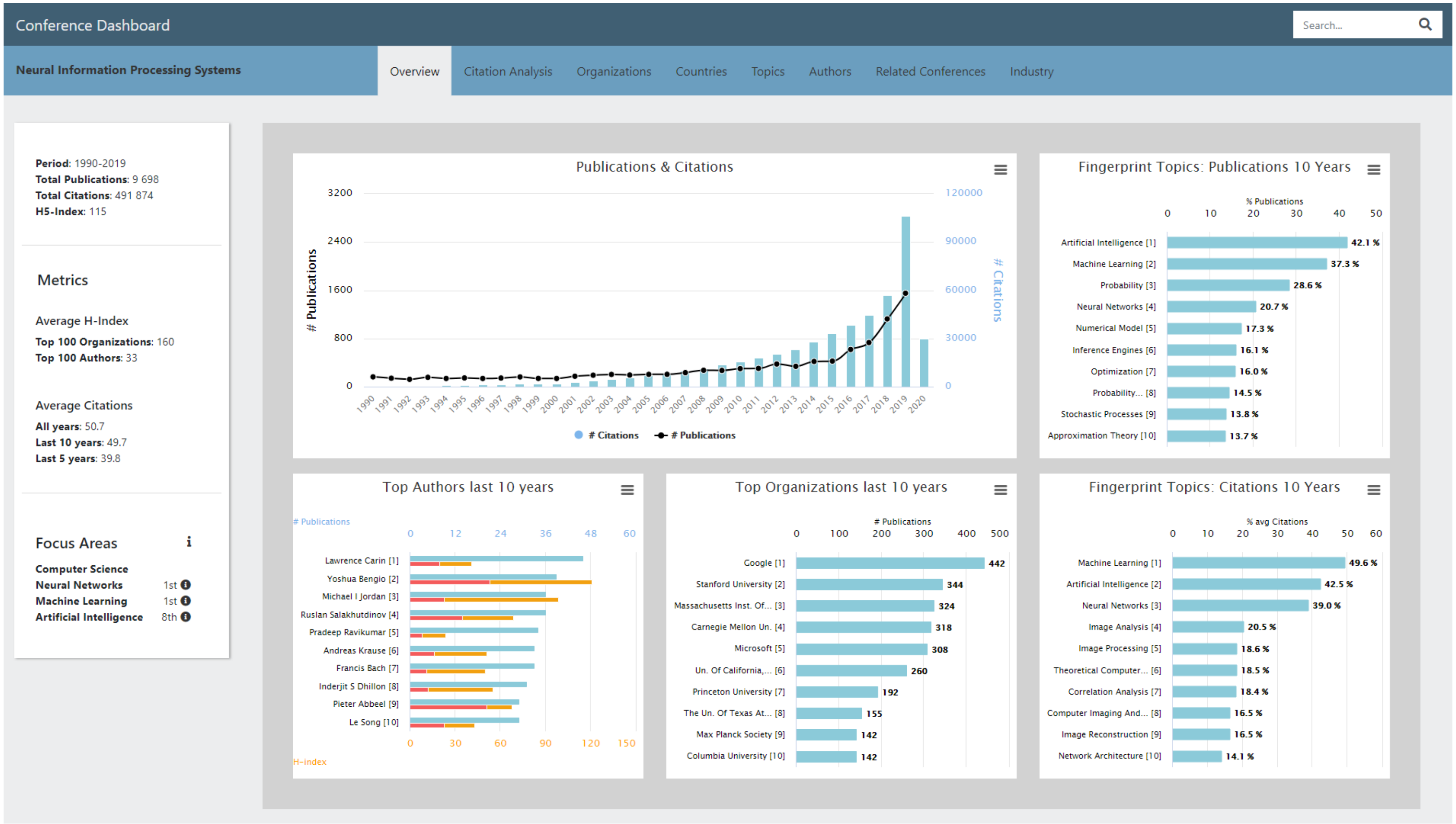
5.3 The AIDA Dashboard: assessing scientific conferences

Scientific conferences play a crucial role in the field of Computer Science by offering high-quality venues for research articles, promoting new collaborations, and connecting research efforts from academia and industry. Understanding and monitoring conferences is thus a crucial task for researchers, editors, funding bodies, and other users in this space. While several academic search engines (e.g., as Microsoft Academic Graph, Semantic Scholar, Scopus) provide basic information about conferences, they do not offer advanced analytics to rank and compare them, assess their main trends, or study their involvement with specific industrial sectors.

To address these limitations, we created the *AIDA Dashboard*, a new web application that takes advantage of AIDA for supporting users in analysing scientific conferences. The AIDA Dashboard was developed in collaboration with Springer Nature, with the primary objective of supporting their team in assessing the quality of a conference in order to inform editorial decisions. However, the analyses supported by the AIDA Dashboard can assist several other stakeholders, including researchers and funding bodies. Specifically, the AIDA Dashboard introduces three novel features that state-of-the-art systems currently lack. First, it characterizes conferences according to the granular representation of topics from AIDA, hence providing high-quality analytics about their research trends over time. Second, it enables users to easily compare conferences in the same fields according to several bibliometrics. Third, it allows users to assess the involvement of commercial organizations in a conference by offering analytics about the academia/industry collaborations and the relevant industrial sectors.

The AIDA Dashboard describe each conference according to eight tabs: *Overview*, *Citation Analysis*, *Organizations*, *Countries*, *Authors*, *Topics*, *Similar Conferences*, and *Industry*. The *Overview* tab (see Figure 10) summarizes the most important information with the aim to allow the user to immediately understand what the conference is about and how it is performing in the last few years. The *Citation Analysis* tab reports several citation-based bibliometrics and highlights how the conference ranks in its main research areas. The *Authors*, *Organizations*, and *Countries* tabs enable users to analyse the actors that produced the articles at different level of granularity (researchers, institutions, and geographical locations). The *Topics* tab allows users to inspect the main research topics and analyse their trends in time. The *Similar Conferences* tab compares the conference under analysis with all the other conferences in the same fields according to different bibliometrics. Finally the *Industry* tab reports the percentage of articles and citations from academia, industry, and collaborative efforts as well as the frequency of the industrial sectors from AIDA.

The AIDA Dashboard is still under development and we aim to release a first stable version in the second part of 2021. A demo of the current prototype is available at `https://aida.kmi.open.ac.uk/dashboard/`.



Period: 1990-2019
Total Publications: 9 698
Total Citations: 491 874
H5-Index: 115

Metrics

Average H-Index
Top 100 Organizations: 160
Top 100 Authors: 33

Average Citations
All years: 50.7
Last 10 years: 49.7
Last 5 years: 39.8

Focus Areas

Computer Science	
Neural Networks	1st
Machine Learning	1st
Artificial Intelligence	8th

Figure 10: The overview page of the NeurIPS conference according to the AIDA Dashboard.

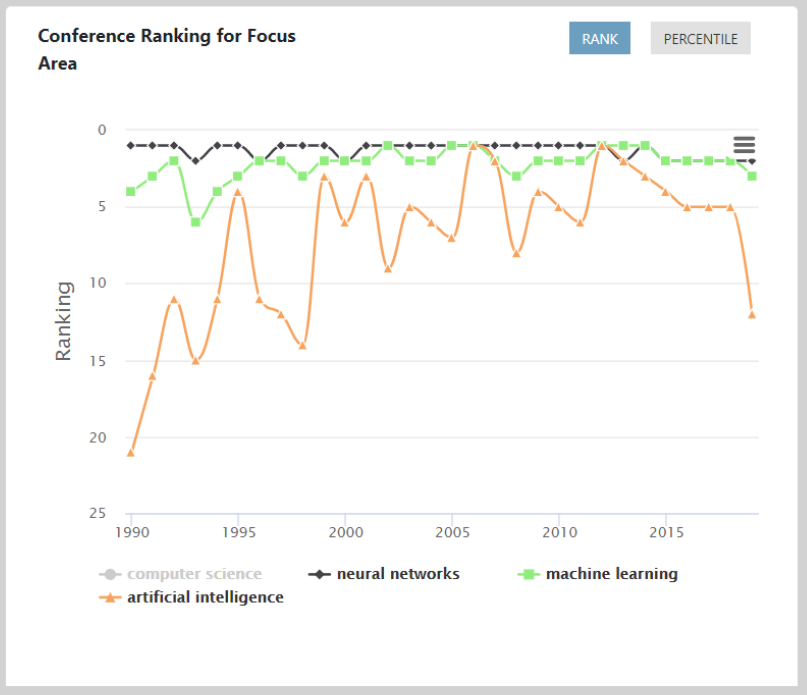


Figure 11: The rank of NeurIPS in its three main focus areas (neural networks, machine learning, artificial intelligence) across time. The conferences are ranked according to their average citations per article.

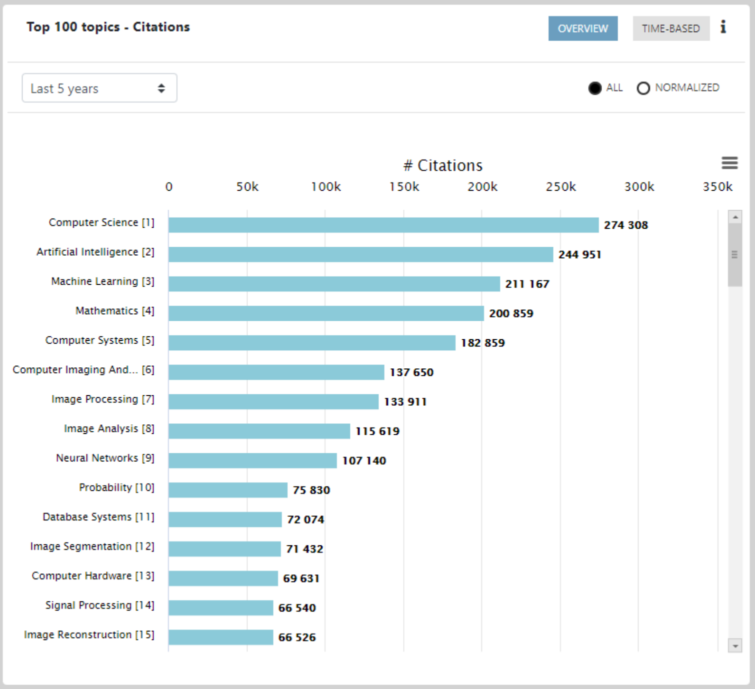
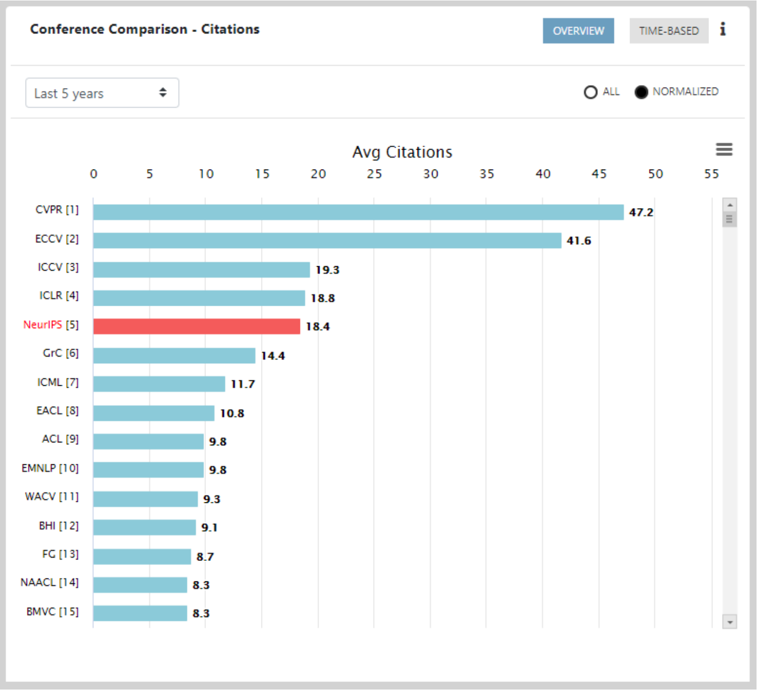


Figure 12: The most cited topics in NeurIPS during the last five years.

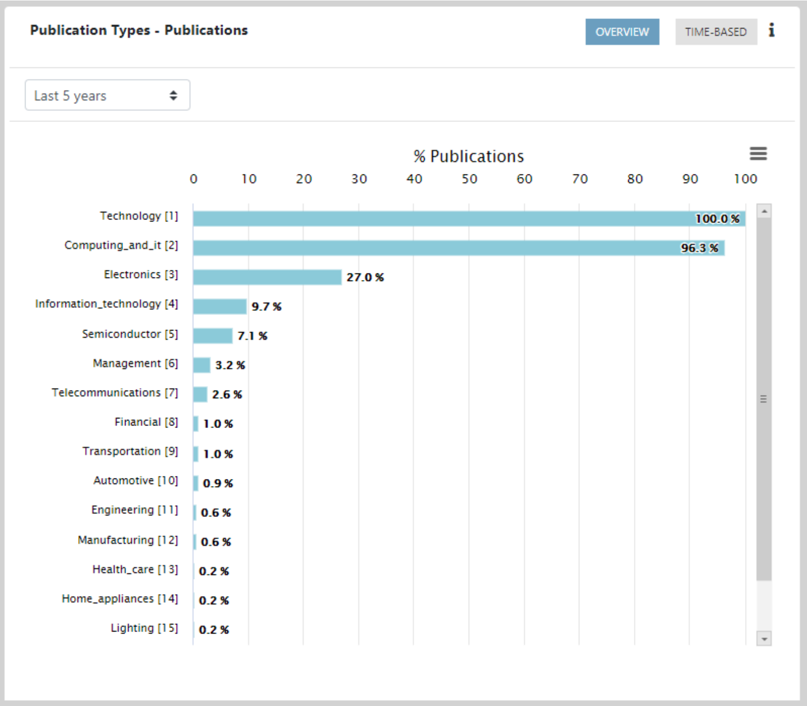


Conference Comparison - Citations

Last 5 years

○ ALL ● NORMALIZED

Figure 13: The best Artificial Intelligence conferences in terms of average citations in the last five years. NeurIPS is in fifth position, highlighted in red.



Publication Types - Publications

Last 5 years

Figure 14: Most frequent industrial sectors in NeurIPS during the last five years.

In order to showcase the functionalities of the AIDA Dashboard, Figures 10-14 illustrate some of the analytics generated for one of the main conferences in the field of Neural Networks: *the Neural Information Processing Systems Conference (NeurIPS)*.

The users can search any conference from the main page. After they select a conference (e.g., NeurIPS) they are redirected to its *Overview* tab. Figure 10 shows the *Overview* tab of NeurIPS which displays several high level information including basic bibliometrics and the main authors, organizations, and topics. We can note the presence of organizations such as Google, Stanford, and MIT and of a Turing Award winner (Yoshua Bengio) and many world-leading researchers in neural networks in the main authors. In the bottom left side, the AIDA Dashboard reports the *focus areas* of NeurIPS: *Neural Networks*, *Machine Learning* and *Artificial Intelligence*. These are high-level fields used to categorize and compare conferences. They are computed automatically by analysing the topic distribution of the conference in AIDA.

The line chart in Figure 11, from the *Citation Analysis* tab, shows how NeurIPS ranks in terms of average citations per paper in the three focus areas. In the last ten years, NeurIPS has always been rippling between the first and second position in the fields of neural networks and machine learning.

The plot in Figure 12 is from in the *Topics* tab and shows the topics which received most citations in the conference. In addition to the focus areas of the conference (Neural Networks, Machine Learning, Artificial Intelligence) we can see many other relevant high-level topics (e.g., Mathematics, Probability, Signal Processing) as well as some important domains of application (e.g., Image Processing, Human Computer Interaction).

Figure 13, from the *Related Conferences* tab, shows the comparison between NeurIPS and all the other conferences in Artificial Intelligence in terms of average citations in the last 5 years. As we can see, NeurIPS ranks fifth with an average of 18.4 citations for article.

Finally, the bar chart in Figure 14, from the *Industry* tab, shows the percentages of the published articles relevant to several industrial sectors from the INDUSO ontology. For NeurIPS, 96.3% of the articles are from Computing and IT, 27% from Electronics, 9.7% from Information Technology, and so on. The *Industry* tab also shows the frequencies of articles published by i) authors exclusively from academia, ii) authors exclusively from industry, and iii) from a joint collaboration of authors from both academia and industry. In Table 9 we report the percentage of articles based on their affiliation. While most articles are from academia, the percentage of industrial and collaborative articles is significantly higher in the last 5 years, suggesting a growing interest by commercial organizations. The overview page, shown in Figure 10, shows some of the companies involved in this shift. The user can also use the *Organizations* tab to display in a line chart the growing number of publications associated with commercial organizations such as Google, Microsoft, IBM, and Facebook.

6 LIMITATIONS

In this section, we discuss some limitations of the current pipeline, and describe our plans to address them in the future.

A first challenge regards improving the scalability. A significant bottleneck of the current version is that it uses the DBpedia REST API for identifying industrial sectors. This solution relies on REST requests on the web and therefore it is quite slow. We plan to switch to a local DBpedia instance in order to solve this issue. In addition, we are currently working

Table 9: Percentages of articles written by Academia/Industry/Collaborative in NeurIPS.

	All Years	Last 5 Years
Academia	80.48%	71.59%
Industry	5.40%	6.61%
Collaborative	14.11%	21.79%

on a new version of the CSO Classifier that uses a smarter cache in the semantic module to improve scalability. We believe that these changes may be able to cut the computational time by half or more.

A second limitation regards the fact that only a subset of the documents (5.1M articles and 5.6M patents) are mapped to GRID and can thus be assigned with the types of affiliations and industrial sectors. We plan to address this issue from different directions. First, we intend to directly map the name of the organizations to DBpedia and knowledge bases of companies using entity-linking solutions. We are also working on link prediction techniques for graph completion, that can be used to automatically classify the affiliations according to contextual information in the knowledge graph. An interesting challenge in this regard is that AIDA contains several N to M relations with N$\gg$M. Given a triple (h, r, t), this situation arises when the cardinality of the entities in the head position (h) for a certain relation (r) is much higher than the one of the entities in the tail position (t). This is actually the case for most scholarly knowledge graphs (Ammar et al., 2018; Knoth & Zdrahal, 2011; Peroni & Shotton, 2020; K. Wang et al., 2020; Y. Zhang et al., 2018) that usually categorize millions of documents (e.g., papers, patents) according to a relatively small set of categories (e.g., topics, countries, chemical compounds). Another important requirement is the scalability of these methods, since we need to be able to process million of entities. We are thus focusing on the creation of link prediction approaches that perform well in this space. The first output of this research line was Trans4E (Nayyeri et al., 2021), a scalable model which tackles these issues by providing a very large number of possible vectors ($8^d - 1$, where d is the embedding dimension) to be assigned to entities involved in N to M relations.

A final important limitation is that the current version of the pipeline uses MAG as source for research articles. Unfortunately, during the writing of this paper, Microsoft decided to decommission the MAG project after 2021[35]. In order to react timely, we worked on this issue with Springer Nature data science team and devised a strategy to obtain the article metadata from Dimensions. We chose this knowledge graph due to its wide coverage of Computer Science and low cost of integration (AIDA already uses Dimensions for patents). Since Dimensions does not disambiguate conferences, we also plan to leverage the conference representation of DBLP, which currently includes 5,438 conferences in Computer Science. Preliminary experiments show that most conferences available in MAG are also covered by DBLP. We plan to integrate Dimensions and DBLP using the paper DOIs. For the few conferences and workshops that do not assign DOIs to articles, we will map the

[35] Next Steps for Microsoft Academic – Expanding into New Horizons - https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach/

papers across the two datasets by computing the string similarity of their titles and authors, after applying filters that normalise, uniform cases, and remove punctuation. We will also leverage additional fields, such as the year of publication and the proceedings title, in order to reduce the number of papers to compare and provide further confirmation of the alignments. We plan to switch to this new solution before the end of 2021.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we have introduced AIDA, the Academic/Industry DynAmics Knowledge Graph. This resource characterizes 21M publications and 8M patents according to the research topics drawn from the Computer Science Ontology (CSO). 5.1M publications and 5.6M patents are also classified according to the type of the author's affiliations and industrial sectors. To characterize documents according to their industrial sectors, we designed the Industrial Sectors Ontology (INDUSO), which describes 66 sectors in a two-level taxonomy.

AIDA was generated using an automatic pipeline that merges and integrates information from Microsoft Academic Graph, Dimensions, DBpedia, the Computer Science Ontology, and the Global Research Identifier Database. It allows researchers to analyze the evolution of research topics across academia and industry as well as to understand their dynamics within several industrial sectors. It can be used to identify the research trends of different industries and how and when academia and/or industry tackle these in particularly significant ways, thus facilitating a granular analysis of the interaction between these two worlds. Moreover, AIDA can also be employed to investigate authors, citations, countries, and other entities already present in Microsoft Academic Graph.

In order to showcase how AIDA can be used by the wider community, we also presented some exemplary studies that take advantage of AIDA for producing advanced bibliometric analysis and introduced the AIDA Dashboard, a novel tool that aims to support Springer Nature editors in assessing the quality of scientific conferences

The process for producing AIDA is general and can be applied to other domains of science. In this case, the CSO Classifier, which is the main computer science specific portion of our pipeline, needs to be tailored to the new field. In order to do so, it is necessary to replace CSO with a different domain ontology and retrain the word2vec model with a corpus of documents that fits the new domain. This procedure is detailed in https://doi.org/10.5281/zenodo.3459286.

We evaluated different parts of the pipeline using a manually created gold standard and obtaining very competitive results. We also evaluated the impact of AIDA on forecasting systems for predicting the impact of research trends on the industry. In particular, we found that a forecaster based on LSTM neural networks and exploiting the full representation of articles and patents from AIDA yielded significantly better performance ($p<0.0001$) than alternative methods. Besides, the version of this classifier using the full set of features (84.6%) gained almost 10% in terms of F1 in comparison with the one using only the number of patents across time (74.8%). This substantiates the hypothesis that adopting a more granular representation of articles and patents is critical for this task.

The resource presented in this paper opens up several interesting directions of work. First, we will produce a comprehensive analysis of AIDA and the most significant research trends in academia and industry. We also intend to use AIDA to support systems for predicting the impact of specific areas of industry research.

We plan to further improve AIDA using graph completion and link prediction techniques. Since many state-of-the-art solutions in this space may suffer when dealing with KGs that categorize a very large number of entities (e.g., research articles, patents, persons), we are currently investigating new scalable approaches that can deal with this situation (Nayyeri et al., 2021). We are also exploring the possibility of using other KGs, such as Wikidata and BabelNet, to further improve the performance of graph completion techniques on AIDA.

We plan to explore the application of our pipeline to other fields, such as Biology and Engineering. To this purpose we intend to develop a new version of our classifier, testing also a range of recent word embeddings solutions, such as BERT and SciBERT. One more direction regards a further classification of papers in peer reviewed and not peer reviewed.

As far as the dashboard is concerned, we are currently performing a comprehensive evaluation with different kinds of users and will make available the results in a future paper. Finally, we are going to employ AIDA for human-robot interaction and develop a robot that can answer questions about the scholarly domain in natural language.

ACKNOWLEDGEMENTS

We gratefully acknowledge Springer Nature for funding this research. We also thank Dimensions for sharing their large dataset of patents. Last but not least, we gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU used for this research.

AUTHOR CONTRIBUTIONS

Simone Angioni: Writing – original draft, Data curation, Formal Analysis; **Angelo Salatino**: Writing – original draft, Formal Analysis, Data curation; **Francesco Osborne**: Writing – original draft, Formal Analysis, Project administration; **Diego Reforgiato Recupero**: Writing – original draft, Formal Analysis, Project administration, Validation; **Enrico Motta**: Writing – review & editing, Supervision, Project administration.

REFERENCES

Altuntas, S., Dereli, T., & Kusiak, A. (2015). Analysis of patent documents with weighted association rules. *Technological Forecasting and Social Change*, *92*, 249–262.

Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., … others (2018). Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262*.

Anderson, M. S. (2001a). The complex relations between the academy and industry: Views from the literature. *The journal of higher education*, *72*(2), 226–246.

Anderson, M. S. (2001b). The complex relations between the academy and industry: Views from the literature. *The Journal of Higher Education*, *72*(2), 226–246. Retrieved from `http://www.jstor.org/stable/2649323`

Angioni, S., Salatino, A. A., Osborne, F., Recupero, D. R., & Motta, E. (2020). Integrating knowledge graphs for analysing academia and industry dynamics. In *Adbis, tpdl and eda 2020 common workshops and doctoral consortium* (pp. 219–225).

Ankrah, S., & Omar, A.-T. (2015). Universities–industry collaboration: A systematic review. *Scandinavian Journal of Management*, *31*(3), 387–408.

Ankrah, S. N., Burgess, T. F., Grimshaw, P., & Shaw, N. E. (2013). Asking both university and industry actors about their engagement in knowledge transfer: What single-group studies of motives omit. *Technovation*, *33*(2-3), 50–65.

Beck, M., Rizvi, S. T. R., Dengel, A., & Ahmed, S. (2020). From automatic keyword detection to ontology-based topic modeling. In *International workshop on document analysis systems* (pp. 451–465). doi: 10.1007/978-3-030-57058-3_32

Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., & Morissette,

J. (2008). Bio2rdf: towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics*, *41*(5), 706–716.

Bikard, M., Vakili, K., & Teodoridis, F. (2019). When collaboration bridges institutions: The impact of university–industry collaboration on academic productivity. *Organization Science*, *30*(2), 426–445.

Bird, S., Dale, R., Dorr, B. J., Gibson, B., Joseph, M. T., Kan, M.-Y., … Tan, Y. F. (2008). The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*(null), 993–1022.

Borges, M. V. M., & dos Reis, J. C. (2019). Semantic-enhanced recommendation of video lectures. In *2019 ieee 19th international conference on advanced learning technologies (icalt)* (Vol. 2161, pp. 42–46). doi: 10.1109/ICALT.2019.00013

Chatzopoulos, S., Vergoulis, T., Kanellos, I., Dalamagas, T., & Tryfonopoulos, C. (2020a). Artsim: improved estimation of current impact for recent articles. In *Adbis, tpdl and eda 2020 common workshops and doctoral consortium* (pp. 323–334). doi: 10.1007/978-3-030-55814-7_27

Chatzopoulos, S., Vergoulis, T., Kanellos, I., Dalamagas, T., & Tryfonopoulos, C. (2020b). Artsim: Improved estimation of current impact for recent articles. In L. Bellatreche et al. (Eds.), *Adbis, tpdl and eda 2020 common workshops and doctoral consortium* (pp. 323–334). Cham: Springer International Publishing.

Chicaiza, J., & Reátegui, R. (2020). 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). doi: 10.1007/978-3-030-65384-2_13

Choi, S., & Jun, S. (2014). Vacant technology forecasting using new bayesian patent clustering. *Technology Analysis & Strategic Management*, *26*(3), 241–251.

Chung, P., & Sohn, S. Y. (2020). Early detection of valuable patents using a deep learning model: Case of semiconductor industry. *Technological Forecasting and Social Change*, *158*, 120146.

Costa, J. P., Rei, L., Stopar, L., Fuart, F., Grobelnik, M., Mladenic, D., … Wallace, J. (2021). Newsmesh: A new classifier designed to annotate health news with mesh headings. *Artificial Intelligence in Medicine*, *114*, 102053. Retrieved from https://www.sciencedirect.com/science/article/pii/S0933365721000464 doi: https://doi.org/10.1016/j.artmed.2021.102053

Deng, W., Huang, X., & Zhu, P. (2019). Facilitating technology transfer by patent knowledge graph. In *Proceedings of the 52nd hawaii international conference on system sciences.*

Dessì, D., Osborne, F., Recupero, D. R., Buscaldi, D., & Motta, E. (2021). Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain. *Future Generation Computer Systems*, *116*, 253–264. doi: 10.1016/j.future.2020.10.026

Dörpinghaus, J., & Jacobs, M. (2020). Knowledge detection and discovery using semantic graph embeddings on large knowledge graphs generated on text mining results. In *2020 15th conference on computer science and information systems (fedcsis)* (p. 169-178). doi: 10.15439/2020F36

Färber, M. (2019). The microsoft academic knowledge graph: A linked data source with 8 billion triples of scholarly data. In *International semantic web conference* (pp. 113–129).

Fathalla, S., Auer, S., & Lange, C. (2020). Towards the semantic formalization of science. In *Proc. of 35th annual acm symposium on applied comp.* (pp. 2057–2059).

Grimpe, C., & Hussinger, K. (2013). Formal and informal knowledge and technology transfer from academia to industry: Complementarity effects and innovation performance. *Industry and Innovation*, *20*(8), 683-700. Retrieved from https://doi.org/10.1080/13662716.2013.856620 doi: 10.1080/13662716.2013.856620

Groth, P., Gibson, A., & Velterop, J. (2010). The anatomy of a nanopublication. *Information Services & Use*, *30*(1-2), 51–56.

Hanieh, A. A., AbdElall, S., Krajnik, P., & Hasan, A. (2015). Industry-academia partnership for sustainable development in palestine. *Procedia CIRP*, *26*, 109–114.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., … others (2021). Knowledge graphs. *ACM Computing Surveys (CSUR)*, *54*(4), 1–37.

Huang, M.-H., Yang, H.-W., & Chen, D.-Z. (2015). Industry–academia collaboration in fuel cells: A perspective from paper and patent analysis. *Scientometrics*, *105*(2), 1301–1318.

Jaradeh, M. Y., Auer, S., Prinz, M., Kovtun, V., Kismihók, G., & Stocker, M. (2019). Open research knowledge graph: Towards machine actionability in scholarly communication. *arXiv preprint arXiv:1901.10816*.

Jose, V., Jagathy Raj, V. P., & George, S. K. (2021). Ontology-based information extraction framework for academic knowledge repository. In X.-S. Yang, S. Sherratt, N. Dey, & A. Joshi (Eds.), *Proceedings of fifth international congress on information and communication technology* (pp. 73–80). Singapore: Springer Singapore.

Knoth, P., & Zdrahal, Z. (2011). Core: connecting repositories in the open access domain. In *Cern workshop on innovations in scholarly communication (oai7)*. Retrieved from http://oro.open.ac.uk/32560/ (Poster Session ID: 53)

Knoth, P., & Zdrahal, Z. (2012). Core: three access levels to underpin open access. *D-Lib Magazine*, *18*(11/12), 1–13.

Kuhn, T., Chichester, C., Krauthammer, M., Queralt-Rosinach, N., Verborgh, R., Giannakopoulos, G., … Dumontier, M. (2016). Decentralized provenance-aware publishing with nanopublications. *PeerJ C. S.*, *2*, e78.

Kuhn, T. S. (1962). The structure of scientific revolutions: University of chicago press. *Original edition*.

La Bruzzo, S., Manghi, P., & Mannocci, A. (2019). OpenAIRE's DOIBoost - Boosting Crossref for Research. In P. Manghi, L. Candela, & G. Silvello (Eds.), *Digital libraries: Supporting open science* (pp. 133–143). Cham: Springer International Publishing. doi: 10.1007/978-3-030-11226-4_11

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Larivière, V., Macaluso, B., Mongeon, P., Siler, K., & Sugimoto, C. R. (2018). Vanishing industries and the rising monopoly of universities in published research. *PLOS ONE*, *13*, 1-10. Retrieved from https://doi.org/10.1371/journal.pone.0202120

Ley, M. (2009). Dblp: some lessons learned. *Proceedings of the VLDB Endowment*, *2*(2), 1493–1500.

Löffler, F., Wesp, V., Babalou, S., Kahn, P., Lachmann, R., Sateli, B., … König-Ries, B. (2020). Scholarlensviz: A visualization framework for transparency in semantic user profiles. In K. Taylor, R. Gonçalves, F. Lecue, & J. Yan (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).*

Lula, P., Dospinescu, O., Homocianu, D., & Sireteanu, N.-A. (2021). An advanced analysis of cloud computing concepts based on the computer science ontology. *Computers, Materials & Continua*, *66*(3), 2425–2443. doi: 10.32604/cmc.2021.013771

Mannocci, A., Osborne, F., & Motta, E. (2019). The evolution of ijhcs and chi: A quantitative analysis. *International Journal of Human-Computer Studies*, *131*, 23–40.

Marinakis, Y. D. (2012). Forecasting technology diffusion with the richards model. *Technological Forecasting and Social Change*, *79*(1), 172–179.

Michaudel, Q., Ishihara, Y., & Baran, P. S. (2015). Academia–industry symbiosis in organic chemistry. *Accounts of Chemical Research*, *48*(3), 712-721. Retrieved from https://doi.org/10.1021/ar500424a (PMID: 25702529) doi: 10.1021/ar500424a

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). 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). USA: Curran Associates Inc. Retrieved from http://dl.acm.org/citation.cfm?id=2999792.2999959

Nayyeri, M., Cil, G. M., Vahdati, S., Osborne, F., Rahman, M., Angioni, S., … Lehmann, J. (2021). Trans4e: Link prediction on scholarly knowledge graphs. *Neurocomputing*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0925231221009607 doi: https://doi.org/10.1016/j.neucom.2021.02.100

Nuzzolese, A. G., Gentile, A. L., Presutti, V., & Gangemi, A. (2016). Semantic web conference ontology-a refactoring solution. In *European semantic web conference* (pp. 84–87).

Osborne, F., & Motta, E. (2015). Klink-2: integrating multiple web sources to generate semantic topic networks. In *Iswc* (pp. 408–424).

Osborne, F., Muccini, H., Lago, P., & Motta, E. (2019). Reducing the effort for systematic reviews in software engineering. *Data Science*, *2*(1-2), 311–340.

Osborne, F., Salatino, A., Birukou, A., & Motta, E. (2016). Automatic classification of springer nature proceedings with smart topic miner. In P. Groth et al. (Eds.), *The semantic web – iswc 2016* (pp. 383–399). Cham: Springer Int. Publishing.

Peroni, S., & Shotton, D. (2018). The spar ontologies. In *International semantic web conference* (pp. 119–136).

Peroni, S., & Shotton, D. (2020). Opencitations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, *1*(1), 428–444.

Powell, W. W., & Snellman, K. (2004). The knowledge economy. *Annual Review of Sociology*, *30*(1), 199-220. Retrieved from https://doi.org/10.1146/annurev.soc.29.010202.100037 doi: 10.1146/annurev.soc.29.010202.100037

Ramadhan, M. H., Malik, V. I., & Sjafrizal, T. (2018). 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).

Rossanez, A., dos Reis, J. C., & da Silva Torres, R. (2020). Representing scientific literature evolution via temporal knowledge graphs.

Saier, T., & Färber, M. (2020). unarxive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. *Scientometrics*, *125*(3), 3085–3108.

Salatino, A., Osborne, F., & Motta, E. (2020). Researchflow: Understanding the knowledge flow between academia and industry. In *Knowledge engineering and knowledge management – 22nd international conference, ekaw 2020.*

Salatino, A. A., Osborne, F., Birukou, A., & Motta, E. (2019). Improving editorial workflow and metadata quality at springer nature. In C. Ghidini et al. (Eds.), *The semantic web – iswc 2019* (pp. 507–525). Cham: Springer Int. Publishing.

Salatino, A. A., Osborne, F., Thanapalasingam, T., & Motta, E. (2019). The cso classifier: Ontology-driven detection of research topics in scholarly articles. In A. Doucet, A. Isaac, K. Golub, T. Aalberg, & A. Jatowt (Eds.), *Digital libraries for open knowledge* (pp. 296–311). Cham: Springer International Publishing.

Salatino, A. A., Thanapalasingam, T., & Mannocci, A. (2019). *angelosalatino/cso-classifier: CSO Classifier v2.3.2.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.2660819 doi: 10.5281/zenodo.2660819

Salatino, A. A., Thanapalasingam, T., Mannocci, A., Birukou, A., Osborne, F., & Motta, E. (2020). The computer science ontology: A comprehensive automatically-generated taxonomy of research areas. *Data Intelligence*, *2*(3), 379–416.

Salatino, A. A., Thanapalasingam, T., Mannocci, A., Osborne, F., & Motta, E. (2018a). Classifying research papers with the computer science ontology. In *Iswc (p&d/industry/bluesky). ceur workshop proceedings* (Vol. 2180).

Salatino, A. A., Thanapalasingam, T., Mannocci, A., Osborne, F., & Motta, E. (2018b). The computer science ontology: A large-scale taxonomy of research areas. In D. Vrandečić et al. (Eds.), *The semantic web – iswc 2018* (pp. 187–205). Cham: Springer Int. Publishing.

Saricaa, S., Luoab, J., & Woodab, K. L. (2019). Technology knowledge graph based on patent data. *CoRR*.

Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops* (p. 166-171). doi: 10.1109/ICDCSW.2011.20

Schneider, J., Ciccarese, P., Clark, T., & Boyce, R. D. (2014). Using the micropublications ontology and the open annotation data model to represent evidence within a drug-drug interaction knowledge base..

Schwartz, D. L., & Sichelman, T. (2019). Data sources on patents, copyrights, trademarks, and other intellectual property. In *Research handbook on the economics of intellectual property law.* Edward Elgar Publishing.

Shotton, D. (2009). Semantic publishing: the coming revolution in scientific journal publishing. *Learned Publishing*, *22*(2), 85–94.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J., & Wang, K. (2015). An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web* (pp. 243–246).

Stilgoe, J. (2020). Who's driving innovation? *New Technologies and the Collaborative State. Cham, Switzerland: Palgrave Macmillan*.

Thanapalasingam, T., Osborne, F., Birukou, A., & Motta, E. (2018). Ontology-based recommendation of editorial products. In D. Vrandečić et al. (Eds.), *The semantic web – iswc 2018* (pp. 341–358). Cham: Springer Int. Publishing.

Vergoulis, T., Chatzopoulos, S., Dalamagas, T., & Tryfonopoulos, C. (2020a). Veto: Expert set expansion in academia. In M. Hall, T. Merčun, T. Risse, & F. Duchateau (Eds.), *Digital libraries for open knowledge* (pp. 48–61). Cham: Springer International Publishing. doi: 10.1007/978-3-030-54956-5_4

Vergoulis, T., Chatzopoulos, S., Dalamagas, T., & Tryfonopoulos, C. (2020b). Veto: Expert set expansion in academia. In M. Hall, T. Merčun, T. Risse, & F. Duchateau (Eds.), *Digital libraries for open knowledge* (pp. 48–61). Cham: Springer International Publishing.

Visser, M., van Eck, N. J., & Waltman, L. (2020). *Large-scale comparison of bibliographic data sources: Scopus, web of science, dimensions, crossref, and microsoft academic.*

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020). Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, *1*(1), 396–413.

Wang, R., Yan, Y., Wang, J., Jia, Y., Zhang, Y., Zhang, W., & Wang, X. (2018). Acekg: A large-scale knowledge graph for academic data mining. In *Proceedings of the 27th acm international conference on information and knowledge management* (p. 1487–1490). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3269206.3269252 doi: 10.1145/3269206.3269252

Weinstein, L. B., Kellar, G. M., & Hall, D. C. (2016). Comparing topic importance perceptions of industry and business school faculty: Is the tail wagging the dog? *Academy of Educational Leadership Journal*, *20*(2), 62.

Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., . . . others (2013). The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. *Nucleic acids research*, *41*(W1), W557–W561.

Zang, X., & Niu, Y. (2011). The forecast model of patents granted in colleges based on genetic neural network. In *2011 international conference on electrical and control engineering* (pp. 5090–5093).

Zhang, X., Chandrasegaran, S., & Ma, K.-L. (2021). 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).

Zhang, Y., Zhang, F., Yao, P., & Tang, J. (2018). Name disambiguation in aminer: Clustering, maintenance, and human in the loop. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 1002–1011).

APPENDIX

We report in this appendix several exemplary SPARQL queries on AIDA. The aim is to show the flexibility of AIDA and the complexity of the queries that can be formulated. We also hope that these examples will offer a good starting point to the users that intend to reuse AIDA. All the following queries can be run on the AIDA SPARQL endpoint, available at http://w3id.org/aida/sparql.

The following performs a describe query for the paper with id 2040986908.

```
DESCRIBE <http://aida.kmi.open.ac.uk/resource/2040986908>
```

The following query returns all papers written by authors from the industrial sector *computing and it* associated with the topic *robotics*:

```
PREFIX aida-ont:<http://aida.kmi.open.ac.uk/ontology#>
PREFIX aida:<http://aida.kmi.open.ac.uk/resource/>
PREFIX aidaDB: <http://aida.kmi.open.ac.uk/resource/DBpedia/>
PREFIX cso: <http://cso.kmi.open.ac.uk/topics/>

SELECT ?paperId
FROM <http://aida.kmi.open.ac.uk/resource>
WHERE {
    ?paperId aida-ont:hasIndustrialSector aida:computing_and_it .
    ?paperId aida-ont:hasTopic cso:robotics .
}
LIMIT 20
```

The following query counts how many papers have been written by authors from an industrial affiliation.

```
PREFIX aida:<http://aida.kmi.open.ac.uk/ontology#>
SELECT (COUNT(?sub) as ?count)
FROM <http://aida.kmi.open.ac.uk/resource>
WHERE {
    ?sub aida:hasAffiliationType ”industry”
}
```

The next query counts how many authors are affiliated with *The Open University*.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX schema:<http://schema.org/>
SELECT (COUNT(DISTINCT(?sub)) as ?count)
FROM <http://aida.kmi.open.ac.uk/resource>
WHERE {
    ?sub schema:memberOf ?aff .
    ?aff foaf:name ”the_open_university”
}
```

The following query returns the industrial sectors of all the papers having *Semantic Web* as a topic.

```
PREFIX aida:<http://aida.kmi.open.ac.uk/ontology#>
PREFIX cso: <http://cso.kmi.open.ac.uk/topics/>
SELECT DISTINCT ?ind
FROM <http://aida.kmi.open.ac.uk/resource>
WHERE {
    ?sub aida:hasTopic cso:semantic_web .
    ?sub aida:hasIndustrialSector ?ind
}
```

The following query returns the papers associated with the topic *Semantic Web* and written in collaboration by authors from industry and academia, where those from academia are more than 80%.

```
PREFIX aida:<http://aida.kmi.open.ac.uk/ontology#>
PREFIX cso: <http://cso.kmi.open.ac.uk/topics/>
PREFIX schema: <http://schema.org/>
SELECT ?paper ?ind (count(?author) as ?nauthor)
FROM <http://aida.kmi.open.ac.uk/resource>
WHERE {
    ?paper aida:hasTopic cso:semantic_web .
    ?paper aida:hasIndustrialSector ?ind .
    ?paper aida:hasPercentageOfAcademia ?x .
    ?paper schema:creator ?author .
    FILTER(?x>80)
}
ORDER BY ?paper
```

The following query returns the number of publications in a topic (in this case *Neural Networks*) during the last five years. It can be used to analyse the trend of this topic in time.

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX prism: <http://prismstandard.org/namespaces/basic/2.0/>
PREFIX aida:<http://aida.kmi.open.ac.uk/ontology#>
PREFIX cso: <http://cso.kmi.open.ac.uk/topics/>

SELECT ?year (count(?paper) as ?n_publications)
FROM <http://aida.kmi.open.ac.uk/resource>
WHERE {
    ?paper aida:hasTopic cso:neural_networks .
    ?paper prism:publicationDate ?year .
    FILTER(xsd:integer(?year)>=2016 && xsd:integer(?year)<=2020)
} GROUP BY ?year
ORDER BY DESC(?year)
```

The following query returns the topic distribution of a given affiliation (in this case The Open University). It can be used to characterize an organization according to its relevant topics.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX aida:<http://aida.kmi.open.ac.uk/ontology#>
PREFIX schema: <http://schema.org/>
SELECT ?topic (count(distinct(?paper)) as ?count)
FROM <http://aida.kmi.open.ac.uk/resource>
WHERE {
    ?paper schema:creator ?author .
    ?author schema:memberOf ?aff .
    ?aff foaf:name ”the_open_university” .
    ?paper aida:hasTopic ?topic .
} GROUP BY ?topic
ORDER BY DESC(?count)
```

This query ranks affiliations according to their number of publications in a given topic (in this case *Semantic Web*):

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX aida:<http://aida.kmi.open.ac.uk/ontology#>
PREFIX cso: <http://cso.kmi.open.ac.uk/topics/>
PREFIX schema: <http://schema.org/>
SELECT ?aff ?aff_name (count(distinct(?paper)) as ?count)
FROM <http://aida.kmi.open.ac.uk/resource>
WHERE {
    ?paper aida:hasTopic cso:semantic_web .
    ?paper schema:creator ?author .
    ?author schema:memberOf ?aff .
    ?aff foaf:name ?aff_name
} GROUP BY ?aff ?aff_name
ORDER BY DESC(?count)
LIMIT 100
```

This query returns the academic affiliations that collaborates most (in terms of publication number) with industrial organizations:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX aida:<http://aida.kmi.open.ac.uk/ontology#>
SELECT ?aff ?name (COUNT(?paper) as ?n_collaborations)
FROM <http://aida.kmi.open.ac.uk/resource>
WHERE {
    ?paper aida:hasAffiliationType ”collaborative” .
    ?paper aida:hasAffiliation ?aff .
    ?aff aida:hasGridType ”education” .
    ?aff foaf:name ?name .
} GROUP BY ?aff ?name
ORDER BY DESC(?n_collaborations)
```

The following query returns the DBpedia concepts associated to a given paper (id: 2300368847 in this case) using the mapping between CSO and DBpedia.

```
PREFIX aida: <http://aida.kmi.open.ac.uk/ontology#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX aidar: <http://aida.kmi.open.ac.uk/resource/>
SELECT *
FROM <http://aida.kmi.open.ac.uk/resource>
WHERE {
    aidar:2300368847 aida:hasTopic ?topic .
    ?topic owl:sameAs ?obj .
    FILTER(regex(str(?obj), ”dbpedia” ) )
}
```