Integrating Knowledge Graphs for Analysing Academia and Industry Dynamics

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

[1] Department of Mathematics and Computer Science, University of Cagliari (Italy)
{simone.angioni, diego.reforgiato}@unica.it
[2] Knowledge Media Institute, The Open University, Milton Keynes (UK)
{angelo.salatino, francesco.osborne, enrico.motta}@open.ac.uk

Abstract. Academia and industry are constantly engaged in a joint effort for producing scientific knowledge that will shape the society of the future. Analysing the knowledge flow between them and understanding how they influence each other is a critical task for researchers, governments, funding bodies, investors, and companies. However, current corpora are unfit to support large-scale analysis of the knowledge flow between academia and industry since they lack of a good characterization of research topics and industrial sectors. In this short paper, we introduce the Academia/Industry DynAmics (AIDA) Knowledge Graph, which characterizes 14M papers and 8M patents according to the research topics drawn from the Computer Science Ontology. 4M papers and 5M patents are also classified according to the type of the author's affiliations (academy, industry, or collaborative) and 66 industrial sectors (e.g., automotive, financial, energy, electronics) obtained from DBpedia. AIDA was generated by an automatic pipeline that integrates several knowledge graphs and bibliographic corpora, including Microsoft Academic Graph, Dimensions, English DBpedia, the Computer Science Ontology, and the Global Research Identifier Database.

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

1 Introduction

Academia and industry are constantly engaged in a joint effort for producing scientific knowledge that will shape the society of the future. Analysing the knowledge flow between them and understanding how they influence each other is a critical task for researchers, governments, funding bodies, investors, and companies. Researchers have to be aware of how their effort impacts the industrial sectors; government and funding bodies need to shape research policies and funding decisions; companies have to constantly monitor the scientific innovation that may be developed in products or services.

The relationship between academia and industry has been analysed from several perspectives, focusing, for instance, on the characteristics of direct collaborations [4], the influence of industrial trends on curricula [16], and the quality

of the knowledge transfer [5]. Unfortunately, the lack of a large scale corpus for tracking knowledge flow limited the scope of previous works, which are typically restricted to small-scale datasets or focused on very specific research questions [6,2].

In order to analyse the knowledge produced by academia and industry, researchers typically exploit corpora of research articles or patents [4,3]. Today, we have several large-scale knowledge graphs which describe these documents. Some examples include Microsoft Academic Graph[3], Open Research Corpus [1], the OpenCitations Corpus [10], Scopus[4], AMiner Graph [17], the Open Academic Graph (OAG)[5], Core [7], Dimensions Corpus[6], and the United States Patent and Trademark Office Corpus[7]. However, these resources are unfit to support large-scale analysis about the knowledge flow since they suffer from three main limitations: 1) they do not directly classify a document according to its provenance (e.g., academia, industry), 2) they offer only coarse-grained characterizations of research topics, and 3) they do not characterize companies according to their sectors (e.g., automotive, financial, energy, electronics).

In this short paper, we introduce the Academia/Industry DynAmics (AIDA) Knowledge Graph, describing 14M articles and 8M patents (in English) in the field of Computer Science according to the research topics drawn from the Computer Science Ontology. 4M articles and 5M patents are also classified according to the type of the author's affiliations (academy, industry, or collaborative) and 66 industrial sectors (e.g., automotive, financial, energy, electronics) obtained from DBpedia. AIDA was generated by integrating several knowledge graphs and bibliographic corpora, including Microsoft Academic Graph (MAG), Dimensions, English DBpedia [8], the Computer Science Ontology (CSO) [14], and the Global Research Identifier Database (GRID)[8]. It can be downloaded for free from the AIDA website[9] under the CC BY 4.0 license.

AIDA was designed to allow researchers, governments, companies and other stakeholders to easily produce a variety of analytics about the evolution of research topics across academy and industry and study the characteristics of several industrial sectors. For instance, it enables detecting what are the research trends most interesting for the automotive sector are or which prevalent industrial topics were recently adopted and investigated by the academia. Furthermore, AIDA can be used to train machine learning systems for predicting the impact of research dynamics [11]. A preliminary versions of AIDA was used to support a comprehensive analysis of the research trends in the main venues of Human-Computer Interaction [9].

[3] https://aka.ms/msracad
[4] https://www.scopus.com/
[5] https://www.openacademic.ai/oag/
[6] https://www.dimensions.ai/
[7] https://www.uspto.gov/
[8] https://www.grid.ac/
[9] http://aida.kmi.open.ac.uk

2 Knowledge Graph on Academic and Industrial Dynamics

The Academia/Industry DynAmics Knowledge Graph describes a large collection of publications and patents in Computer Science according to the kind of affiliations of their authors (academia, industry, collaborative), the research topics, and the industrial sectors.

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

	Academia	**Industry**	**Collaboration**	**Total classified**	**Total**
Publications	3,043,863	730,332	108,506	3,882,701	14,317,130
Patents	133,604	4,741,695	16,335	4,891,634	7,940,034

Table 1 reports the number of publications and patents from academy, industry, and collaborative efforts. Most scientific publications (78.4%) are written by academic institutions, but industry is also a strong contributor (18.8%). Conversely, 96.9% of the patents are from industry and only 2.7% from academia. Collaborative efforts appears limited, including only 2.8% of the publications and 0.4% of the patents.

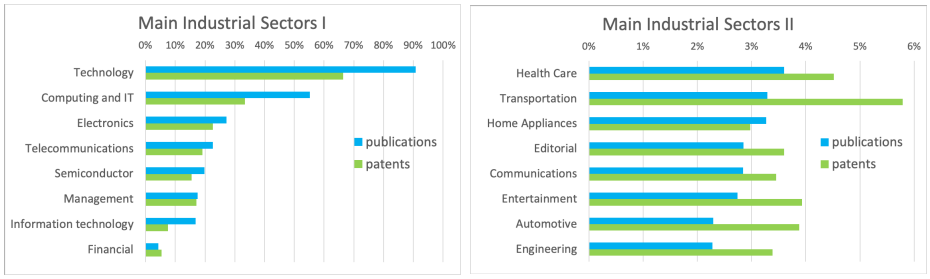

Fig. 1. Distribution of publications and patents in the main 16 industrial sectors.

Figure 1 reports the percentage of publications and patents associated with the most prominent industrial sectors. The most popular sectors in AIDA are directly pertinent to Computer Science (e.g., Technology, Computing and IT, Electronics, and Telecommunications, and Semiconductors), but we can also see many other sectors which adopt Computer Science technologies such as Financial, Health Care, Transportation, Home Appliance, and Editorial. The first group produces a higher percentage of publications, while the second generates more patents.

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

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

– *hasTopic*, which associates to the documents all their relevant topics drawn from CSO;

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

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

2.1 AIDA generation

AIDA was generated using an automatic pipeline that integrates and enriches data from Microsoft Academic Graph, Dimensions, Global Research Identifier Database, DBpedia, CSO [14], and INDUSO. It consists of three steps: i) topics detection, ii) extraction of affiliation types, and iii) industrial sector classification.

Topic Detection - *hasTopic* In this phase, we annotated each document with a set of research topics drawn from CSO: the intent is both to obtain a fine-grained representation of topics, with the aim of supporting large-scale analyses of research trends [12], and to have the same representation across the paper and the patents. The latter is critical since it allows to track the behavior of a topic according to different documents from academia and industry and assess its importance for the different industrial sectors.

As first step, we selected all the publications and patents from MAG and Dimensions within the domain of Computer Science. To achieve this, we extracted from MAG all the papers classified as "Computer Science" according to their classification: the Fields of Science (FoS) [15]. Similarly, we extracted from Dimensions all the patents pertinent to Computer Science according to the International Patent Classification (IPC) and the fields of research (FoR) taxonomy. The resulting dataset consists of 14M publications and 8M patents. Next, we run the CSO Classifier [13] on the title and the abstract of all the 22M documents. In addition to extracting the topics relevant to the text, we also exploited the same tool for including all their super topics according to the CSO. For instance, a paper tagged with *neural networks* was also assigned the topic *artificial intelligence*. This solution enables to monitor more abstracts and high level topics that are not always directly referred in the documents.

Extraction of Affiliation Types - *hasAffiliationType*, *hasAssigneeType* In this step, we classified research papers and patents according to the nature of their authors' affiliation in GRID, which is an open database identifying and typing over 90K organizations involved in research. Specifically, GRID describes research institutions with an identifier, geographical location, date of establishment, alternative labels, external links (including Wikipedia), and type of institution (e.g., Education, Healthcare, Company, Archive, Nonprofit, Government,

Facility, Other). MAG and Dimensions map a good number of affiliations to GRID IDs. We classified a document as 'academia' if all the authors have an educational affiliation and as 'industry' if all the authors have an industrial affiliation. Documents whose authors are from both academia and industry were classified as 'collaborative'.

Extraction of industrial category - *hasIndustrialSector* In this step, we characterised documents from industry according to the Industrial Sectors Ontology (INDUSO)[10], an ontology that we designed for this specific task. We designed INDUSO by merging and arranging in a taxonomy a large set of industrial sectors that we extracted from the affiliations of the paper authors and the patent assignees. First, we used the mapping between GRID and Wikipedia to retrieve the affiliations on DBpedia by extracting the objects of the properties "About:Purpose" and "About:Industry". This resulted in a noisy and redundant set of 699 sectors. We then manually analysed them with the help of domain experts and merged similar industrial sectors, finally obtaining 66 distinct sectors. For instance, the industrial sector "Computing and IT" in the resulting knowledge graph was derived from categories such as "Networking hardware", "Cloud Computing", and "IT service management". Finally, we designed INDUSO by arranging the 66 sectors in a two level taxonomy using the SKOS schema[11]. INDUSO also links the 66 main industrial sectors to the original 699 sectors using the *derivedFrom* relation from PROV-O[12].

Finally, we associated to each document all the industrial sectors that were derived from the DBpedia representation of its affiliations. For instance, the documents with an author affiliation described in DBpedia as 'natural gas utility' were tagged with the second level sector 'Oil and Gas Industry' and the first level sector 'Energy'.

3 Conclusions and Future Work

In this paper we introduced AIDA, the Academic/Industry DynAmics Knowledge Graph. AIDA includes knowledge on research topics of 14M publications extracted from MAG and 8M patents extracted from Dimensions. Moreover, 4M papers and 5M patents have also been classified according to the types of authors' and assignees' affiliations and 66 industrial sectors.

We are currently working on several next steps: i) we will provide our insights and analysis of research topic trends on academia and industry dynamics; ii) we are setting up a public triplestore to allow everyone to perform SPARQL queries to come up with further analytics and analysis out of the generated data; iii) we are setting up a pipeline that will automatically update AIDA with recent data; and iv) we will provide a rigorous evaluation of each component of the AIDA pipeline.

[10] INDUSO - `http://aida.kmi.open.ac.uk/downloads/induso.ttl`
[11] `https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html`
[12] `https://www.w3.org/TR/prov-o/`

References

1. Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., et al.: Construction of the literature graph in semantic scholar. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 84–91. Association for Computational Linguistics (2018)
2. Anderson, M.S.: The complex relations between the academy and industry: Views from the literature. The journal of higher education **72**(2), 226–246 (2001)
3. Angioni, S., Osborne, F., Salatino, A.A., Reforgiato, D., Recupero, E.M.: Integrating knowledge graphs for comparing the scientific output of academia and industry. In: International Semantic Web Conference ISWC 2019. pp. 85–88 (2019)
4. Ankrah, S., Omar, A.T.: Universities–industry collaboration: A systematic review. Scandinavian Journal of Management **31**(3), 387–408 (2015)
5. Ankrah, S.N., Burgess, T.F., Grimshaw, P., Shaw, N.E.: 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 (2013)
6. Bikard, M., Vakili, K., Teodoridis, F.: When collaboration bridges institutions: The impact of university–industry collaboration on academic productivity. Organization Science **30**(2), 426–445 (2019)
7. Knoth, P., Zdrahal, Z.: Core: three access levels to underpin open access. D-Lib Magazine **18**(11/12), 1–13 (2012)
8. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., et al.: Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web **6**(2), 167–195 (2015)
9. Mannocci, A., Osborne, F., Motta, E.: The evolution of ijhcs and chi: A quantitative analysis. International Journal of Human-Computer Studies **131**, 23–40 (2019)
10. Peroni, S., Shotton, D.: Opencitations, an infrastructure organization for open scholarship. Quantitative Science Studies **1**(1), 428–444 (2020)
11. Salatino, A., Osborne, F., Motta, E.: Researchflow: Understanding the knowledge flow between academia and industry (2020), `http://skm.kmi.open.ac.uk/rf-utkfbaai/`
12. Salatino, A.A., Osborne, F., Motta, E.: How are topics born? understanding the research dynamics preceding the emergence of new areas. PeerJ Computer Science **3**, e119 (2017)
13. Salatino, A.A., Osborne, F., Thanapalasingam, T., Motta, E.: The cso classifier: Ontology-driven detection of research topics in scholarly articles pp. 296–311 (2019)
14. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Birukou, A., Osborne, F., Motta, E.: The computer science ontology: A comprehensive automatically-generated taxonomy of research areas. Data Intelligence **0**(0), 1–38 (0). https://doi.org/10.1162/dint_a_00055, `https://doi.org/10.1162/dint_a_00055`
15. Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.J., Wang, K.: An overview of microsoft academic service (mas) and applications. In: Proceedings of the 24th international conference on world wide web. pp. 243–246 (2015)
16. Weinstein, L.B., Kellar, G.M., Hall, D.C.: Comparing topic importance perceptions of industry and business school faculty: Is the tail wagging the dog? Academy of Educational Leadership Journal **20**(2), 62 (2016)
17. Zhang, Y., Zhang, F., Yao, P., Tang, J.: 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 (2018)