IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 8, 2022, accepted March 29, 2022, date of publication April 11, 2022, date of current version April 18, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3166256

The AIDA Dashboard: A Web Application for Assessing and Comparing Scientific Conferences

SIMONE ANGIONI[1], ANGELO SALATINO[2], FRANCESCO OSBORNE[2,3], DIEGO REFORGIATO RECUPERO[1], AND ENRICO MOTTA[2]

[1]Department of Mathematics and Computer Science, University of Cagliari, 09124 Cagliari, Italy
[2]Knowledge Media Institute, The Open University, Milton Keynes MK7 6AA, U.K.
[3]Department of Business and Law, University of Milano Bicocca, 20126 Milan, Italy

Corresponding author: Simone Angioni (simone.angioni@unica.it)

This work was partially supported by the ASTRID project (Fondazione di Sardegna, L.R. 7 agosto 2007, n°7, CUP: F75F21001220007).

ABSTRACT Scientific conferences are essential for developing active research communities, promoting the cross-pollination of ideas and technologies, bridging between academia and industry, and disseminating new findings. Analyzing and monitoring scientific conferences is thus crucial for all users who need to take informed decisions in this space. However, scholarly search engines and bibliometric applications only provide a limited set of analytics for assessing research conferences, preventing us from performing a comprehensive analysis of these events. In this paper, we introduce the AIDA Dashboard, a novel web application, developed in collaboration with Springer Nature, for analyzing and comparing scientific conferences. This tool introduces three major new features: 1) it enables users to easily compare conferences within specific fields (e.g., Digital Libraries) and time-frames (e.g., the last five years); 2) it characterises conferences according to a 14K research topics from the Computer Science Ontology (CSO); and 3) it provides several functionalities for assessing the involvement of commercial organizations, including the ability to characterize industrial contributions according to 66 industrial sectors (e.g., automotive, financial, energy, electronics) from the Industrial Sectors Ontology (INDUSO). We evaluated the AIDA Dashboard by performing both a quantitative evaluation and a user study, obtaining excellent results in terms of quality of the analytics and usability.

INDEX TERMS Scholarly data, knowledge graph, conference analytics, bibliographic data, scholarly ontologies, science of science.

I. INTRODUCTION

Scientific conferences are essential for developing active research communities, promoting the cross-pollination of ideas and technologies, bridging between academia and industry, and disseminating new findings. This is particularly true in the fast-paced field of Computer Science, where conferences are usually the first venue in which researchers present new research efforts [1]. Indeed, each research area in Computer Science is typically associated with a set of conferences that help to define and evolve the main challenges and paradigms. Analyzing and monitoring scientific conferences is thus crucial for all users who need to take informed decision in this space, such as researchers, scientific editors, developers, government, funding bodies, and other relevant stakeholders.

The associate editor coordinating the review of this manuscript and approving it for publication was Sathish Kumar.

Current scholarly search engines and bibliometric applications provide a wide variety of functionalities to support the exploration of research data and produce various kinds of analytics. These include Semantic Scholar,[1] Dimensions[2] Scopus,[3] Web of Science,[4] AMiner,[5] and many others. However, these tools only provide a limited set of analytics and metrics for assessing research conferences, limiting our ability to perform a comprehensive analysis of these events.

In this paper, we focus on three main limitations of these systems. *First*, they do not support a granular comparison of all the conferences in a field according to various metrics in time. Google Scholar allows users to rank a limited set of

[1]Semantic Scholar - https://www.semanticscholar.org/
[2]Dimensions - https://app.dimensions.ai/discover/publication
[3]Scopus - https://www.scopus.com
[4]Web of Science - https://www.webofknowledge.com
[5]AMiner - https://www.aminer.cn/

conferences, but only according to a course-grained taxonomy of fields and one metric (h5-index). For instance, the field of Artificial intelligence is one of the leaf categories and includes only 20 conferences. Conversely, we would like to identify the main conferences in more specific fields, such as Neural Networks or Digital Libraries, how they rank in terms of average citations or number of publications, and how they evolved in the last few years.

Second, current tools do not allow users to analyze the research topics of a conference and their evolution over the years. Conversely, it can be argued that examining these trends is critical to assess the status of a conference and to predict its future performance.

Third, current systems do not take in consideration the industrial involvement in a conference. In particular, they do not report to which degree a conference attracts commercial organizations or what are the relevant industrial sectors. This is a significant missed opportunity since conferences are one of the premium public venues in which industry and academia interact and their analysis can offer important insights on how the research in a field is being carried out, supported, or reused by specific industrial sectors. For instance, large tech companies such as Alphabet (Google's parent company), Facebook (now Meta), Microsoft, and IBM became extremely active producing fundamental approaches in the field of Neural Networks in the last few years [2]. Also worth to note that reporting collaborations with non-academic partners is becoming an important metric for funding agencies. Knowledge institutions have to report those to both their funding agencies and the EU. This creates an incentive for academics to collaborate with the industry and to look for suitable venues.

In order to address these issues, we developed the *AIDA Dashboard*, a web application for analyzing and comparing scientific conferences which combines machine learning solutions, semantic technologies, and visual analytics. The AIDA Dashboard was developed in collaboration with Springer Nature with the aim of assisting editors in assessing conferences for informing editorial and business decisions. However, it evolved in a more general tool that can produce a wide range of analytics and support multiple use cases. In particular, in this paper we will assess the ability of the AIDA Dashboard to support researchers in the field of Computer Science.

The AIDA Dashboard introduces three novel features in order to address the limitations of current tools. First, it provides an interface for comparing and ranking conferences within specific fields (e.g., Digital Libraries) according to different metrics and time-frames (e.g., the last five years).

Second, it characterises conferences according to 14K research topics from the Computer Science Ontology[6] (CSO) [3]. This representation is used to produce several analytics about the evolution and impact of specific research topics. Second, it characterises conferences according to 14K research topics from the Computer Science Ontology (CSO). The reader notes that the CSO allows us to structure the research topics within the conferences according to a very granular representation [4]. For instance, the topic "Machine Learning" is composed of 760 more specific sub-topics, such as "Denoising Autoencoders" and "Fuzzy Neural Networks". This allows us to both offer a high-level representation that can be understood by less expert users, but also zoom in on very specific concepts and analyse their trends in time.

Finally, it enables users to analyse the involvement of industry in a conference by i) assessing the impact of commercial organizations across time, ii) reporting the ratio of publications from industry, academia, or collaborative efforts, and iii) categorising industrial contributions according to 66 industrial sectors (e.g., automotive, financial, energy, electronics) from the Industrial Sectors Ontology (INDUSO).[7]

The AIDA Dashboard builds on the *Academia/Industry DynAmics Knowledge Graph* (AIDA) [5], a new knowledge base that integrates information from Microsoft Academic Graph (MAG), Dimensions,[8] DBpedia,[9] the Computer Science Ontology (CSO), the Industrial Sectors Ontology (INDUSO), and the Global Research Identifier Database (GRID).[10] In order to associate conferences with research topics and industrial sectors we developed two unsupervised classifiers (described respectively in Section III-A and Section III-C) that consider the abstract of the articles, the relevant metadata, and additional information from DBpedia and domain ontologies (CSO, INDUSO).

The prototype of the AIDA Dashboard is available at http://w3id.org/aida/dashboard. The current version covers from 1990 to mid-2021. We are currently working on integrating up-to-date data. We plan to release the first official version in March 2022.

The AIDA knowledge graph can be downloaded under the CC-BY 4.0 license or queried (via SPARQL Endpoint) at http://w3id.org/aida. In order to support bibliometric studies, we also release the *AIDA Conference dataset*, a conference-centric version of the AIDA knowledge graph,[11] enriched with the new data generated for the AIDA Dashboard.

We evaluated the AIDA Dashboard by performing 1) a quantitative evaluation on the two approaches for classifying conferences according to their research topics and industrial sectors and 2) a user study involving ten researchers. The classifiers obtained results consistent with the ones of human experts according to statistical test, yielding an average F1 of respectively 91.8% and 87.5%. In the user study, the AIDA Dashboard obtained excellent results both in terms of quality of the analytics (scored 4.4/5 by researchers) and usability (87.5/100 according to the SUS questionnaire [6]). The data produced in the evaluation are publicly available at https://w3id.org/aida/downloads.

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

[7]INDUSO - http://w3id.org/aida/downloads/induso.ttl
[8]Dimensions - https://app.dimensions.ai
[9]DBpedia - https://wiki.dbpedia.org
[10]GRID - https://www.grid.ac
[11]AIDA Downloads - http://w3id.org/aida/downloads

In summary, our main contributions include:

- the AIDA Dashboard, a new web application for analyzing and comparing conferences in Computer Science;
- a pipeline for automatically generating several semantically-enriched analytics of scientific conferences which include two approaches for classifying conferences according to research areas and industrial sectors;
- a quantitative evaluation of the two approaches for associating research topics and industrial sectors to conferences;
- a qualitative evaluation of the usability of the AIDA Dashboard involving ten senior researchers;
- the AIDA Conference dataset, a resource describing 3,509 conferences in Computer Science according to all the data produced by the AIDA Dashboard back-end.

The rest of the paper is organized as follows. In Section II, we review the literature on systems and datasets for assessing scientific conferences. In Section III, we introduce AIDA and the pipeline used to generate it. In Section IV, we describe the AIDA Dashboard in details. Section V presents the quantitative evaluation and the user study. Section VI discusses a sustainability plan for the following years. Finally, Section VII ends the paper and outlines future directions of research.

II. RELATED WORK

In this section, we review the relevant literature focusing on three aspects: i) the knowledge graphs describing scholarly data, ii) tools for supporting the assessment of scientific conferences, and iii) scientometrics tools for assessing research trends.

A. SCIENTIFIC KNOWLEDGE GRAPHS

In recent years, we witnessed the emergence of several knowledge graphs describing research publications and their metadata, including Microsoft Academic Graph (MAG) [7], AMiner [8], ScholarlyData[12] [9], PID Graph[13] [10], SciGraph,[14] Open Research Knowledge Graph[15] [11], OpenCitations[16] [12], and OpenAIRE research graph[17] [13]. However, only few of these knowledge graphs contain information about scientific conferences.

MAG [7] is a heterogeneous, pan-publisher scholarly knowledge graph produced and actively maintained by Microsoft, which contains scientific publication records, citation relations, authors, institutions, and fields of study. Its metadata cover also journals and conferences, including conference series (e.g., NeurIPS) and specific conference editions (e.g., NeurIPS 2020). It is one of the most extensive datasets of scholarly data publicly available, and, as of March 2021, it contains more than 250 million publications [14].

The Semantic Scholar Open Research Corpus[18] [15] is a dataset of about 185M publications released by Semantic Scholar, an academic search engine provided by the Allen Institute for Artificial Intelligence. Information about conferences is available but not disambiguated in conference series and editions. The OpenCitations Corpus [12] is released by OpenCitations, which is 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. Information about venues is not often available.

Scopus is a well-known dataset maintained by Elsevier, which includes more than 80M publications. It is often used by governments and funding bodies to compute performance metrics. Although it is well-curated, its paper coverage is not as comprehensive as MAG [14], besides it mostly focuses on journals and less on conference proceedings.

The AMiner Graph [8] is a corpus of more than 200M publications generated and used by the AMiner system. AMiner is a free online academic search and mining system that also extracts researchers' profiles from the Web and integrates them into the metadata. It includes also conferences and journals metadata.

The Open Academic Graph[19] is a large knowledge graph integrating Microsoft Academic Graph and AMiner Graph. The current version contains 208M papers from MAG and 172M from AMiner and 91M links between the two graphs. This release includes also information about venues.

DBLP[20] [16] is a dataset of publications in Computer Science, which was originally created by the University of Trier and is now managed by Schloss Dagstuhl. It currently includes metadata about 5.5M articles, 2.7M authors, 5.4K conferences, and 1.7K journals.

CORE[21] [17] is a repository that integrates 24M open access research outputs from repositories and journals worldwide. The OpenAIRE dataset DOIboost[22] [18] is a similar integration effort that provides an enhanced version of Crossref and combines information from Unpaywall,[23] ORCID[24] and MAG, covering author identifiers, affiliations, organization identifiers, and abstracts. Conferences are currently not covered.

The Dimensions dataset is another well-known corpus which is produced by Digital Science, and interlinks 119M research publications, 6M grants, and 137M patents. Although Dimensions corpus includes a wide variety of metadata, it does not provide identifiers for conferences.

Another category of knowledge graphs offer a semantic representation of the content of scientific articles.

[12] ScholarlyData - http://www.scholarlydata.org
[13] PID Graph - https://www.project-freya.eu/en/pid-graph/the-pid-graph
[14] SciGraph datasets - https://sn-scigraph.figshare.com
[15] Open Research Knowledge Graph - https://www.orkg.org/orkg
[16] OpenCitations - https://opencitations.net
[17] OpenAIRE research graph - https://graph.openaire.eu
[18] ORC - http://s2-public-api-prod.us-west-2.elasticbeanstalk.com/corpus/
[19] Open Academic Graph - https://www.openacademic.ai/oag/
[20] DBLP - https://dblp.org
[21] CORE - https://core.ac.uk/
[22] DOIboost laster release - https://zenodo.org/record/3559699
[23] Unpaywall - https://unpaywall.org
[24] ORCID - https://orcid.org

The Semantic Web community has been working for a while on this direction, fostering the Semantic Publishing paradigm [19], creating bibliographic repositories in the Linked Data Cloud [20], generating knowledge bases of biological data [21], formalising research workflows [22], extracting knowledge graphs from research papers [23], [24], implementing systems for managing nano-publications [25], [26] and micropublications [27], and producing a variety of ontologies to describe scholarly data, e.g., SWRC,[25] BIBO,[26] BiDO,[27] FABIO,[28] SPAR,[29] CSO,[30] and SKGO[31] [30].

To develop the AIDA Dashboard we used MAG as main source for the articles since i) it is the most comprehensive publicly available knowledge graph [14], and ii) it includes a good representation of both conference editions and conference series. Since Microsoft decided to decommission MAG after 2021, we plan to switch to a combination of Dimensions and DBLP, as discussed in the sustainability plan (Section VI).

B. TOOLS FOR ASSESSING CONFERENCES

Several academic search engines and bibliometric tools allow users to explore the conference space. Microsoft Academic, which builds on MAG, provides several analytics about conferences. These include number of papers, citations, related conferences, main topics, publications, authors, and main institutions. However, it does not allow users to compare conferences or to analyse the evolution of research topics in time. AMiner and Semantic Scholar allow users to browse conferences, but they report only the most prominent authors and the relevant papers. Scholia[32] [31] is a Web service that creates scholarly profiles for topics, people, organizations, and venues according to the information in Wikidata.[33] When a conference is selected, Scholia reports all relevant proceedings, the main articles ranked by their citations, and the main topics, authors, and organizations. However, the data in Wikidata is far from being comprehensive. Moreover, the topics are associated with the conference series as a whole and thus they cannot be used to assess the evolution of the conference across time. The Scopus web application offers several analytics regarding researchers and articles. It links papers to conference proceedings, but does not aggregate the latter in conference items. Therefore, it is unable to support significant analyses on conferences. Lens.org[34] [32] is a web application that integrates data from MAG, Crossref, Core, and PubMed. It supports the analysis of several scholarly entities such as authors, institutions, contries, journal, conferences, topics, and others. Being based on MAG, it offers the same advantages and limitations of Microsoft Academic.

Overall, all these systems are limited by background data that offer only a coarse-grained representation of conferences and their relevant actors (e.g., authors, organizations, countries). For this reason, our first step in the creation of the AIDA Dashboard was the integration and enrichment of several knowledge graphs with the aim of creating more comprehensive metadata about scientific conferences.

Our aim, differently from the previous works, is to identify the main conferences in specific fields (e.g., Neural Networks or Digital Libraries instead of the general ones, like Artificial Intelligence), and analyse how they rank in terms of number of publications or average citations as well as whether their scope has changed over the years. To this end, given a conference, we determine its research topics and how they develop over time, so as to understand its status and support stakeholders in making data-informed decisions.

C. OTHER SCIENTOMETRIC TOOLS

In this section, we report additional state-of-the-art tools, which do not directly support the assessment of conferences but have the potential to be extended towards such a direction [33]–[37].

Van Eck *et al.* [33] developed VOSviewer, a tool for creating and visualising networks of publications, researchers, organizations, countries, keywords, and journals. VOSviewer takes as input bibliographic database files (e.g., from Dimensions or Scopus) and builds co-authorship, co-occurrence, citation, bibliographic coupling, or co-citation networks. Ideally, one can download a small dataset concerning a given conference and use such a tool to gain early insights on that conference.

Guilarte *et al.* [34] developed an interactive tool that leverages citations to visualise branches of science and identify main experts. Specifically, this tool has been applied to the problem of finding potential experts that act as peer reviewers of a target paper. This approach is based on the premise that if a target paper shares similar scientific issues or concerns with some of its references, then the authors of such references can be considered experts. This approach can be potentially extended to analyse whole conference proceedings, to assess the potential experts of that given conference, and even suggest who can act as a programme committee member.

Tosi *et al.* [35] developed SciKGraph, an approach that takes advantage of semantic technologies and natural language processing to structure research fields from research papers. Specifically, given a corpus of papers, it identifies their concepts and builds a knowledge graph based on their co-occurrence in papers. Concepts are then clustered to show how a scientific area is organised. This approach can be adapted to work on research papers of a single conference to identify its main areas and sub-areas, or analyse research papers of several conferences and identify the similar ones through their topical characterisation.

[25] SWRC - http://ontoware.org/swrc
[26] BIBO - http://bibliontology.com
[27] BiDO - http://purl.org/spar/bido
[28] FABIO - http://purl.org/spar/fabio
[29] SPAR - http://www.sparontologies.net/ [28]
[30] CSO - https://cso.kmi.open.ac.uk/ [29]
[31] SKGO - https://github.com/saidfathalla/Science-knowledge-graph-ontologies
[32] Scholia - https://scholia.toolforge.org
[33] Wikidata - https://www.wikidata.org
[34] Lens.org - https://www.lens.org

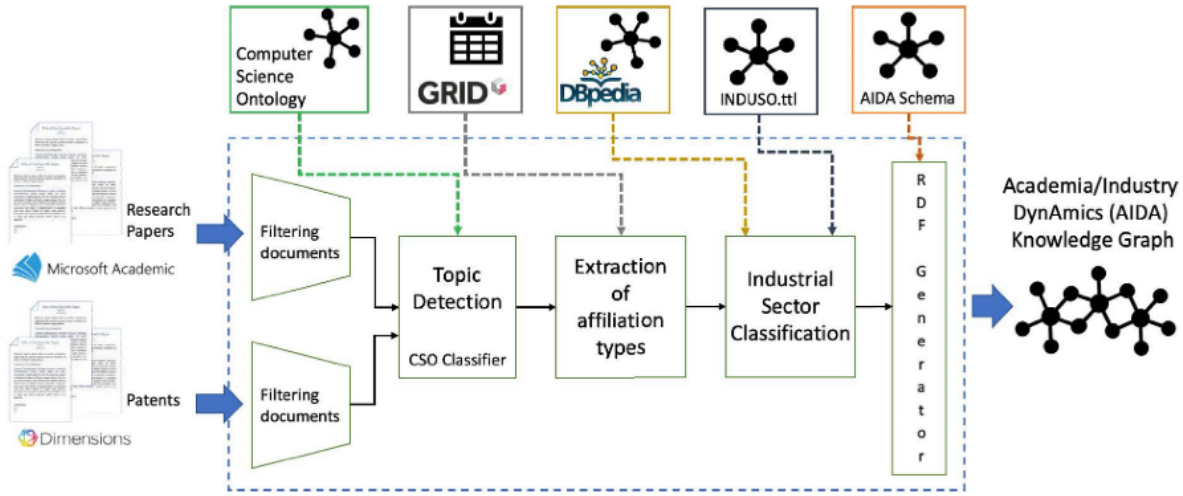

FIGURE 1. Workflow for the generation of the AIDA knowledge graph.

In general, although these approaches mainly focus on tasks that are different from analysing conferences, with a little adaptation they can support users in improving their understanding of conferences. On the other hand, the AIDA Dashboard focuses specifically on conferences and offers a more integrated suite of analytics in this space.

Furthermore, the above systems do not take into account how much a conference attracts industrial organizations or what relevant industrial sectors are attending the conference. Another goal of the presented dashboard is to analyse the involvement of the industrial sectors within conferences and research topics to provide useful information also to funding agencies.

III. THE AIDA KNOWLEDGE GRAPH

The Academia/Industry DynAmics knowledge graph (AIDA) describes 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). We generated this resource to support the computation of advanced analytics that are not available in current systems. Specifically, AIDA includes 21M publications from MAG and 8M patents from Dimensions according to the research topics drawn from the Computer Science Ontology (CSO). The 5.1M publications and 5.6M patents are also classified according to the type of the author's affiliations (e.g., academia, industry) and 66 industrial sectors from INDUSO.

The current version of the AIDA Dashboard focuses only on publications and does not use patents. These were included to enable more comprehensive analyzes, such as understanding the relationship between academia and industry [38], but are not yet available in the current version of the dashboard.

Figure 1 shows the automatic pipeline for generating the AIDA Knowledge Graph. The process includes three main steps: i) classification of articles with research topics from CSO, ii) detection of affiliation types, and iii) classification of articles with industrial sectors from INDUSO.

A. TOPIC CLASSIFICATION

In this phase, each document is classified with a set of research topics drawn from CSO. The purpose is to obtain a fine-grained representation of the topics with the aim of supporting large-scale analyses of research trends [39].

CSO is a large-scale ontology of research areas in the field of Computer Science that we developed in collaboration with Springer Nature [3]. The current version of CSO[35] includes 14K semantic topics and 159K relationships. CSO is used by several tools and proved to effectively support a wide range of tasks, such as exploring and analysing scholarly data (e.g., Rexplore [40], ScholarLensViz [41], ConceptScope [42]), identifying domain experts (e.g., VeTo [43]), recommending video lessons [44], and predicting academic impact (e.g., ArtSim [45], Augur [29]). CSO also supports several applications used by Springer Nature editorial team, such as Smart Topic Miner [46], a tool for assisting the classification of proceedings books, and the Smart Book Recommender [47], a recommender systems for scientific volumes.

As a first step, we select 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 the Fields of Science (FoS) [48] and from Dimensions all patents pertinent to Computer Science according to the International Patent Classification (IPC)[36] and the fields of research (FoR) taxonomy.[37] The resulting dataset consists of 20M publications and 8M patents.

In order to classify the documents according to CSO, we developed the CSO Classifier [49], an unsupervised classifier that we also released as a open-source Python tool.[38] This approach first identifies all topics that are explicitly

[35] CSO is available for download at https://w3id.org/cso/downloads

[36] International Patent Classification - https://www.wipo.int/classifications/ipc/en/

[37] Fields of Research - https://www.abs.gov.au/Ausstats/abs@.nsf/Latestproducts/6BB427AB9696C225CA2574180004463E

[38] CSO Classifier - https://pypi.org/project/cso-classifier/

mentioned in the paper (syntactic module) and then detects further semantically related topics by utilising part-of-speech tagging and word embeddings (semantic module). The syntactic module splits the text into n-grams and compares them with the topic labels in CSO using the Levenshtein similarity. The semantic module uses part-of-speech tagging to identify candidate terms composed by a combination of nouns and adjectives and retrieves their most similar words from a Word2Vec model [50]. We trained this model[39] using titles and abstracts of over 4.6 million English publications in the field of Computer Science from MAG. If the candidate terms are not available within the model vocabulary, the classifier uses the average of the embedding vectors of all its tokens. Then, it computes the relevance score for each topic in the ontology as the product between the number of times it was identified in those candidate terms (frequency) and the number of unique candidate terms that led to it (diversity). Finally, it uses the elbow method [51] for selecting the set of most relevant topics.

We run the CSO Classifier on the title and the abstract of all the 28M documents, associating each of them with the set of relevant topics.

We also include in the final representation all the super-topics according to the CSO. For instance, a paper tagged with *neural networks* was also assigned the topics *machine learning* and *artificial intelligence*. This solution enables monitoring more abstracts and high-level topics that are not always directly referred in the documents.

B. DETECTION OF AFFILIATION TYPES

In this step, the research papers and patents are classified according to the nature of their authors' affiliation in GRID, an open database identifying and typing[40] over 99K research organizations. 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 high number of affiliations to GRID IDs. A document is then classified as 'academia' or 'industry' based on whether all its authors have either an educational or industrial affiliation. Documents whose authors have mixed affiliations from both academia and industry (e.g., one author from academia and one author from industry or one author with multiple affiliations in academia and industry), are classified as 'collaborative effort'.

C. INDUSTRIAL SECTOR CLASSIFICATION

In this step, we classify the documents from industry according to INDUSO, an ontology of industrial sectors that we designed for this specific task.

We created INDUSO by integrating and structuring a large set of industrial sectors that we extracted from the affiliations of both papers' authors and patents' assignees. We took advantage of the mapping between GRID and Wikipedia to first retrieve the affiliations sectors from DBpedia using the properties "About:Purpose", "About:Industry".

This resulted in a redundant and noisy set of 699 sectors. With the help of domain experts, we manually analyzed and merged similar industrial sectors, obtaining a final set of 66 distinct sectors. For instance, the industrial sector "Computing and IT" in the resulting representation was derived from categories such as "Networking hardware", "Cloud Computing", and "IT service management". Then, using the SKOS schema,[41] we designed INDUSO by arranging the 66 sectors in a two-level hierarchy, with 27 first level sectors subsuming 39 second level sectors. These 66 main industrial sectors are also linked to the original 699 DBpedia sectors using the *wasDerivedFrom* relation within PROV-O.[42]

In order to associate research papers with the corresponding industrial sectors in INDUSO, we developed a simple unsupervised classifier. This approach retrieves the strings representing the affiliations of the authors in AIDA-KG and matches them to the most relevant entries in GRID, and then retrieves the relevant entities in DBpedia. For instance, given an author affiliated with "Apple Inc, Cupertino", it detects the corresponding GRID institution "Apple" (https://www.grid.ac/institutes/grid.455360.1) and retrieves its Wikipedia URL (https://en.wikipedia.org/wiki/Apple_Inc.). This is used to query the relevant DBpedia entity (https://dbpedia.org/data/Apple_Inc..ttl) and retrieving the relevant triples (e.g., *<dbr:Apple_Inc., dbp:industry, dbr:Computer_hardware>*, *<dbr:Apple_Inc., dbp:industry, dbr:Consumer_electronics>*). Finally, it uses the mapping between DBpedia and INDUSO to assign to the article the corresponding sectors (e.g., "Electronics", "Manufacturing", "Financial", "Computing and IT") as well as all their super-categories in INDUSO (e.g., "Technology", "Engineering"). Therefore, each document may be associated with multiple categories in INDUSO, depending on the sectors of its industrial affiliations.

IV. THE AIDA DASHBOARD

The AIDA Dashboard is a web application which builds on AIDA to generate several interactive analytics about conferences in Computer Science.

One of the main concerns when creating the AIDA dashboard was its scalability. Our objective was to develop a system that could run on an average server and at the same time be used by a large number of users with no significant slowdowns. To this purpose, we adopted a simple architecture composed by a back-end developed in Python and a front-end developed in HTML5 and JavaScript. Periodically, the *back-end* precomputes all the analytics and represents

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

[40] With typing we mean assign a type to a given entity.

[41] SKOS - https://www.w3.org/2004/02/skos/

[42] The PROV Ontology - https://www.w3.org/TR/prov-o/

each conference as a single JSON file. It can be queried for a specific conference and it will return the relevant JSON file. The *front-end* is a web application that allows the user to select a conference, loads the required JSON file from the back-end, and produces a set of interactive views. This solution is extremely lightweight and easy to maintain. The obvious drawback is that it is not possible to run live queries. Therefore, we precomputed a large number of analytics and rankings in order to support most use cases in this space.

In the following we will describe the back-end (Section IV-A) and front-end (Section IV-B) of the AIDA Dashboard.

A. THE BACK-END: GENERATION OF THE AIDA CONFERENCE DATASET

The back-end of the AIDA Dashboard iterates on the conferences in AIDA, for each of them computing a set of analytics, and storing the outcome in a collection of JSON files. All the information about a specific conference is thus contained in a single file identified by the conference ID in AIDA. We plan to perform this computation every two months. We label the resulting dataset *The AIDA Conference Dataset* and release it to the wide community. The aim is to support other tools as well as further scientometrics analysis. The current version is available at http://aida.kmi.open.ac.uk/downloads. We plan to release regular updates of this dataset, every six months.

The AIDA Conference dataset describes a conference according to: 1) a set of general metrics, 2) the top authors, organizations, countries, and topics associated with different metrics in time, 3) information about the dynamics between academia and industry in the conference, and 4) the focus areas of the conference. The *focus areas* are a set of high-level topics that the AIDA Dashboard uses for comparing similar conferences. In the following we will detail the process for generating these data from the AIDA knowledge graph.

First, given an input conference, we query the AIDA knowledge graph to gather information such as the name of the conference, its acronym, when it was held, and the total number of publications and citations received by the articles published in its proceedings over the years.

The latter are used to compute h-index, h5-index and the impact factor (over the last 2 years). We compute all these metrics considering the set of papers accepted by the conference, following the same procedure of other systems in this space such as Google Metrics. For instance, we calculate the h5-index over the set of articles published in the conference during the last 5 years.

We then count the number of publications and citations associated with four categories of scholarly items: authors, organizations, countries, and topics. Next, we select the top 100 of each category in terms of publications and the top 100 in terms of citations. Each of the resulting item is associated with its number of publications and citations across the years. For some categories (e.g., authors, organizations) their h-index and h5-index were also computed. Since the distribution of the main topics tends to include several generic high-level topics even when they are under-represented in the specific conference, we also extract an additional set labelled *fingerprint topics*. These are the top 100 topics that in the conference received a percentage of publications and citations higher than their average in the whole Computer Science domain. They are selected by computing the difference between the distribution of topics in the conference and the distribution of the same topics in the whole computer science domain. For instance, the topic *machine learning* is assigned 40% for NeurIPS (Neural Information Processing Systems) because in this conference it appears in about 60% of the articles, while it appears in 20% of the papers in Computer Science.

We then compute the number of publications and citations received from the research papers written by academia, industry, and collaborations, and by the most active industrial sectors.

Finally, we associated the input conference with its main focus areas. Each conference receives a rank in each of these areas based on their average citations in a time interval. For instance, NeurIPS was associated with the focus areas: Neural Networks (2nd overall in the last five years), Machine Learning (2nd), and Artificial Intelligence (5th). The rank allows the users to easily determine the importance of a conference in a field.

In the next paragraph we will describe the algorithm to generate focus areas of a given conference.

FOCUS AREAS GENERATION

Algorithm 1 shows the pseudo-code for identifying the focus areas of an input conference. The main purpose of this approach is to determine the research topic that is the most representative of the conference and then returns it together with its super-topics. Simply selecting the topic with the highest frequency is not a good solution since high-level topics are associated with all the publications of their sub-topics. For instance, a naïve algorithm based on frequency may assign to NeurIPS the focus area *artificial intelligence*, ignoring what component of AI is more prominent in this case. Conversely, we may detect that the large number of publications associated with AI is mainly due to the prominence of the sub-topic *machine learning*, and in turn that the majority of articles associated with this area are from the specific sub-topic *neural networks*. Therefore, our approach first orders the topics according to their number of publications (line 1). Topics are then (line 2) filtered by using a whitelist. Next, we fetch (line 3) the total number of publications of the conference. The algorithm iterates on all the topics (line 6) and selects the first topic as candidate focus area (lines 9-11). For the other topics, it checks whether it is a descendant of the current candidate (first condition, line 12), and if it is the main reason for its high frequency of publications (second condition, line 12). It does so by assessing if the percentage of the candidate publications associated also with the sub-topics is higher than a threshold (line 12, $subtopic\_thr = 0.6$ in the prototype). If this is the case, it selects the sub-topic as new

Algorithm 1 Focus Areas Generation

Input : Conference ID *conference*, Threshold for taking a sub-area *subtopic_thr*, Whitelist of areas *whitelist*
Output: Set of Focus Areas *focus_areas*

```
1  topics ← getConfSortedTopics (conference)
2  topics ← filter (whitelist, topics)
3  publications_c ← getTotalPublications (conference)
4  candidate ← NULL
5  candidate_impact ← 0
6  foreach topic in topics do
7      publications_t ← getNumberOfPubs (topic,
       conference)
8      impact = publications_t/publications_c
9      if candidate is NULL then
10         candidate ← topic
11         candidate_impact ← impact
12     else if (topic is descendant of the candidate) AND
       (impact/candidate_impact > subtopic_thr) then
13         candidate ← topic
14         candidate_impact ← impact
15 focus_areas ← expand (candidate)
16 return focus_areas
```

candidate (lines 13-14). Finally, it returns (lines 15-16) the last candidate topic (e.g., neural networks) and all its super topics (e.g., machine learning, artificial intelligence). When computing the focus areas for all conferences in Computer Science, the whitelist was first initialised to the full set of topics in CSO. We then analyzed the distribution of the resulting focus areas and generated a whitelist including the 166 focus areas that were associated with at least 5 conferences. The purpose of this operation is to discard minor areas not useful for comparing a fair number of conferences and obtain a representative whitelist which we feed to sequent executions of the algorithm. This whole process takes a few minutes on an average machine and it is processed offline once a year.

B. THE WEB INTERFACE

The Web interface of the AIDA Dashboard allows users to search for the full name or the acronym of a conference using an autocomplete field. When a conference is selected, it loads the corresponding JSON file from the back-end. It then produces interactive views of the resulting analytics structured in eight tabs: *Overview*, *Citation Analysis*, *organizations*, *Countries*, *Authors*, *Topics*, *Related Conferences*, and *Industry*.

The **Overview** tab is the introductory page of a conference, where the user is first redirected. It provides general information about the conference performance and trends. Figure 2 shows as example the Overview tab of the NeurIPS conference. This page is organized in two sections. The bar on the left gives information and metrics (e.g., the period of activity, the total number of publications and citations, the h5-index) about the underlying conference. It also provides general information about the average h-index of the organizations and authors who published in the conference as well as the average citations received by the published papers. In the lower part, it reports the focus areas and the rank of the conference in each of them (according to the average citations in the last 5 years). The section on the right provides several charts about the number of publications and citations over the years, the main authors and organizations in terms of publications (in the last 10 years), and the top fingerprint topics in terms of publications and citations (in the last 10 years).

The **Citation Analysis** tab reports the evolution in time of several citation-based metrics such as the impact factor and the average citations for paper. It also shows the evolution of the rank and the percentile of the conference in the focus areas. For instance, in Figure 3 we can see that NeurIPS has been among the top two conferences in Neural Networks and Machine Learning and the top ten conferences in Artificial Intelligence for the last 20 years. This visualization is typically used by Springer Nature editors to assess the performance of conferences within different communities.

The **Organizations** tab shows several analytics about the main institutions active in the conference. In this section the users can assess the main organizations according to their number of publications, citations, and average citation. Organizations can also be filtered according to their types (academia, industry, or all). The default interface used by the dashboard for reporting these data is a bar chart in which each item is associated with the total of the metric in a period (e.g., last five years). The user can also change this view (using the 'time-based' button) to a line-chart showing the same data across the years, which allows users to easily analyze trends in time.

The **Authors** tab uses the same interface for displaying the main researchers associated with their number of publications, citations, and average citations. The researchers can also be sorted by their overall H-index and H5-index, in order to quickly identify high impact researchers. Figure 4 shows the authors from NeurISP ordered according to their number of citations in the last five years. Editors at Springer Nature typically use the Organizations and Authors tabs to assess the quality of researchers and organizations attracted by the conferences. This is particularly important for assessing relatively young conferences that may not yet have developed a strong citation record.

The **Countries** tab allows the users to analyze the contribution of specific countries. The user can switch between the Chart view and the Map view. The first one shows the set of countries according to their number of publications, citations, and average citations. The second view arranges the information about the frequency of articles by country in a world map.

The **Topic** tab allows the users to analyze the topic trends over time. Specifically, it shows two selections of topics: main topics and fingerprint topics, discussed earlier in the paper. Figure 5 shows the main topics of NeurIPS. On the left side we indicate the percentage of publications in which

FIGURE 2. AIDA Dashboard - the overview tab of the NeurIPS conference.

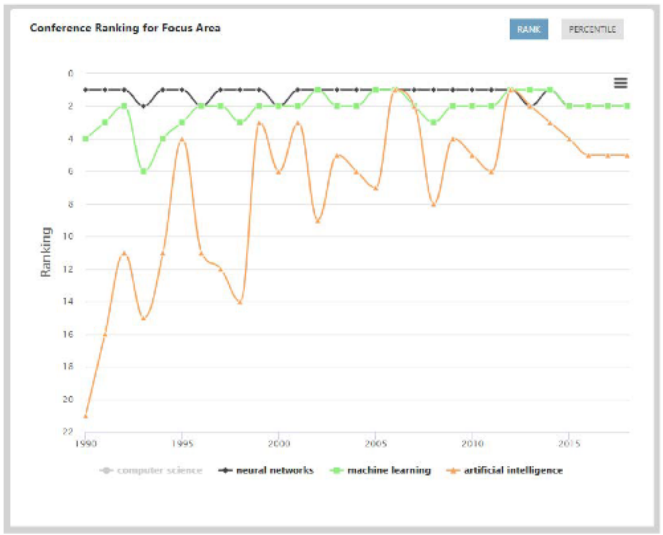

FIGURE 3. Portion of the citation analysis tab - The ranking of NeurIPS its focus areas.

FIGURE 4. Portion of the Authors tab - Authors ranked by citations in NeurIPS.

the underlying topic appears. On the right side we show the number of citations received by articles in which the topic appears.

The **Related Conferences** tab allows the users to compare the underlying conference against all the others in the same fields according to their number of publications, citations, and average citations for paper. The user can contextualise the comparison to different fields. For example, the NeurIPS conference can be compared with all the other conferences in the fields of Neural Networks, Machine Learning, and Artificial Intelligence. Figure 6 shows the comparison of NeurIPS with the other top conferences in Artificial Intelligence. The conference in analysis is highlighted in red.

Finally, the **Industry** tab reports the number of publications and citations from academia, industry, and collaborative efforts as well as the industrial sectors analysis. The latter shows the percentage of produced publications and citations received by companies in different industrial sectors. Figure 7 shows the trend of publications received by companies in different industrial sectors.

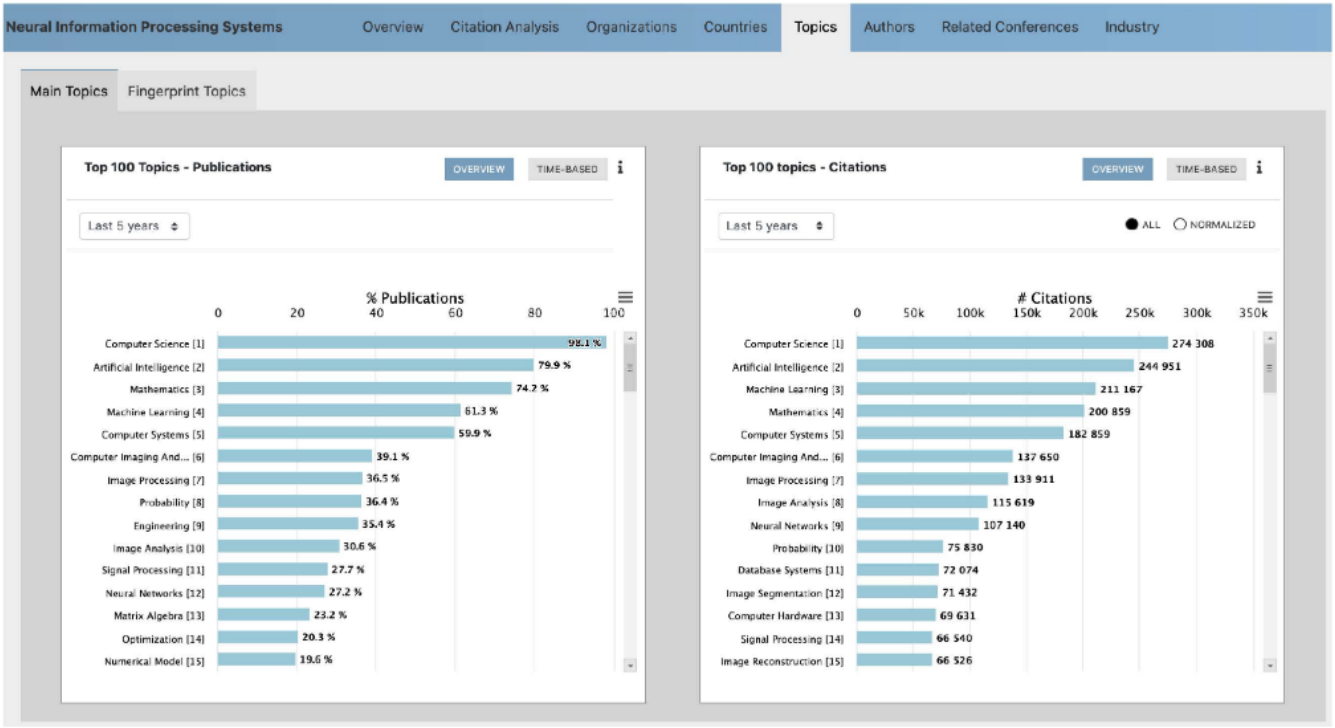

FIGURE 5. AIDA dashboard - the Topics tab of the NeurIPS conference.

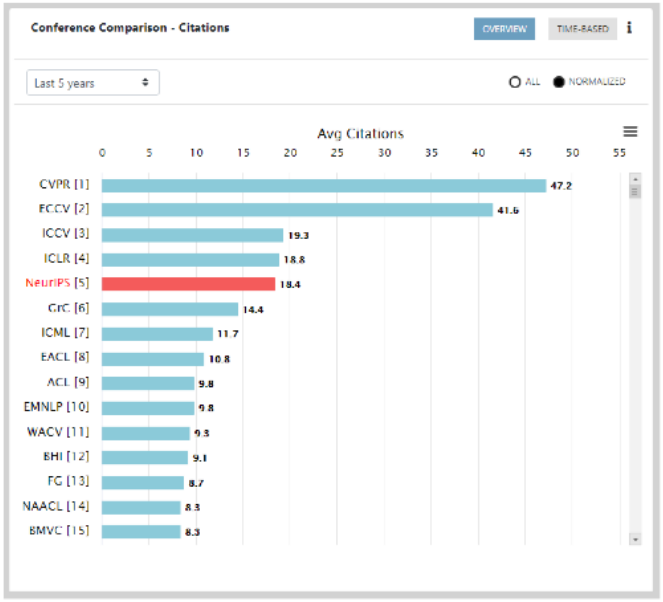

FIGURE 6. Portion of the related conferences tab - Conferences in artificial intelligence ranked by average citations.

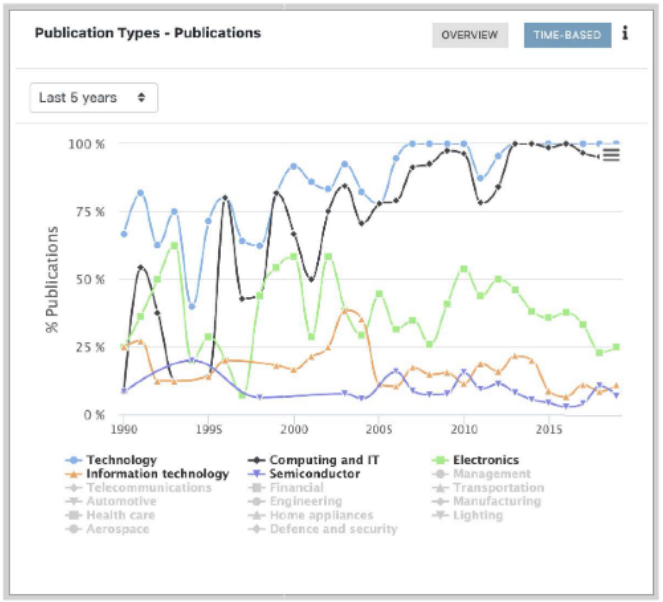

FIGURE 7. Portion of the industry tab - The main industrial sectors in NeurIPS across time. The percentage indicates the fraction of papers published in the corresponding year by companies of the underlying industrial sector.

V. EVALUATION

In this section, we present the quantitative evaluation of the two approaches for classifying conferences according to research topics and industrial sectors (Section V-A) and discuss the results of a user study involving 10 senior researchers (Section V-B).

A. QUANTITATIVE EVALUATION

We report the evaluation of the two approaches for classifying conferences according to their focus areas (Section V-A1) and for classifying articles according to industrial sectors (Section V-A2). We do not present here an evaluation of the approach for classifying articles according to the different topics since the CSO Classifier was comprehensively evaluated in [49], yielding excellent results against several state-of-the-art alternatives.

1) CLASSIFICATION OF FOCUS AREAS

In order to evaluate our approach to identify the focus areas of a conference we created two human-crafted gold

TABLE 1. Performance of the focus area classification task.

	Type	Precision	Recall	F1-score
Group 1	Artificial Intelligence	0.900	0.750	0.818
	Computer Hardware	0.900	1.000	0.947
	Human Computer Interaction	1.000	0.909	0.952
	Software Engineering	1.000	1.000	1.000
	Internet	0.700	0.875	0.777
	Macro Average	0.900	0.906	0.899
	Weighted Average	0.910	0.900	0.900
Group 2	Machine Learning	1.000	0.769	0.869
	Genetic Algorithm	0.700	1.000	0.823
	Formal Logic	1.000	1.000	1.000
	Knowledge Based Systems	1.000	1.000	1.000
	Multi-agent System	1.000	1.000	1.000
	Macro Average	0.940	0.953	0.938
	Weighted Average	0.958	0.940	0.941

standards. The first one focuses on high-level fields and contains 50 conferences manually annotated with five direct sub-topics of Computer Science in CSO (Artificial Intelligence, Computer Hardware, Software Engineering, Software Engineering, Internet). The second one addresses more specific fields and includes 50 conferences manually annotated with five direct sub-topics of Artificial Intelligence in CSO (Machine Learning, Genetic Algorithm, Formal Logic, Knowledge Based Systems, Multi-agent System).

Each conference was annotated by three senior researchers in Computer Science. In case of disagreement we used the majority voting strategy for defining the correct assignment. There were no cases in which the three annotators chose three different options.

Table 1 reports the performance of the approach described in Section IV-A on the two gold standards. Our solution performs very well obtaining an average F1 of respectively 89.9% and 93.8%. We performed a statistical analysis to test the hypothesis that the automatic classifier performed in line with the human expert. To this end we computed the average Cohen's concordance coefficient k for mixed pairs human-algorithm obtaining 0.83 (95% CI[43] 0.76-0.91). This result is consistent with the results for pairs of human readers $k = 0.77$ (95% CI 0.68-0.86). Since the two 95% CIs overlap, we can conclude that the results of the automatic approach are not significantly different from the ones of the experts.

2) CLASSIFICATION OF INDUSTRIAL SECTORS

In order to correctly classify the industrial sectors of a document we collected 100 organizations equally split by using the process mentioned within Section III-C (i.e., 20 organizations per industrial sector) among five classes: telecommunication, healthcare, automotive, computing and information technology, and electronic. Three senior researchers were asked to assign one of the classes above to each organization (or *other* in case the organization had different typology). We then employed the majority voting technique to come up with the gold standard: e.g., if a certain company was labelled with electronic by at least two annotators then we assigned the electronic class to it. We did not have any case where the three annotators gave three different classes. Table 2 shows the precision, recall and F1-score of our method using the gold standard just described. As before, our solution obtains very good results with an average F1 of 87.5%. The average Cohen's concordance coefficient for mixed pairs human-algorithm is 0.79 (95% CI 0.69-0.88) while that between human readers is 0.86 (95% CI 0.79-0.84). This suggests that also in this case our approach yields results not significantly different (the 95% CI overlap) from those of the experts.

TABLE 2. 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

B. USER STUDY

We performed a user study on the AIDA Dashboard in order to assess the quality and usefulness of the analytics as well as the usability of the user interface. To this end, we organized individual sessions with 10 researchers in Computer Science.

In each session, we first presented the AIDA Dashboard describing the new functionalities for about 15 minutes. We then assigned to them the task of analysing two conferences within their expertise in order to asses the quality of the resulting analytics. After the hands-on session the researchers filled a five-parts survey about their experience. The first part assessed the editor background and expertise. The second part was a standard System Usability Scale (SUS)[44] [6] questionnaire to assess the usability of the AIDA dashboard. The third section asked the researchers to rate the quality of the analytics for the two chosen conferences on a [1-5] scale. The fourth part included seven open questions about strengths and weaknesses of the AIDA Dashboard. Finally, the fifth part asked the user to list at least three of the most useful functionalities of the dashboard.

In the following sections we discuss the results in details.

1) USER BACKGROUND

Users were chosen among senior researchers within the Computer Science departments of the Open University (UK) (2 researchers), the University of Cagliari (IT) (3 researchers), the National Council of Research (IT) (1 researcher), FIZ Karlsruhe – Leibniz (DE) (2 researchers), University of Paris 13 (FR) (1 researcher), and the Institute for Applied Informatics (DE) (1 researcher). As far as the gender distribution is concerned, four of them were women and six men.

[43] A 95% CI (Confidence Interval) is the interval within which we can be 95% confident that the true population value for k is actually included.

[44] System Usability Scale (SUS) - https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

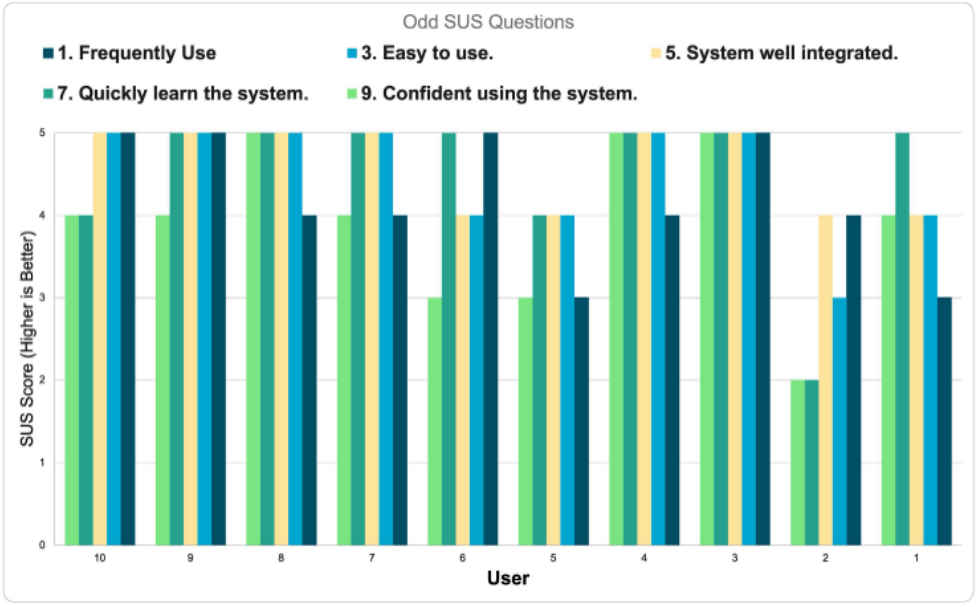

FIGURE 8. SUS questionnaire results (positive questions).

They had on average 7.8 years of experience as researcher. Four out of ten had at least 13 years of experience. Five of them had also experience in organizing conferences, workshops, special issues, and similar events. Their topics of expertise included Artificial Intelligence, Natural Language Processing, Digital Libraries, Semantic Web, Robotics, Information Retrieval, Human Computer Interaction, Computer Vision, and Theoretical Computer Science.

2) SUS QUESTIONNAIRE

The SUS questionnaire provided excellent results scoring 87.5/100, which is equivalent to A+ grade and places AIDA Dashboard in the 98% percentile rank.[45]

Figure 8 and Figure 9 show the answers of the users to the SUS questionnaire. The odd questions are positive (a higher score is better) while the even ones are negative (a lower score is better).

The users considered the AIDA Dashboard easy to use (with an average score of 4.5 ± 0.71) and believed its functions were well-integrated (4.6 ± 0.51). They thought it was not complex to use (1.5 ± 0.71) and that they would not need help to use it in the future (1.2 ± 0.42). The SUS also reported that most of the users felt very confident when using the dashboard (3.9 ± 0.99) and would be happy to use it frequently (4.2 ± 0.78). In addition, most users thought that they could learn to use the dashboard very quickly (4.5 ± 0.97) since it does not require learning a lot of new concepts (1.4 ± 0.97). Finally, they thought that the system was not inconsistent (1.4 ± 0.52) nor cumbersome (1.2 ± 0.63).

3) QUALITY ASSESSMENT

We asked the researchers to evaluate the quality of the analytics produced by the AIDA Dashboard for the two chosen conferences and rank them on a scale from 1 to 5.

TABLE 3. Conference quality assessment.

User	*Conference 1*	*Vote*	*Conference 2*	*Vote*
1	KDIR	4	ESWC	4
2	K-CAP	5	ESWC	4
3	EMNLP	5	ACL	5
4	IC3	5	ICNLP	5
5	KDIR	4	AAAI	3
6	ICRA	5	IJCAI	5
7	EACL	4	SIGIR	4
8	ISWC	5	TheWebConf	5
9	AAAI	4	ISWC	4
10	TheWebConf	4	TPDL	4

The list of conferences included top venues in the fields of Artificial Intelligence (AAAI, IJCAI, IC3), Natural Language Processing (EMNLP, ICNLP, ACL, EACL), Semantic Web (ISWC, ESWC, K-CAP), Information Retrieval (SIGIR, KDIR), the Web (TheWebConf), Robotics (ICRA), and Digital Library (TPDL). The average score was 4.4 ± 0.6, suggesting that the users were positively impressed by the usefulness of the functionalities and attractiveness of the analytics.

4) OPEN QUESTIONS

In this section we summarise the answers to the open questions.

a: Q1. HOW DO YOU FIND THE INTERACTION WITH THE AIDA DASHBOARD INTERFACE?

Five researchers considered it "easy to use", two "user friendly" and "intuitive", and two researchers were positive about it. One of them suggested that the abundance of functionality and analytics could lead to confusion and suggested to add more tooltips to explain all the available options.

[45] Interpreting a SUS score - https://measuringu.com/interpret-sus-score/

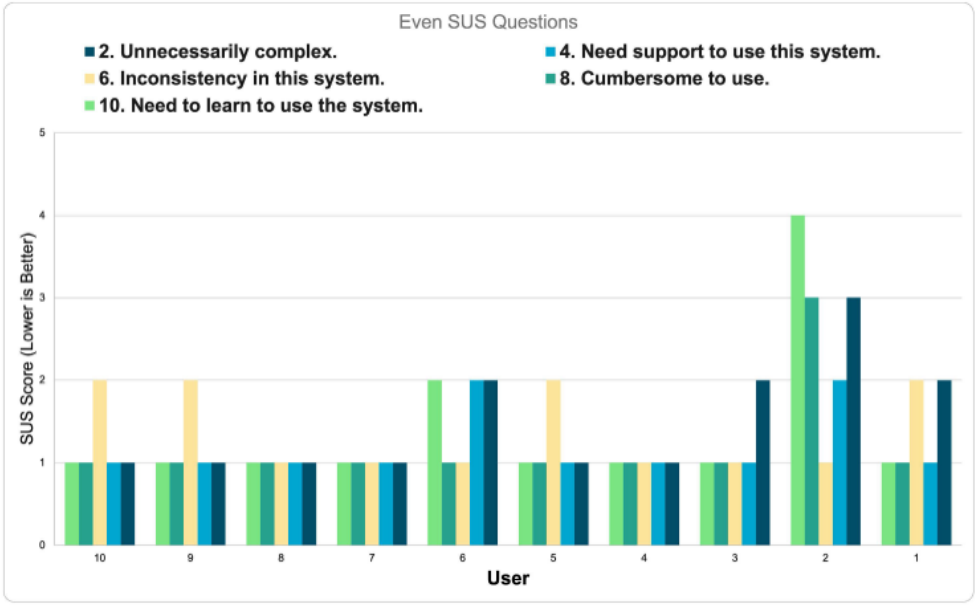

FIGURE 9. SUS questionnaire results (negative questions).

b: Q2. HOW EFFECTIVELY DID AIDA DASHBOARD SUPPORT YOU IN ASSESSING CONFERENCES?

All the researchers stated that the AIDA dashboard had an extremely positive effect on their use case. Two of them commented that it is an effective tool to check how their organizations performed; one user added that it could be very useful for early career researchers who are not familiar yet with their fields. Others appreciated that it provides a data-driven approach to evaluate conferences in alternative to standard rankings.

c: Q3. WHAT ARE THE MAIN STRENGTHS OF AIDA DASHBOARD?

Seven researchers pointed out that the main strengths of AIDA Dashboard are its simplicity and usability. One researcher appreciated the granularity of the information displayed, and another stated that it was ''easier and smoother than the competitors''.

d: Q4. WHAT ARE THE MAIN WEAKNESSES OF AIDA DASHBOARD?

Two researchers mentioned the lack of customization in the overview and in the related conferences tab. In particular, they would prefer to be able to rank the conferences according to arbitrary topics of interest. Two users criticised the inability to obtain further information about items such as authors, topics, or organizations via relevant pages or external links. One researcher suggested as main limitation the fact that the AIDA Dashboard is restricted to the Computer Science domain. One claimed that the user interface was sometimes too complex, making it hard to find all the functionalities. Another reported some disambiguation issues, in particular regarding authors with similar names. Finally, two researchers did not report any particular weakness.

e: Q5. CAN YOU THINK OF ANY ADDITIONAL FEATURES TO BE INCLUDED IN AIDA DASHBOARD?

The suggested features were: 1) the ability of bookmarking some conferences as favourite (two researchers), 2) the ability of ranking topics alphabetically (two researchers), 3) an even more granular representation of the topics (two researchers), 4) an info page explaining in details all the analytics (two researchers), 5) the ability to search all conferences in a focus areas, 6) a new rank system for conferences based on the dashboard data, and 7) the ability to generate a Map View about specific time periods.

f: Q6. HOW COMPREHENSIVE/ACCURATE DO YOU CONSIDER THE LIST OF FOCUS AREAS ASSOCIATED WITH THE CONFERENCES IN AIDA DASHBOARD?

Eight researchers found the list of focus areas very accurate. Two mentioned that the areas may have been too high-level and suggested to add the ability to allow the users to define new focus areas. Other two users believed that the inspected conference was missing a focus area.

g: Q7. HOW COMPREHENSIVE/ACCURATE DO YOU CONSIDER THE CONFERENCE COMPARISON FOR FOCUS AREAS IN AIDA DASHBOARD?

Eight researchers found the comparison very accurate and comprehensive. Two out of ten researchers pointed out some missing conferences in specific fields.

5) BEST FUNCTIONALITIES

We asked the researchers to list at least three of the most useful sections of the AIDA Dashboard. Table 4 reports the user preferences. The *Related Conference* tab, that allows users to compare conferences within a focus area, obtained nine preferences out of ten. This highlights how comparing conferences is a critical task that was not well supported by previous solutions. The analytics about citations, authors,

TABLE 4. Most useful sections for researchers.

Section	*Preferences*
Related Conferences	9
Citation Analysis	7
Authors	7
Organizations	6
Topics (Fingerprint Topics)	4
Industry (Publication Types)	4
Topics (Main Topics)	3
Industry (Industrial Sectors)	3
Countries (Map View)	2
Countries (Normal Views)	1

and organizations also obtained the majority of preferences. Four users mentioned the novel analytics about topics and industrial sectors.

VI. SUSTAINABILITY PLAN

Our goal is to keep the AIDA Dashboard up and running for the foreseeable future. We plan to update both its data and its functionalities with the support and feedback of Springer Nature and the scientific community. In particular, we plan to enhance the AIDA Dashboard by allowing the users to analyse also journals, authors, and organizations.

As first step, we are focusing on evolving our data pipeline. Indeed, at the time of writing this manuscript, Microsoft decided to decommission[46] the MAG project after 2021. To overcome this problem we devised a new strategy. We will obtain the research paper metadata from Dimensions 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 plan to leverage the conference representation of DBLP [16], which is a bibliographic database of Computer Science conferences, workshops, and journals. The current version includes 5,438 conferences.

We plan to integrate Dimensions and DBLP using the paper DOIs. For the few conferences and workshops that do not assign DOIs to articles (e.g., NeurIPS, INTERSPEECH), we will map the 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 resulting alignments.

We are currently working on this plan in collaboration with Springer Nature data science team and soon we will switch to this new solution.

VII. CONCLUSION AND FUTURE WORK

In this paper we proposed the AIDA Dashboard, a web application developed in collaboration with Springer Nature for analyzing and comparing scientific conferences which combines machine learning solutions, Semantic Web technologies, and visual analytics. It has been built on top of the Academia/Industry Dynamics Knowledge Graph [5], [52], a knowledge base that integrates information from Microsoft Academic Graph, Dimensions, DBpedia, the Computer Science Ontology, the Industrial Sectors Ontology, and the Global Research Identifier Database.

The AIDA Dashboard obtained excellent results both regarding the quality of the analytics and the usability of the interface. In particular, the qualitative evaluation showed that the automatic classification of focus areas and industrial sectors yielded an average F1 of respectively 91.8% and 87.5%. In both cases the performances of our approach are consistent with the ones of human experts according to the statistical tests. The accuracy of the background data is also reflected by the quality of the analytics that were scored 4.4/5 by the researchers in the user study. In terms of usability, the SUS questionnaire yielded a first-tier usability score (87.5/100, A+). In particular, the researchers found the AIDA Dashboard very easy to use and stated that they would be happy to use it regularly. The user study also highlighted that the most useful functionality according to the users is the comparison of conferences according to their focus areas, which is in line with our preliminary analysis that led to the development of the conference dashboard.

The evaluation highlighted also some limitations of the current prototype that we plan to address in future versions. A first concern is that some analytics (e.g., fingerprint topics) are not well explained in the dashboard. We are thus introducing additional tooltips to clarify them further. Another issue arising from the user study is that it is not possible to use topics as entry point by directly searching for all conferences addressing a focus area. We thus plan to modify the initial page and allow users to search and browse conferences according to their main topics.

Furthermore, the AIDA Dashboard is currently limited to conferences in the domain of Computer Science. We plan to improve its coverage by gradually introducing both other scholarly items, starting with journal and authors, and other academic fields. To this end, we are working on new version of our classifiers able to integrate taxonomies of research areas from other domains. Finally, at the time of writing this paper, Microsoft decided to decommission MAG. Therefore, we are now working on replacing MAG with Dimensions and DBLP, in order to produce new versions of AIDA and the AIDA Dashboard that can be sustainable and easily updated in the following years.

The AIDA dashboard, now publicly released, will benefit researchers, editors, and funding agencies by allowing them to perform granular comparisons of the conferences in a given field, analyse their research topics over the years, and assess the role of commercial organisations and industrial sectors. We plan to enhance it further in the following months by including new types of entities to analyse (e.g., journals) and developing new functionalities (e.g., prediction of topic trends).

[46]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/

REFERENCES

[1] M. Franceschet, "The role of conference publications in CS," *Commun. ACM*, vol. 53, no. 12, pp. 129–132, 2010.

[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[3] A. A. Salatino, T. Thanapalasingam, A. Mannocci, F. Osborne, and E. Motta, "The computer science ontology: A large-scale taxonomy of research areas," in *Proc. Int. Semantic Web Conf.* Cham, Switzerland: Springer, 2018, pp. 187–205.

[4] B. Rahdari, P. Brusilovsky, and A. J. Sabet, "Connecting students with research advisors through user-controlled recommendation," in *Proc. 15th ACM Conf. Rec. Syst.* New York, NY, USA: Association for Computing Machinery, 2021, pp. 745–748.

[5] S. Angioni, A. Salatino, F. Osborne, D. R. Recupero, and E. Motta, "AIDA: A knowledge graph about research dynamics in academia and industry," *Quant. Sci. Stud.*, vol. 2, no. 4, pp. 1356–1398, 2022, doi: 10.1162/qss_a_00162.

[6] J. Brooke, "Sus: A 'quick and dirty' usability scale," *Usability Eval. Ind.*, vol. 189, no. 3, pp. 1–8, 1996.

[7] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, and A. Kanakia, "Microsoft academic graph: When experts are not enough," *Quant. Sci. Stud.*, vol. 1, no. 1, pp. 396–413, Feb. 2020.

[8] Y. Zhang, F. Zhang, P. Yao, and J. Tang, "Name disambiguation in AMiner: Clustering, maintenance, and human in the loop," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1002–1011.

[9] A. G. Nuzzolese, A. L. Gentile, V. Presutti, and A. Gangemi, "Conference linked data: The scholarlydata project," in *Proc. Int. Semantic Web Conf.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 9982, 2016, pp. 150–158.

[10] M. Fenner and A. Aryani, "Introducing the PID Graph," 2019, doi: 10.5438/JWVF-8A66.

[11] M. Y. Jaradeh, A. Oelen, K. E. Farfar, M. Prinz, J. D'Souza, G. Kismihók, M. Stocker, and S. Auer, "Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge," in *Proc. 10th Int. Conf. Knowl. Capture (K-CAP)*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 243–246.

[12] S. Peroni and D. Shotton, "OpenCitations, an infrastructure organization for open scholarship," *Quant. Sci. Stud.*, vol. 1, no. 1, pp. 428–444, Feb. 2020.

[13] P. Manghi, C. Atzori, A. Bardi, M. Baglioni, J. Schirrwagen, H. Dimitropoulos, S. La Bruzzo, I. Foufoulas, A. Mannocci, M. Horst, A. Czerniak, K. Kiatropoulou, A. Kokogiannaki, M. De Bonis, M. Artini, E. Ottonello, A. Lempesis, A. Ioannidis, N. Manola, and P. Principe, "OpenAIRE research graph dump (4.0) [data set]," Zenodo, 2021, doi: 10.5281/zenodo.5801283.

[14] M. Visser, N. J. van Eck, and L. Waltman, "Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic," *Quant. Sci. Stud.*, vol. 2, no. 1, pp. 20–41, 2021, doi: 10.1162/qss_a_00112.

[15] W. Ammar *et al.*, "Construction of the literature graph in semantic scholar," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 84–91.

[16] M. Ley, "DBLP: Some lessons learned," *Proc. VLDB Endowment*, vol. 2, no. 2, pp. 1493–1500, Aug. 2009.

[17] P. Knoth and Z. Zdrahal, "CORE: Connecting repositories in the open access domain," in *Proc. CERN Workshop Innov. Scholarly Commun. (OAI7)*, Geneva, Switzerland, Jun. 2011.

[18] S. La Bruzzo, P. Manghi, and A. Mannocci, "Openaire's doiboost—Boosting crossref for research," in *Digital Libraries: Supporting Open Science*, P. Manghi, L. Candela, and G. Silvello, Eds. Cham, Switzerland: Springer, 2019, pp. 133–143.

[19] D. Shotton, "Semantic publishing: The coming revolution in scientific journal publishing," *Learned Publishing*, vol. 22, no. 2, pp. 85–94, Apr. 2009.

[20] A. G. Nuzzolese, A. L. Gentile, V. Presutti, and A. Gangemi, "Semantic web conference ontology—A refactoring solution," in *Proc. Eur. Semantic Web Conf.* Cham, Switzerland: Springer, 2016, pp. 84–87.

[21] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2RDF: Towards a mashup to build bioinformatics knowledge systems," *J. Biomed. Informat.*, vol. 41, no. 5, pp. 706–716, Oct. 2008.

[22] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, J. Bhagat, K. Belhajjame, F. Bacall, A. Hardisty, A. N. de la Hidalga, M. P. B. Vargas, S. Sufi, and C. Goble, "The Taverna workflow suite: Designing and executing workflows of web services on the desktop, web or in the cloud," *Nucleic Acids Res.*, vol. 41, no. W1, pp. W557–W561, Jul. 2013.

[23] F. Osborne and E. Motta, "Klink-2: Integrating multiple web sources to generate semantic topic networks," in *Proc. ISWC*. Cham, Switzerland: Springer, 2015, pp. 408–424.

[24] D. Buscaldi, D. Dessì, E. Motta, F. Osborne, and D. R. Recupero, "Mining scholarly data for fine-grained knowledge graph construction," in *Proc. CEUR Workshop*, vol. 2377, 2019, pp. 21–30.

[25] P. Groth, A. Gibson, and J. Velterop, "The anatomy of a nanopublication," *Inf. Services Use*, vol. 30, nos. 1–2, pp. 51–56, Sep. 2010.

[26] T. Kuhn, C. Chichester, M. Krauthammer, N. Queralt-Rosinach, R. Verborgh, G. Giannakopoulos, A.-C. N. Ngomo, R. Viglianti, and M. Dumontier, "Decentralized provenance-aware publishing with nanopublications," *PeerJ Comput. Sci.*, vol. 2, p. e78, Aug. 2016.

[27] J. Schneider, P. Ciccarese, T. Clark, and R. D. Boyce, "Using the micropublications ontology and the open annotation data model to represent evidence within a drug-drug interaction knowledge base," in *Proc. Workshop Linked Sci. Making Sense Out Data (LISC) ISWC*, Riva del Garda, Italy, Oct. 2014. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01076282

[28] S. Peroni and D. Shotton, "The spar ontologies," in *Proc. Int. Semantic Web Conf.* Cham, Switzerland: Springer, 2018, pp. 119–136.

[29] A. A. Salatino, T. Thanapalasingam, A. Mannocci, F. Osborne, and E. Motta, "The computer science ontology: A large-scale taxonomy of research areas," in *The Semantic Web—ISWC*, D. Vrandečić, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee, and E. Simperl, Eds. Cham, Switzerland: Springer, 2018, pp. 187–205.

[30] S. Fathalla, S. Auer, and C. Lange, "Towards the semantic formalization of science," in *Proc. 35th Annu. ACM Symp. Appl. Comput.*, Mar. 2020, pp. 2057–2059.

[31] F. A. Nielsen, D. Mietchen, and E. Willighagen, "Scholia, scientometrics and wikidata," in *The Semantic Web: ESWC 2017 Satellite Events*, E. Blomqvist, K. Hose, H. Paulheim, A. Ławrynowicz, F. Ciravegna, and O. Hartig, Eds. Cham, Switzerland: Springer, 2017, pp. 237–259.

[32] O. A. Jefferson, D. Koellhofer, B. Warren, and R. Jefferson, "The lens MetaRecord and LensID: An open identifier system for aggregated metadata and versioning of knowledge artefacts," Nov. 2019, doi: 10.31229/osf.io/t56yh.

[33] N. J. van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," *Scientometrics*, vol. 84, no. 2, pp. 523–538, 2010.

[34] O. F. Guilarte, S. D. J. Barbosa, and S. Pesco, "RelPath: An interactive tool to visualize branches of studies and quantify the expertise of authors by citation paths," *Scientometrics*, vol. 126, no. 6, pp. 4871–4897, Jun. 2021.

[35] M. D. L. Tosi and J. C. dos Reis, "SciKGraph: A knowledge graph approach to structure a scientific field," *J. Informetrics*, vol. 15, no. 1, Feb. 2021, Art. no. 101109.

[36] S. Khalid, S. Wu, A. Wahid, A. Alam, and I. Ullah, "An effective scholarly search by combining inverted indices and structured search with citation networks analysis," *IEEE Access*, vol. 9, pp. 120210–120226, 2021.

[37] Z. Ali, I. Ullah, A. Khan, A. Ullah Jan, and K. Muhammad, "An overview and evaluation of citation recommendation models," *Scientometrics*, vol. 126, no. 5, pp. 4083–4119, May 2021.

[38] A. Salatino, F. Osborne, and E. Motta, "ResearchFlow: Understanding the knowledge flow between academia and industry," in *Knowledge Engineering and Knowledge Management*. Cham, Switzerland: Springer, 2020.

[39] A. A. Salatino, F. Osborne, and E. Motta, "How are topics born? Understanding the research dynamics preceding the emergence of new areas," *PeerJ Comput. Sci.*, vol. 3, p. e119, Jun. 2017.

[40] F. Osborne, E. Motta, and P. Mulholland, "Exploring scholarly data with rexplore," in *The Semantic Web—ISWC*, H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. Noy, C. Welty, and K. Janowicz, Eds. Berlin, Germany: Springer, 2013, pp. 460–477.

[41] F. Löffler, V. Wesp, S. Babalou, P. Kahn, R. Lachmann, B. Sateli, R. Witte, and B. König-Ries, "Scholarlensviz: A visualization framework for transparency in semantic user profiles," in *Proc. ISWC Demos Ind. Tracks, From Novel Ideas Ind. Practice Co-Located With 19th Int. SemanticWeb Conf. (ISWC)*, K. Taylor, R. Gonçalves, F. Lecue, and J. Yan, Eds., Nov. 2020, pp. 20–25.

[42] X. Zhang, S. Chandrasegaran, and K.-L. Ma, "ConceptScope: Organizing and visualizing knowledge in documents based on domain ontology," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–13.
[43] T. Vergoulis, S. Chatzopoulos, T. Dalamagas, and C. Tryfonopoulos, "VeTo: Expert set expansion in academia," in *Digital Libraries for Open Knowledge*, M. Hall, T. Merčun, T. Risse, and F. Duchateau, Eds. Cham, Switzerland: Springer, 2020, pp. 48–61.
[44] M. V. M. Borges and J. C. dos Reis, "Semantic-enhanced recommendation of video lectures," in *Proc. IEEE 19th Int. Conf. Adv. Learn. Technol. (ICALT)*, Jul. 2019, pp. 42–46.
[45] S. Chatzopoulos, T. Vergoulis, I. Kanellos, T. Dalamagas, and C. Tryfonopoulos, "ArtSim: Improved estimation of current impact for recent articles," in *Proc. ADBIS, TPDL EDA Common Workshops Doctoral Consortium*. Cham, Switzerland: Springer, 2020, pp. 323–334.
[46] A. A. Salatino, F. Osborne, A. Birukou, and E. Motta, "Improving editorial workflow and metadata quality at springer nature," in *The Semantic Web—ISWC*. Cham, Switzerland: Springer, 2019, pp. 507–525.
[47] T. Thanapalasingam, F. Osborne, A. Birukou, and E. Motta, "Ontology-based recommendation of editorial products," in *The Semantic Web—ISWC*, D. Vrandečić, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee, and E. Simperl, Eds. Cham, Switzerland: Springer, 2018, pp. 341–358.
[48] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. Hsu, and K. Wang, "An overview of Microsoft academic service (MAS) and applications," in *Proc. 24th Int. Conf. World Wide Web*, May 2015, pp. 243–246.
[49] A. A. Salatino, F. Osborne, T. Thanapalasingam, and E. Motta, "The CSO classifier: Ontology-driven detection of research topics in scholarly articles," in *Digital Libraries for Open Knowledge*, A. Doucet, A. Isaac, K. Golub, T. Aalberg, and A. Jatowt, Eds. Cham, Switzerland: Springer, 2019, pp. 296–311.
[50] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2. Red Hook, NY, USA: Curran Associates, 2013, pp. 3111–3119.
[51] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a 'kneedle' in a haystack: Detecting knee points in system behavior," in *Proc. 31st Int. Conf. Distrib. Comput. Syst. Workshops*, Jun. 2011, pp. 166–171.
[52] S. Angioni, A. Salatino, F. Osborne, D. R. Recupero, and E. Motta, "Integrating knowledge graphs for analysing academia and industry dynamics," in *Proc. ADBIS, TPDL EDA Common Workshops Doctoral Consortium*. Cham, Switzerland: Springer, 2020, pp. 219–225.

SIMONE ANGIONI received the B.S. and M.S. degrees in computer science from the University of Cagliari, Italy, where he is currently pursuing the Ph.D. degree. He is the Main Developer of the Academia/Industry DynAmics (AIDA) Knowledge Graph, an innovative resource for studying the relationship between academia and industry. His research interests include science of science, scientometrics, information extraction, semantic web, and robotics.

ANGELO SALATINO received the Ph.D. degree in early detection of research trends. He is currently a Research Associate at the Intelligence Systems and Data Science (ISDS) Group, Knowledge Media Institute (KMi), The Open University. In particular, his project aimed at identifying the emergence of new research topics at their embryonic stage. His research interests include semantic web, network science, and knowledge discovery technologies, with focus on the structures and evolution of science.

FRANCESCO OSBORNE is currently a Research Fellow at the Knowledge Media Institute, The Open University, U.K., where he leads the Scholarly Data Mining Team. He is also an Assistant Professor at the University of Milano Bicocca. He collaborates with major publishers, universities, and companies in the space of innovation for producing a variety of innovative services for supporting researchers, editors, and research polities makers. He released many well-adopted resources, such as the computer science ontology and the artificial intelligence knowledge graph. His research interests include artificial intelligence, information extraction, knowledge graphs, science of science, and semantic web. He has authored more than 90 peer-reviewed publications in top journals and conferences of these fields.

DIEGO REFORGIATO RECUPERO received the Ph.D. degree in computer science from the University of Naples Federico II, Italy, in 2004. From 2005 to 2008, he was a Postdoctoral Researcher at the University of Maryland, College Park, MD, USA. He has been an Associate Professor at the Department of Mathematics and Computer Science, University of Cagliari, Italy, since December 2015. He co-founded six companies within the ICT sector. He is actively involved in European projects and research (with one of his companies he won more than 40 FP7 and H2020 projects). His current research interests include sentiment analysis, semantic web, natural language processing, human–robot interaction, financial technology, and smart grid. He is the author of more than 170 conference and journal papers in these research fields, with more than 2000 citations. He won different awards in his career, such as the Marie Curie International Reintegration Grant, the Marie Curie Innovative Training Network, the Best Researcher Award from the University of Catania, the Computer World Horizon Award, the Telecom Working Capital, Startup Weekend, and the Best Paper Award.

ENRICO MOTTA received the Laurea degree in computer science from the University of Pisa, Italy, and the Ph.D. degree in artificial intelligence from The Open University, U.K. He was the Former Director of the Knowledge Media Institute (KMi), The Open University, from 2000 to 2007. He is currently a Professor in knowledge technologies at the Knowledge Media Institute (KMi), The Open University. Over the years, he has led KMi's contribution to numerous high-profile projects, receiving over £10.4 million in external funding from a variety of institutional funding bodies and commercial organizations, since 2000. His research interests include the intersection of large-scale data integration and modeling, semantic and language technologies, intelligent systems, and human–computer interaction.