EVENTS: A Dataset on the History of Top-Prestigious Events in Five Computer Science Communities

Said Fathalla[1,2] and Christoph Lange[1,3]

[1] Smart Data Analytics (SDA), University of Bonn, Germany
{fathalla,langec}@cs.uni-bonn.de
[2] Faculty of Science, University of Alexandria, Egypt
[3] Fraunhofer IAIS, Germany

Abstract Information emanating from scientific events, journal, organizations, institutions as well as scholars become increasingly available online. Therefore, there is a great demand to assess, analyse and organize this huge amount of data produced every day, or even every hour. In this paper, we present a dataset (EVENTS) of scientific events, containing historical data about the publications, submissions, start date, end date, location and homepage for 25 top-prestigious event series (718 editions in total) in five computer science communities. The dataset is publicly available online in three different formats (i.e., CSV, XML, and RDF). It is of primary interest to the steering committees or program chairs of the events to assess the progress of their event over time and compare it to competing events in the same field, and to potential authors looking for events to publish their work. In addition, we shed light on these events by analyzing their metadata over the last 50 years. Our transferable analysis is based on exploratory data analysis.

Keywords: Scientific Events Dataset, Scholarly Communication, Digitization, Metadata Analysis

1 Introduction

Digitization is of crucial importance to all areas of scholarly communication. Therefore, over the last two decades, many organizations and institutes have begun to organize and establish new scientific events. This paper discusses some facts and figures representing 50 years[4] of history of computer science events, where conferences, symposia, and workshops are of paramount importance and a major means of scholarly communication. A key question is: How does digitization affect scholarly communication in computer science? In particular, we address the following questions:

a) What is the trend of submissions and acceptance rates?

[4] The oldest data points.

b) How did the number of publications change?
c) Is there an augmentation of publications of a computer science sub-community?
d) Has the geographical distribution of events changed across various regions of the world?
e) Which events are more geographically diverse than others?

We target some of these questions by analyzing comprehensive scholarly communication metadata from computer science events in the last 50 years. Our analysis methodology is based on exploratory data analysis, which aims at analyzing data to explore the main characteristics, oftentimes with visual methods. We analyze the key characteristics of scientific events over time, including their CORE[5], Qualis (Q)[6] and GII rankings[7], geographic distribution, average acceptance rate, time distribution over the year, submissions and publications. We selected five top-prestigious events in five CS communities derived from analyzing the topics covered by each event series, then mapping the event series to the ACM Computing Classification System (CCS)[8]: Information systems (IS), Security and privacy (SEC), Artificial intelligence (AI), Computer systems organization (CSO) and Software and its engineering (SE). Events will only be referred to using their acronym. We believe that the EVENTS dataset will have a great impact on scholarly communication community, particularly for the following stakeholders(cf. [6]): *a*) *event organizers*: to trace their events' progress/impact, *b*) *authors*: identify prestigious events to submit their research results to, *c*) *proceedings publishers*: to know the impact of the events whose proceedings they are publishing.

This article is organized as follows: Section 2 gives an overview of related work. Section 3 presents the main characteristics of the dataset. Section 4 explains the curation process of creating and evolving the dataset. Section 5 discusses the results of our analysis of the dataset. Section 6 concludes and outlines our future work.

2 Related Work

In our recent review of the literature [1, 4, 5, 9, 8] found that most studies tended to focus on grabbing information about scholarly communication from bibliographic metadata. Ameloot et al. presented a comprehensive analysis of the Principles of Database Systems (PODS) conference series including word clouds of most PODS researchers and newcomers, longest streaks and locations of PODS in the period 2002–2011 [2]. Similarly, Aumüller and Rahm [3] analyzed affiliations of database publications using author information from DBLP. Fathalla et al. [7] provided an analysis of 40 computer science conference series in terms of continuity, time and geographic distribution, submissions and publications. Barbosa et al. [4] analysed the metadata of 340 full papers published

[5] `http://www.core.edu.au/`
[6] `http://qualis.ic.ufmt.br/`
[7] `http://valutazione.unibas.it/gii-grin-scie-rating/`
[8] `https://dl.acm.org/ccs/ccs.cfm`

in 14 editions of the Brazilian Symposium on Human Factors in Computing Systems (IHC). Vasilescu et al. [10] presented a dataset of eleven software engineering conferences, containing historical data about publications and program committees in the period 1994–2012. Agarwal et al. [1] presented a bibliometric analysis of the metadata of seven ACM conferences covering different CS fields such as information management, data mining, digital libraries and information retrieval.

3 Characteristics of the EVENTS Dataset

EVENTS dataset covers historical information about 25 top-prestigious events of the last five decades, including (where available) an event's full title, acronym, start date, end date, number of submissions, number of accepted papers, city, state, country, event type, field and homepage. These global indicators have been used to spot and interpret peculiarities on the temporal and geographical evolution of event series. There are two types of events: conferences and symposia[9]. Table 1 provides high-level statistics for the 25 event series in the five CS communities of IS, SEC, AI, CSO, and SE. Entries refers to all available attributes of all events.

Table 1: EVENTS dataset: high-level statistics.

Metrics	**value**	**Metrics**	**value**
series	25	event types	2
editions	718	communities	5
entries	9,460	duration (years)	50
attributes	15	available formats	3

Use Cases. Using this dataset, event organizers and chairs will be able to assess their selection process, e.g., to keep, if desired, the acceptance rate stable even when the submissions increase, to make sure the event is held around the same time each year, and to compare against other competing events. Furthermore, we believe this dataset will assist researchers who want to submit a paper to be able to decide to which events they could submit their work, e.g., answering questions such as "which events have a high impact in a particular CS field?". Moreover, when a specific conference is held each year, it helps them to prepare their research within the conference's usual timeline. section 5 presents a part of the analysis that could be performed by using the EVENTS dataset.

Extensibility. EVENTS can be extended in three dimensions to meet future requirements by 1) adding more events in each community, 2) adding events in other communities and 3) adding more attributes such as hosting university or

[9] It would be correct to label a symposium as a small scale conference as the number of participants is smaller.

organization, sponsors, and event steering committees or program committee chairs.

Availability. EVENTS is published at `https://saidfathalla.github.io/EVENTS-Dataset/EVENTS.html`. It is subject to the Creative Commons Attribution license, as documented at `https://saidfathalla.github.io/EVENTS-Dataset/EVENTS_Licence.html`. The RDF version has been validated using W3C Validation Service[10]. The following listing shows the information about the AAAI conference of 2017 in RDF. We defined new vocabularies in the OpenResearch namespace[11].

```
<rdf:RDF xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
xmlns="http://example.org/data/PERCOM.csv#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:or="http://openresearch.org/vocab/">

<rdf:Description rdf:about="http://openresearch.org/vocab/AAAI">
  <Event_Title>23rd Annual National Conference on Artificial Intelligence</Title>
  <Event_Acronym>AAAI</Event_Acronym>
  <Event_Field>Artificial Intelligence</Event_Field>
  <Event_Homepage>aaai.org/Conferences/AAAI/aaai17.php</Event_Homepage>
  <Event_Series>AAAI2017</Event_Series>
  <Event_Year rdf:datatype="&xsd;date">2017</Event_Year>
  <Event_Start_date rdf:datatype="&xsd;dateTime">2017-02-04</Event_Start_date>
  <Event_End_date rdf:datatype="&xsd;dateTime">2017-02-09</Event_End_date>
  <Event_City>San Francisco</Event_City>
  <Event_State>California</Event_State>
  <Event_Country>USA</Event_Country>
  <Event_Submission_Rate rdf:datatype="&xsd;decimal">24.2\%</Event_Submission_Rate>
  <Event_Submitted_papers rdf:datatype="&xsd;integer">2590</Event_Submitted_papers>
  <Event_Accepted_papers rdf:datatype="&xsd;integer">638</Event_Accepted_papers>
  <Event_Type>Conference</Event_Type>
</rdf:Description>
```

4 Data Curation

While we collected the data for the dataset, we faced several technical problems, such as having to eliminate irrelevant and redundant data, to unify event name, to complete missing data, and to correct incorrect data. Therefore, a data curation process is required. The EVENTS dataset is being maintained over time according to the curation process described later in this section.

4.1 Data Acquisition

After identifying top events, metadata (raw data) of these events is collected either from structured or unstructured data sources. The metadata of selected events has been manually collected from various sources such as IEEE Xplore Digital Library[12], ACM Digital Libraries[13], DBLP, OpenResearch.org and events websites. The selection is based on several criteria such as CORE ranking, Qualis ranking, GII ranking and Google h-index (the largest number h such that h articles published in the last 5 complete years have at least h citations each).

[10] `https://www.w3.org/RDF/Validator/`

[11] `or:http://openresearch.org/`

[12] `http://ieeexplore.ieee.org`

[13] `https://dl.acm.org/`

4.2 Data preprocessing

The main objective of the data preprocessing phase is to fill in missing data, to identify and correct incorrect data, to eliminate irrelevant data and to resolve inconsistencies. In order to prepare the raw data for analysis, we carried out three preprocessing processes: *data integration*, *data cleansing*, *data transformation* and *Event name unification.*.

Data integration. This process involves combining data from multiple sources into meaningful and valuable information. In addition, this process also involves eliminating redundant data which occur during the integration process.

Data cleansing. This process involves detecting and correcting incorrect or inaccurate records. For instance, we found several websites providing incorrect information about events' submissions and accepted papers. We double checked this information against the events' official websites or proceedings published in digital libraries.

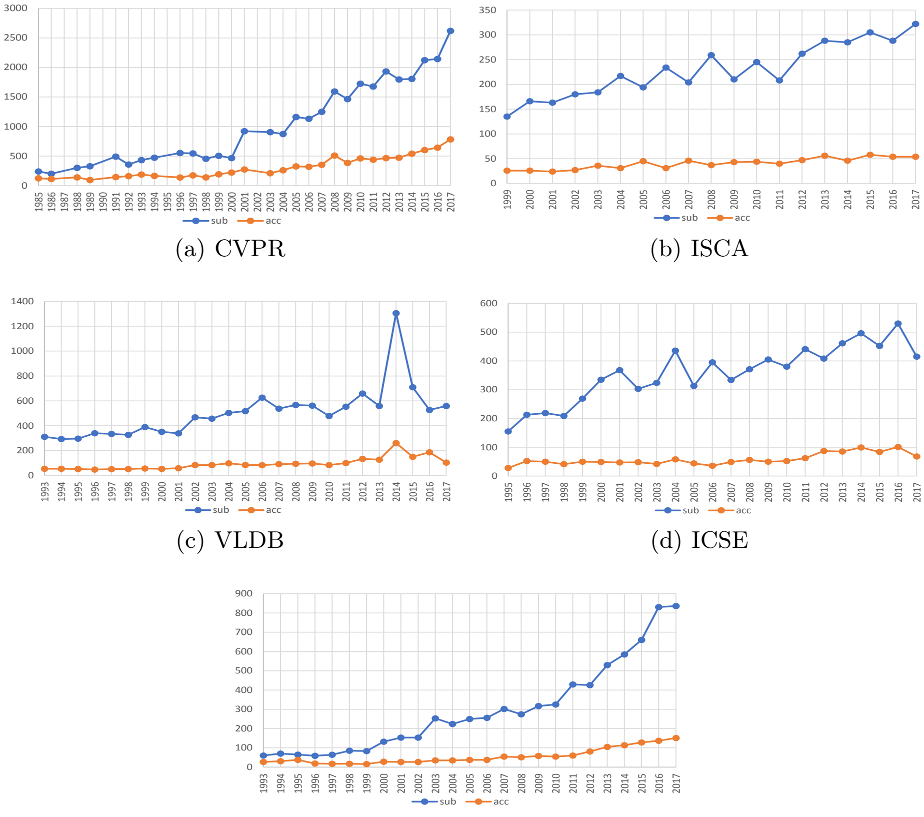
Data transformation. This process involves converting cleaned data values from unstructured formats into a structured one. For instance, data collected from events websites as text (i.e. unstructured format) is manually transformed to CSV (i.e. structured format) and consequently to XML and RDF.

Event name unification. This process involves integrating all editions of an event series, which had changed its name since the beginning under its most recent name because it is important for the researchers to know the recent name rather than the old name. However, the old name remains important for a researcher who wants to get an overview of the history of an event. For example, PLDI is the unified name of the *Conference on Programming Language Design and Implementation*, which was named *Symposium on Compiler Construction* in the period 1979–1986, *Symposium on Interpreters and Interpretive Techniques* in 1987 and finally it assumed its recent name in the period 1989–2018, i.e., for 30 years. With the completion of these steps, we are now ready to perform our exploratory data analysis.

5 Data Analysis And Results

Over the last 50 years, we have analyzed metadata of CS events in the EVENTS dataset including the h5-index, the average acceptance rate, the number of editions of each event, the country that hosted most editions of the event, the month in which the event is usually held each year, the year of the first edition, and the publisher of the proceedings.

Submissions and publications. Figure 1 presents accepted and submitted papers measures for the top events, i.e. high-ranked events in terms of h5-index and events ranking services, in the five CS communities from 1985 to 2017. For the CVPR conference, the numbers of submitted and accepted papers were very close in the first edition in 1985, and the gap between them began to slightly increase until 2000, then it increased noticeably until the end of the time span, i.e., 2017. The gap between submission and accepted papers refers to how far the number of submissions from the accepted papers.



(a) CVPR

(b) ISCA

(c) VLDB

(d) ICSE

(e) CCS

Figure 1: Variation of the number of submitted and accepted papers of the top event in each CS community.

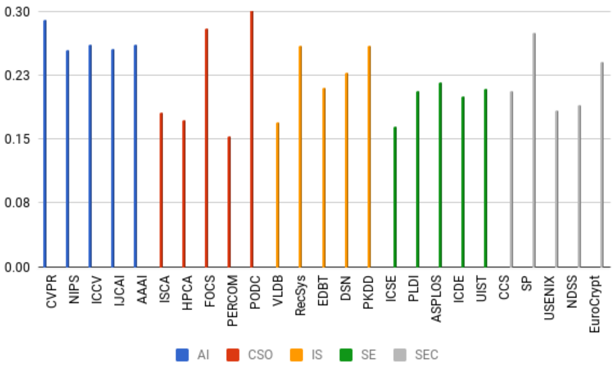


Figure 2: Average acceptance rate of all events

However, the gap between the number of submitted papers and accepted papers in VLDB remained the same during the whole time span. Overall, we can see a clear upward trend in the number of submitted and accepted papers during the whole time span. The reason is that digitization makes more research papers available to the whole community and submitting papers and even contacting papers' reviewers has become much easier and efficient.

Time distribution. We observed that the organizers of the prestigious events always try to keep holding their events around the same month each year, which helps researchers who want to submit their work to expect the date of the next edition of an event. Namely, PLDI has been held 30 times (out of 36) in June and SP has been held 31 times (out of 39) since 1989 in May.

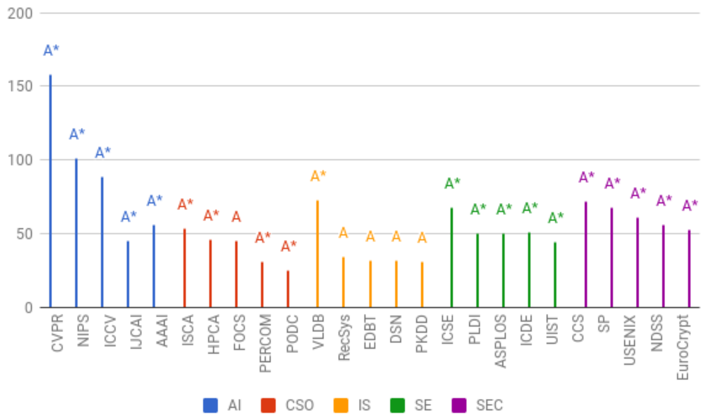


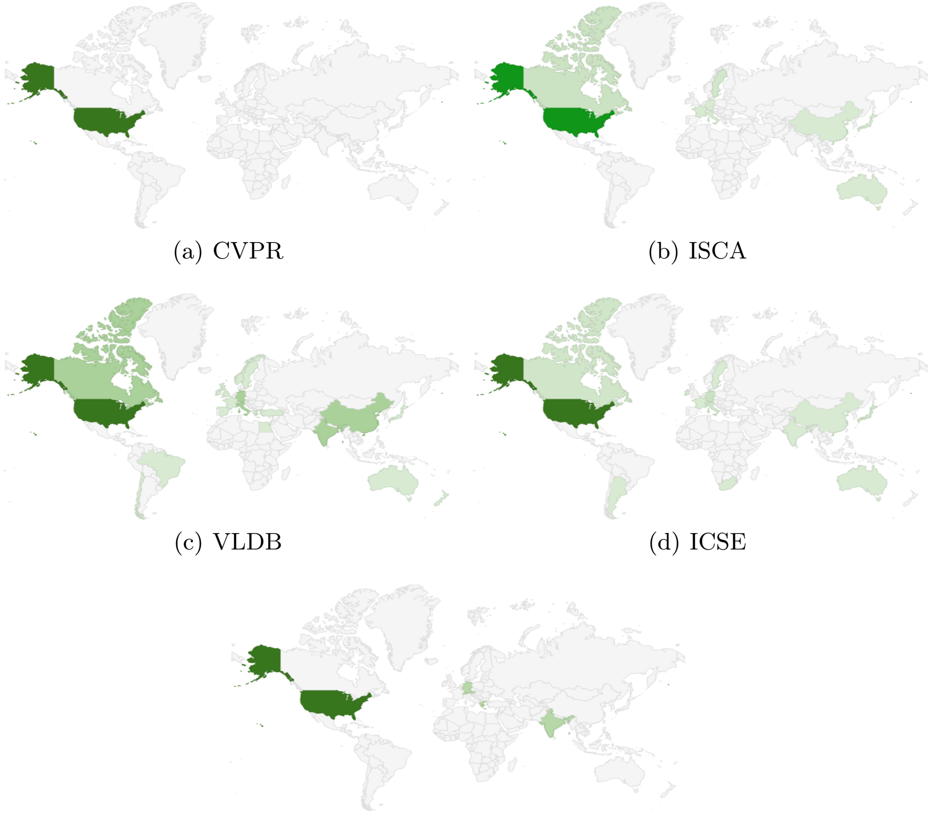
Figure 3: H5-index of all events along with CORE 2018 ranking

Acceptance rate. We analyzed the acceptance rate of the events involved in the study over the last 50 years. As shown in Figure 2, for each event, we computed the average of the acceptance rate of each event since beginning[14]. Interestingly, we found that the average acceptance rate for all events, since the first edition, falls into the range 15% to 31% in the time window of 50 years. Overall, the largest acceptance rate is the one of PODC of 31%, while PERCOM has the smallest one of 15%.

H5-index. Figure 3 presents the h5-index of all event series along with their CORE 2018 ranking. The highest h5-index is the one of CVPR of 158, while PODC has the smallest one of 25.

Geographical distribution. We analyzed the geographical distribution of each event in the dataset. The key question is which countries hosted most of

[14] these values are included into the dataset, so that others wouldn't have to recompute them.



(a) CVPR

(b) ISCA

(c) VLDB

(d) ICSE

(e) CCS

Figure 4: Geographical distribution of the top event in each CS community since 1973.

the top events in the EVENTS dataset, and how frequently a country has hosted an event during the last five decades. Figure 4 shows how frequently different countries around the world have hosted a top event in the five CS communities considered in the study. We observed that USA leads by far, having hosted most editions of CVPR, ISCA, VLDB, ICSE, and CCS. Canada comes second, hosting most editions of ISCA, VLDB, and ICSE.

Table 2 shows the scientometric profile of all events in the EVENTS dataset in the five considered CS communities ordered by descending h5-index for each community. AI community has the largest average h5-index of 89.9; SEC comes second with 62. Surprisingly, despite the Qualis ranking of RecSys as *B1*, the h5-index of RecSys is relatively high, and it is ranked as *A* by CORE and as *A-* by GII. Regarding publishers, we observed that ACM publishes most of the events proceedings, and IEEE comes next. However, we observed that some events such as NDSS and USENIX publish their proceedings on their own website. In terms of the number of editions, ISCA has the longest history with 45 editions since 1969, while RecSys is the newest one, with 12 editions since 2007. Although RecSys is a relatively new conference, it has a good reputation and it is highly-ranked in CORE, GII, and Qualis.

Table 2: Scientometric profile of all events in EVENTS dataset in five CS communities. N is the number of editions in 2018

Acronym	**Comm.**	**CORE 2018**	**GII**	**Q**	**h5**	**N**	**Avg. AR**	**Most freq. Country**	**Usual Month**	**Usual Month Freq.**	**Since**	**Publisher**
CVPR	AI	A*	A+	A1	158	28	0.33	US	Jun	26	1985	IEEE
NIPS		A*	A++	A1	101	32	0.25	US	Dec	18	1987	NIPS
ICCV		A*	A++	A1	89	17	0.26	Japan	Oct	5	1987	IEEE
IJCAI		A*	A++	A1	45	27	0.26	US	Aug	16	1969	AAAI
AAAI		A*	A++	A1	56	32	0.26	US	Jul	20	1980	AAAI
ISCA	CSO	A*	A++	A1	54	45	0.18	US	Jun	27	1973	IEEE
HPCA		A*	A+	A1	46	24	0.20	US	Feb	17	1995	ACM
FOCS		A	A++	A1	45	30	0.28	US	Oct	25	1989	IEEE
PERCOM		A*	A+	A1	31	16	0.15	US	Mar	16	2003	IEEE
PODC		A*	A+	A1	25	37	0.30	Canada	Aug	19	1982	ACM
VLDB	IS	A*	A++	A1	73	33	0.18	US	Aug	20	1985	VLDB
RecSys		A	A-	B1	34	12	0.26	US	Oct	7	2007	ACM
EDBT		A	A	A2	32	21	0.20	Italy	Mar	21	1988	OP
DSN		A	A	A1	32	19	0.23	US	Jun	18	2000	IEEE
PKDD		A	A	A2	31	22	0.25	France	Sep	19	1997	ACM
ICSE	SE	A*	A++	A1	68	24	0.17	US	May	25	1975	ACM
PLDI		A*	A++	A1	50	33	0.21	US	Jun	33	1979	ACM
ASPLOS		A*	A++	A1	50	23	0.22	US	Mar	10	1982	ACM
ICDE		A*	A+	A1	51	34	0.20	US	Feb	14	1984	IEEE
UIST		A*	A+	A1	44	31	0.21	US	Oct	18	1988	ACM
CCS	SEC	A*	A++	A1	72	25	0.22	US	Oct	12	1993	ACM
SP		A*	A++	A1	68	39	0.28	US	May	31	1980	IEEE
USENIX		A*	A-	A1	61	27	0.19	US	Aug	17	1990	USENIX
NDSS		A*	A+	A1	56	25	0.20	US	Feb	24	1993	NDSS
EuroCrypt		A*	A++	A1	53	37	0.24	France	May	23	1982	Springer

6 Conclusions and Future work

In this paper, we present a dataset (EVENTS) of metadata about conferences and symposia, containing historical data about 25 top prestigious events in five computer science communities. We presented our methodology of creating the dataset, starting from identifying prestigious events, data acquisition and preprocessing to finally publishing the dataset. To the best of our knowledge, this is the first time a dataset is published that contains metadata of top prestigious events in Information systems, Security and privacy, Artificial intelligence, Computer systems organization and Software and its engineering. This dataset is used to compare scientific events in the same community, which is useful for both events organizers and less-expertise researchers. In summary, we made the following observations:

- During data acquisition, we observed that there is not much information about events prior to 1990, in particular on the number of submissions and accepted papers,
- organizers of the prestigious events try to keep the events held around the same month each year,
- There is a clear upward trend in the number of submitted and accepted papers during the whole time span due to the digitization of scholarly communication. However, the digitization of scholarly communication also has negative impacts, most significantly the proliferation of submissions, which significantly increases the reviewing workload,
- Among all countries, USA hosted about 76% of the events in the dataset in the last five decades.

To further our research, we are planning to systematically investigate review quality, to update EVENTS to meet future requirements by adding more events in each community and more attributes such as hosting university or organization, sponsors, and event steering committees or program committee chairs. Furthermore, we plan to perform more exploratory analysis by applying more metrics such as geographical distribution and publications by continents, event continuity, event progress rate and acceptance rate stability.

Acknowledgments

This work has been supported by the DFG under grant agreement AU 340/9-1 (OSCOSS).

References

1. Agarwal, S., Mittal, N., Sureka, A. A glance at seven acm sigweb series of conferences. In: ACM SIGWEB Newsletter(Summer) (2016), p. 5.
2. Ameloot, T. J., Marx, M., Martens, W., Neven, F., Wees, J. van. 30 Years of PODS in facts and figures. In: SIGMOD Record 40(3) (2011).
3. Aumüller, D., Rahm, E. Affiliation analysis of database publications. In: SIGMOD Record 40(1) (2011).
4. Barbosa, S. D. J., Silveira, M. S., Gasparini, I. What publications metadata tell us about the evolution of a scientific community: the case of the Brazilian human–computer interaction conference series. In: Scientometrics 110(1) (2017), pp. 275–300.
5. Biryukov, M., Dong, C. Analysis of computer science communities based on DBLP. In: *TPDL*. Springer. 2010.
6. Bryl, V., Birukou, A., Eckert, K., Kessler, M. What's in the proceedings? Combining publisher's and researcher's perspectives. In: *SePublica*. CEUR-WS.org 1155. 2014.
7. Fathalla, S., Vahdati, S., Lange, C., Auer, S. Analysing Scholarly Communication Metadata of Computer Science Events. In: *International Conference on Theory and Practice of Digital Libraries*. Springer. 2017, pp. 342–354.

8. Fathalla, S., Vahdati, S., Auer, S., Lange, C. Towards a Knowledge Graph Representing Research Findings by Semantifying Survey Articles. In: *International Conference on Theory and Practice of Digital Libraries*. Springer. 2017, pp. 315–327.
9. Hiemstra, D., Hauff, C., De Jong, F., Kraaij, W. SIGIR's 30th anniversary: an analysis of trends in IR research and the topology of its community. In: *ACM SIGIR Forum*. Vol. 41. 2. ACM. 2007, pp. 18–24.
10. Vasilescu, B., Serebrenik, A., Mens, T. A historical dataset of software engineering conferences. In: *Proceedings of the 10th Working Conference on Mining Software Repositories*. IEEE Press. 2013, pp. 373–376.