



EVENTSKG: A 5-Star Dataset of Top-Ranked Events in Eight Computer Science Communities

Said Fathalla^{1,2}(✉) , Christoph Lange^{1,3} , and Sören Auer⁴

¹ Smart Data Analytics (SDA), University of Bonn, Bonn, Germany
`{fathalla,langec}@cs.uni-bonn.de`

² Faculty of Science, University of Alexandria, Alexandria, Egypt

³ Fraunhofer IAIS, Sankt Augustin, Germany

⁴ TIB Leibniz Information Centre for Science and Technology, L3S Research Center,
University of Hannover, Hannover, Germany
`soeren.auer@tib.eu`

Abstract. Metadata of scientific events has become increasingly available on the Web, albeit often as raw data in various formats, disregarding its semantics and interlinking relations. This leads to restricting the usability of this data for, e.g., subsequent analyses and reasoning. Therefore, there is a pressing need to represent this data in a semantic representation, i.e., Linked Data. We present the new release of the EVENTSKG dataset, comprising comprehensive semantic descriptions of scientific events of eight computer science communities. Currently, EVENTSKG is a 5-star dataset containing metadata of 73 top-ranked event series (almost 2,000 events) established over the last five decades. The new release is a Linked Open Dataset adhering to an updated version of the Scientific Events Ontology, a reference ontology for event metadata representation, leading to richer and cleaner data. To facilitate the maintenance of EVENTSKG and to ensure its sustainability, EVENTSKG is coupled with a Java API that enables users to add/update events metadata without going into the details of the representation of the dataset. We shed light on events characteristics by analyzing EVENTSKG data, which provides a flexible means for customization in order to better understand the characteristics of renowned CS events.

Keywords: Scientific Events Ontology · Scholarly data ·
Linked open data · EVENTSKG · Metadata Analysis · 5-star dataset

1 Introduction

Recently, large collections of events metadata have become publicly available on the Web. However, this data is not well-organized, distributed over digital libraries and event websites, and not integrated. The existence of such data freely available online has motivated us to create a comprehensive dataset for renowned

computer science events. A good practice in the Semantic Web community is to publish datasets as Linked Data. Therefore, this paper introduces the second release of the EVENTSkg dataset, which is the new release of the EVENTSkg dataset [5], as Linked Data. Currently, EVENTSkg contains 73 event series (i.e., 75% series in addition to the first release) from eight CS communities¹: Artificial Intelligence (AI), Software and its engineering (SE), World Wide Web (WEB), Security and Privacy (SEC), Information Systems (IS), Computer Systems Organization (CSO), Human-Centered Computing (HCC) and Theory of Computation (TOC). The latter two communities are new in the current release. Further new features of the new release include the use of the latest version of the Scientific Events Ontology (SEO)² (more details in Sect. 3), a Java API that has been developed for maintaining and updating the dataset, and a public Virtuoso SPARQL endpoint that has been established for querying the new release. EVENTSkg is a 5-star dataset [3], i.e., following a set of design principles for sharing machine-readable interlinked data on the Web, which enable data publishers to link their data to linked open data sources to provide context. Therefore, more *related* data can be discovered, enabling data consumers to directly learn about the data, thus increasing the value of the data and sharing the benefits from data already defined by others, i.e., enabling incremental work rather than working from scratch. In EVENTSkg, we map research fields and both countries and cities to SEO and DBpedia respectively. Events are linked by research fields, hosting country, and publishers. A key overarching research question that motivates our work is: *What is the effect of digitization on scholarly communication in computer science events?* In particular, we address specific questions such as the following:

- *What is the trend of submissions and publications of renowned CS events?*
- *Which CS communities have attracted increasing attention in the last decade?*
- *Have top-ranked events changed their publishers?*
- *Which continent hosts most events of a given CS community?*

A key benefit of this work is the availability of the dataset as LOD, as well as a collection of open source tools for maintaining and updating the dataset, with the goal to ensure the sustainability and usability of the dataset and to support the analysis of scholarly events metadata. EVENTSkg can answer the following competency questions:

- *What is the average acceptance rate of renowned Software Engineering events?*
- *To what venues can I submit my work to be published by Springer?*
- *Which CS communities have a growing popularity over the last decade?*

The analysis results presented in this work give some insights to answer these questions. The dataset documentation page (cf. Table 2) describes the dataset

¹ These communities have been identified using the ACM Computing Classification System: <https://dl.acm.org/ccs/ccs.cfm>.

² <https://w3id.org/seo#>.

structure and its releases. It also contains a description of each release and a chart comparing statistics of each release. The URI of each resource, i.e., of an individual events or an event series, is formed of the dataset URL (<http://w3id.org/EVENTSKG-Dataset/ekg>) followed by the event's acronym and the year, e.g., <http://w3id.org/EVENTSKG-Dataset/ekg#ESWC2018> is the URI of the 2018 ESWC conference. EVENTSKG stores data relevant to these events in RDF, and each event's metadata is described appropriately by means of the data and object properties in the Scientific Events Ontology (SEO). All data within EVENTSKG is available as dumps in the JSON-LD, Turtle, and RDF/XML serializations, and via our SPARQL endpoint. Previous versions of EVENTSKG are archived in data dumps in both CSV and RDF formats. CSV data is available in ZIP archives, with one CSV file per event series. Updating resources and adding new ones to a Linked Dataset is a time consuming and error-prone task. EVENTSKG is coupled with a Java API for this purpose (more details in Sect. 5). To illustrate the potential use of EVENTSKG for tracking the evolution of scholarly communication practices, we analyzed the key characteristics of scholarly events over the last five decades, including their geographic distribution, time distribution over the year, submissions, publications, ranking in several ranking services, publisher, and progress ratio (cf. Sect. 2). An exploratory data analysis is performed aiming at inferring facts and figures about CS scholarly events over the last five decades. We believe that EVENTSKG will bridge the gap between stakeholders involved in the scholarly events life cycle, starting from event establishment through paper submission till proceedings publishing, including events organizers, potential authors, publishers, and sponsors. This is, therefore, an area of particular interest for: (1) *event organizers* to measure the impact of their events in comparison with other events in their community or events in other communities by identifying success factors, (2) *potential authors* to assess the characteristics of high-impact events for deciding to what events to submit their work, (3) *scientometrics researchers* to identify metrics to consider when ranking scholarly events, and (4) *proceedings publishers* to study the impact of their events, or of other events they would be interested to publish.

The remainder of the article is structured as follows: Sect. 2 presents a brief review of the related work. Section 3 outlines the SEO ontology. Section 4 presents the main characteristics of EVENTSKG. Section 5 explains its curation process. Section 6 presents some examples of queries that EVENTSKG can answer. Section 7 discusses the results of analyzing the EVENTSKG data. Section 8 concludes and outlines possible future work.

2 Related Work

The past decade has witnessed increased attention to providing a comprehensive semantic description of scholarly events and their related entities [2, 4, 6, 11, 13]. Recently, publishing scholarly events metadata as Linked Data has become of prime interest to several publishers, such as Springer and Elsevier. Few researchers have addressed the problem of identifying the characteristics of

renowned events in CS overall or within a particular CS communities. However, none of them provides services to ease the process of gaining an overview of a field, which is the contribution of this work. Overall, we found that the characteristics of these events have not been dealt with in depth. We have divided the literature on this topic into two areas: datasets, and analysis of scholarly events metadata.

Datasets. The Semantic Web Dog Food (SWDF) dataset is one of the pioneers of datasets of comprehensive scholarly communication metadata [13]. The first attempt to create a dataset containing metadata of top-ranked computer science events categorized by five communities is represented by our own EVENTS dataset [4]. EVENTS contains metadata of 25 event series in terms of 15 attributes, such as the geographical distribution (by hosting country) and the time distribution over the year. The main shortcoming of this dataset is that it is published as individual RDF dumps, which are not linked using well-formed URIs in a linked data style. This results in losing the links between dataset elements, such as events addressing topics in the same field or being hosted in the same country. Vasilescu et al. [15] presented a dataset of just eleven renowned software engineering conference series, such as ICSE and ASE, containing accepted papers along with their authors, programme committee members and the number of submissions each year. Luo and Lyons [12] presented a dataset with the metadata, including authors' names, the number of papers, and the number of workshops of every edition of the annual conference of the IBM Centre for Advanced Studies (CAS) in the period 1993–2017.

Metadata Analysis. Osborne et al. [14] developed the Rexplore tool for exploring and making sense of scholarly data through integrating visual and statistical analytics. Hiemstra et al. [11] analyzed the trends in information retrieval research community through co-authorship analysis of ACM SIGIR conference proceedings. Barbosa et al. [2] studied the evolution of Human-computer interaction research field in Brazil through analyzing the metadata of full papers of 14 editions of the Brazilian HCI conference (IHC). Agarwal et al. [1] analyzed the bibliometric metadata of seven ACM conferences in information retrieval, data mining, and digital libraries. Fathalla et al. [6, 7] analyzed the evolution of key characteristics of CS events over a period of 30 years using descriptive data analysis, including continuity, geographic and time distribution, and submission and acceptance numbers

Despite these continuous efforts, most of the previous work has only focused on analyzing metadata of events of one series. What additionally distinguishes our work from the related work mentioned above, including our own previous version, is the creation of a Linked Dataset, with dereferenceable IRIs, under a persistent URL following W3C standards and best practices. In addition, EVENTSKG can be queried through a SPARQL endpoint.

3 Scientific Events Ontology

The Scientific Events Ontology (SEO) [8] is our ontology of choice to describe scientific events because it integrates the state-of-the-art ontologies for events in addition to its own vocabularies. SEO is the ontology of the OpenResearch³ platform for curating scholarly communication metadata. It does not only represent what happened, i.e., the scholarly event and its date and location, but also the roles that each agent played, and the time at which a particular role was held by an agent at an event. Best practices within the Semantic Web community (cf., e.g., [9]) have been considered when designing and publishing the ontology. SEO reuses several well-designed ontologies, such as the Conference Ontology⁴, FOAF, SIOC, Dublin Core and SWRC (Semantic Web for Research Communities), and defines some of its own vocabularies. All namespace prefixes are used according to prefix.cc⁵. The OR-SEO concepts used to represent events metadata in EVENTSKG are: `OrganisedEvent` (and its subclasses), `Site`, `EventSeries` (and its subclasses), `ResearchField` and `Agent`. Furthermore, several properties have been used, including data properties and object properties (Table 1).

Table 1. Events properties

Property	Type	Source	Description
<code>acronym</code>	datatype	conf-onto	The acronym of an event
<code>endDate</code>	datatype	conf-onto	The date of the last day of an event
<code>startDate</code>	datatype	conf-onto	The date of the first day of an event
<code>field</code>	datatype	seo	The research field which the event belongs to
<code>country</code>	datatype	DBpedia	The country hosting the event
<code>state</code>	datatype	seo	The state hosting the event (if applicable)
<code>city</code>	datatype	seo	The city hosting the event
<code>submittedPapers</code>	datatype	seo	The number of papers submitted to an event
<code>acceptedPapers</code>	datatype	seo	The number of papers accepted at an event
<code>acceptanceRate</code>	datatype	seo	The acceptance rate of an event
<code>eventWebsite</code>	datatype	seo	The website of an event
<code>belongsToSeries</code>	object	seo	The series which an event belongs to
<code>colocatedWith</code>	object	seo	Links an event to another, co-located one
<code>hasTrack</code>	object	seo	Specifies the different tracks associated to an event

³ <http://openresearch.org/>.

⁴ <http://www.scholarlydata.org/ontology/doc/>.

⁵ A namespace look-up tool for RDF developers: <http://prefix.cc/>.

4 EVENTSKG Characteristics

Currently, EVENTSKG covers three types of computer science events since 1969⁶: conferences, workshops, and symposia. EVENTSKG contains metadata of 73 events series, representing 1951 events with 17 attributes each. The total number of triples is 29,255, i.e., counting all available attributes of all events. EVENTSKG is a 5-star dataset [3]. Each resource is denoted by a URI and links to other datasets on the Web, such as DBpedia (to represent countries) and SEO entities (to represent terms such as “Symposium”), to provide context. The locations of the further EVENTSKG-related resources mentioned below are given in Table 2.

Availability and Best Practices: EVENTSKG is available as a Linked Dataset, with dereferenceable IRIs, under the persistent URL (<http://w3id.org/EVENTSKG-Dataset/ekg>), and as structured CSV tables. In addition, we established a SPARQL endpoint (using Virtuoso) to enable users to query the dataset. EVENTSKG is licensed under the terms of Creative Commons Attribution 4.0 Unported (CC-BY-4.0).

Extensibility: There are three dimensions to extend EVENTSKG to meet future requirements: (a) add more events in each community, (b) cover more CS communities and (c) add event properties, such as deadlines, registration fees, and chairs.

Documentation: The documentation of the dataset is available online⁷ and has been checked using the W3C Markup Validation Service⁸.

Sustainability: To ensure the sustainability of EVENTSKG, we developed an API for updating and maintaining the dataset. The dataset is replicated on its GitHub repository and our servers.

Announcement: we announced EVENTSKG on several mailing lists, such as the W3C LOD list⁹, the discussion list of the open science community¹⁰, and discussion forums, such as those of the Open Knowledge Foundation. We got valuable feedback, addressing issues such as inconsistencies in the data (in values, not in the semantics), from several parties, including researchers in our community and also librarians, e.g., from the German national library.

Quality assurance: the Vapour Linked Data validator is used to check whether EVENTSKG is correctly published according to the Linked Data principles and related best practices [9].

5 Data Curation

The lack of clear guidelines for data generation and maintenance has motivated us to propose a workflow for the curation process of EVENTSKG to serve as a guideline for linked datasets generation and maintenance. EVENTSKG is generated from metadata collected from several data sources (e.g., DBLP, WikiCFP, and digital libraries). Therefore, a data curation process is crucial. The curation

⁶ the date of the oldest events in the dataset.

⁷ <http://kddste.sda.tech/EVENTSKG-Dataset/>.

⁸ <https://validator.w3.org/>.

⁹ public-lod@w3.org.

¹⁰ open-science@lists.okfn.org.

Table 2. EVENTSKG-related resources

Resource	URL
Turtle file	http://kddste.sda.tech/EVENTSKG-Dataset/EVENTSKG_R2.ttl
RDF/XML file	http://kddste.sda.tech/EVENTSKG-Dataset/EVENTSKG_R2.rdf
JSON-LD file	http://kddste.sda.tech/EVENTSKG-Dataset/EVENTSKG_R2.json
SEO Ontology	https://w3id.org/seo/#
Issue Tracker	https://github.com/saidfathalla/EVENTSKG-Dataset/issues/
API	https://github.com/saidfathalla/EVENTSKG_API
GitHub repository	https://github.com/saidfathalla/EVENTS-Dataset
SPARQL endpoint	http://kddste.sda.tech/sparql
DataHub	https://datahub.ckan.io/dataset/eventskg
VoID	http://kddste.sda.tech/EVENTSKG-Dataset/VoID.nt
Documentation	http://kddste.sda.tech/EVENTSKG-Dataset/

of EVENTSKG dataset is an incremental process starting from the identification of top-ranked events in each CS community until the maintenance phase, which is performed continuously. It has been carried out comprising eight steps as shown in Fig. 1. During the curation process, several problems have been encountered, such as (1) identification of top-ranked events in each CS community, (2) data collection problems, such as data duplication, inconsistencies, and erroneous data, (3) data integration problems, such as integrating data about the same event collected from various data sources and unifying event names, (4) data transformation problems, such as converting unstructured to structured data, i.e., from text to CSV and consequently to RDF, and (5) LD generation, interlinking and validation. In the following subsections, we report only the major problems we faced, and how we solved these problems; mainly, they were data preprocessing problems.

Events Identification: At the very beginning, we should identify the top-ranked events in each CS community. To identify a subset of these events to be added to EVENTSKG, we used the following metrics, which are used widely by CS communities to identify top-ranked events in various CS communities. CORE¹¹: Computing Research and Education Association of Australasia uses community-defined criteria for ranking journals and events in the computing disciplines. The rankings have periodic rounds, usually every year, of updates for adding or re-ranking conferences. Based on these metrics an event can be ranked into eight classes – in decreasing order: **A***, **A**, **B**, **C**, **Australian**, **National**, **Regional**, and **un-ranked**. *QUALIS*: It uses the h-index as a performance measure for conferences. Based on the h-index percentiles, the conferences are ranked into

¹¹ <http://www.core.edu.au/>.

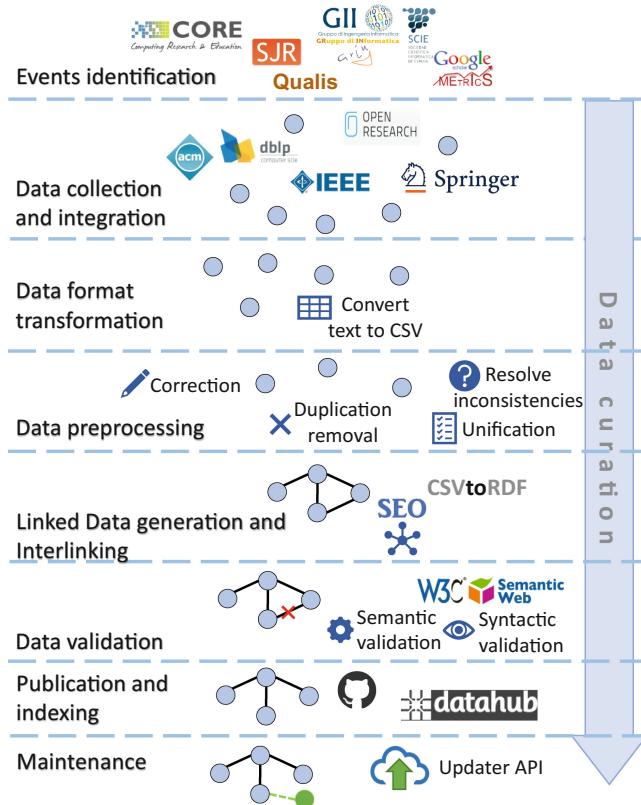


Fig. 1. Data curation of EVENTSKG.

seven classes – in decreasing order: A1, A2, B1, …, and B5. *ERA*¹²: Excellence in Research for Australia ranking is created by the Australian Research Council. The classes, in decreasing order, are: A+, B, and C. *GGS*¹³: The ratings are generated by an automatic algorithm based on existing international classifications. The classes are, in decreasing order: A++, A+, A, A-, B, B-, and C. While identifying top-ranked events in each community, we observed a heterogeneity of the ranking of them in the aforementioned services, e.g., FOGA is ranked A* in CORE, i.e., ranked 1st, while ranked B3 in QUALIS, i.e., ranked 5th. In addition, the rank of FSE in CORE is B, while its rank in GGS is A+ and in ERA it is A. Therefore, we propose the Scientific Events Ranking (SER) (available at <http://kddste.sda.tech/SER-Service/>), in which we unified the ranking of each event in the dataset using the sum of weight method. SER is represented by the function $SER: C \times Q \times E \times G \rightarrow S$, where C is the set of CORE classes, Q is the set of QUALIS classes, E is the set of ERA classes, G is the set of GGS classes,

¹² <https://www.arc.gov.au/excellence-research-australia/era-2018>.

¹³ <http://gii-grin-scie-rating.scie.es/ratingSearch.jsf>.

and S is the set of SER classes. The range of $SER(x)$ is defined in Eq. 1, where x is the sum of weights of each class in CORE, QUALIS, ERA, and GGS for each event series. We only choose the top-5 events according to SER.

$$SER(x) = \begin{cases} A+ & \text{if } 100 < x \leq 75 \\ A & \text{if } 75 < x \leq 50 \\ B+ & \text{if } 50 < x \leq 25 \\ B & \text{if } 25 < x \leq 0 \end{cases} \quad (1)$$

Data Collection and Integration: Still, metadata collection is considered a time-consuming task because of the diversity of data sources available on the Web. Actually, data collection for EVENTSKG is a semi-automated process in which the OpenResearch.org data crawlers are executed monthly to collect metadata of scientific events. In addition, we collected data from different unstructured and semi-structured data sources, such as IEEE Xplore, ACM DL, DBLP, and web pages. Therefore, this data should be integrated and cleaned to be exposed as Linked Data. Then, we initiate a data integration process, which involves integrating collected data from disparate sources into a unified view.

Data Preprocessing: The goal of the data preprocessing phase is to prepare the collected data for performing the analysis by integrating data from several data sources, eliminating irrelevant data and resolving inconsistencies. Three preprocessing tasks have been carried out: *data cleansing and completion*, *data structure transformation* and *event name unification*.

- *Data cleansing and completion*: involves removing duplicates, and identifying and correcting unsound data. Where we found incomplete information for some events, we complemented it as available. For instance, we double checked this information against the events' official websites or proceedings published in digital libraries, where they are trusted information sources. In addition, we periodically explore online digital libraries for the missing information.
- *Data structure transformation*: involves transforming cleaned data into a structured format, i.e., CSV.
- *Event name unification*: for the analysis purpose, we unified the names of all editions of an event series to the most recent name. For example, the unified name of The Web conference is *TheWeb*, formerly the World Wide Web conference (WWW).

Linked Data Generation and Interlinking: The adoption of the Linked Data best practices has led to the enrichment of data published on the Web by linking data from diverse domains, such as scholarly communication, digital libraries, and medical data [10]. The objective of this phase is to generate linked data from the less reusable, intermediate CSV representation. Using an ad-hoc transformation tool¹⁴, we transformed the CSV data to a RDF graph, after mapping several events attributes given in the CSV file to the corresponding OR-SEO properties.

¹⁴ <http://levelup.networkedplanet.com/>.

Using a comprehensive ontology as a dataset’s schema gives the ability to obtain insights from the data by applying inference engines. Interlinking is required to achieve the 5th star of the 5-star deployment scheme proposed by Berners-Lee [3].

Data Validation: The next step is to semantically and syntactically validate the RDF graph to ensure the quality of the data produced. The validation has been carried out using the W3C RDF online validation service¹⁵ to ensure conformance with the W3C RDF standards. The Hermit Reasoner is used to detect inconsistencies. Detecting inconsistencies is important because they result in a false semantic understanding of the knowledge. We resolve detected inconsistencies and periodically run the reasoner to ensure that no other inconsistencies arise after the interlinking process.

Data Publication: The objective of data publication is to enable humans and machines to share structured data on the Web. Therefore, EVENTSKG is published according to the Linked Data best practices [10] and it is registered in a GitHub repository (cf. Table 2). The commonly used way to let make a dataset easier to find, share and download is to index it in a public data portal, e.g., DataHub (cf. Table 2). A complete resource of the AAAI 2017 conference in the RDF/XML serialization can be found on the EVENTSKG documentation page.

Maintenance: To maintain EVENTSKG and to keep it sustainable, there are several challenges to be considered; here is how we address them: (1) A Java API for updating and maintaining the dataset has been developed, source code is available on GitHub (cf. Table 2). It facilitates the modification of EVENTSKG resources without going into the details of how this data is represented in the dataset since it has a natural language interface, in which casual users use only text fields, calendars and lists for modifying data, and it also facilitates the addition of new events to the dataset. For instance, metadata for each individual event, e.g., TheWeb, can be easily updated or added using a friendly user interface, and (2) GitHub Issue tracker: EVENTSKG has an issue tracker on GitHub, enabling the community to report bugs or to request features.

6 Use Case

This section presents some competency queries ($Q_1 - Q_4$) that EVENTSKG can answer. A concrete use case for querying EVENTSKG is to disclose the hidden characteristics of top-ranked events and also to help researchers in taking decisions on what event to submit their work to, or whether to accept invitations for being a chair or PC member. Event chairs will be able to assess their selection process, e.g., to keep 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 it against other competing events. For instance, “ Q_1 : What is the Average Acceptance Rate for a particular conference series, e.g., ESWC, in the last

¹⁵ <https://www.w3.org/RDF/Validator/>.

decade?” In addition, the productivity and the popularity of a CS community over time can be analyzed by studying the number of accepted and submitted papers respectively. For instance, “*Q₂: Compare the popularity of the CS communities in the past decade*” (Listing 1). Regarding country-level analysis, the popularity of a CS community in a particular country can be determined by such a query: “*Q₃: What are the top-5 countries hosting most of the events belonging to Security and Privacy in the past decade?*” Listing 2 shows such a query. In fact, EVENTSKG is not only able to answer quantitative questions, but it also provides qualitative information, such as countries hosted most events related to a particular community.

Listing 1. SPARQL query for comparing the popularity of the CS communities.

```
SELECT ?field (SUM(?sub) AS ?numOfSubmissions)
WHERE{
?e seo:field ?field.
?e conference-ontology:startDate ?d.
FILTER (?d >="2009-01-01T00:00:00.0000000+00:00"^^xsd:dateTime)
?e seo:submittedPapers ?sub.
}
ORDER BY DESC(?numOfSubmissions)
```

Listing 2. SPARQL query for finding top-5 countries host most of the events belonging to *Security and Privacy* in the past decade.

```
SELECT ?country (count(?country) as ?numOfEvents)
WHERE{
?e seo:heldInCountry ?country.
?e seo:field <https://w3id.org/seo#SecurityAndPrivacy>.
?e conference-ontology:startDate ?sd.
FILTER(?sd >="2009-01-01T00:00:00.0000000+00:00"^^xsd:dateTime)
}
GROUP BY(?country)
ORDER BY DESC(?numOfEvents)
LIMIT 5
```

7 Metadata Analysis

In this section, we present a part of the exploratory data analysis we performed on EVENTSKG. Further details can be found in [4,5,7]. The objective here is to emphasize the usefulness of EVENTSKG in exploring new features and unknown relationships in the data to provide recommendations. Furthermore, we summarize the main characteristics of top-ranked CS events using visual methods. We report the results of analyzing metadata of events, including the proceedings publishers, time distribution, geographical distribution (with two different granularities) and events progress ratio. These results provide some insights towards answering the research questions mentioned in Sect. 1.

Time distribution (TD): refers to the month of each year in which an event takes place. We computed the frequency of occurrence, in terms of the month of the year, of top-5 events (identified using the SER ranking) for each event since its establishment. Figure 2 shows the most frequent month in which events take place along with the number of editions of each event. We observed that most of the renowned events usually took place around the same month each year. For instance, CVPR has been held 28 times (out of 31) in June and PLDI has been held 33 times (out of 36) in June. This helps potential authors to expect when the event will take place next year, which helps with the submission schedule organization.

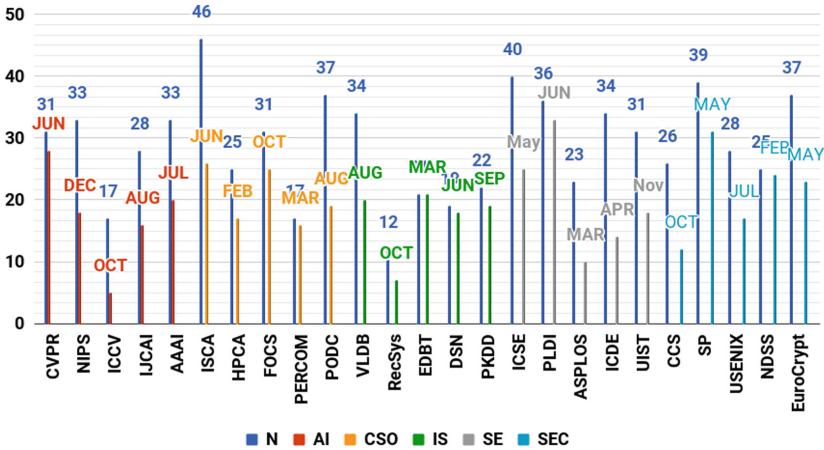


Fig. 2. TD of all events in terms of the most months where the event was held.

Geographical distribution (GD): refers to the distribution of events among countries (country-level GD) and continent (continent-level GD) each year since the beginning. We recorded, for each distinct location (either a country or continent), the number of times the event took place there. Events in EVENTSKG were distributed among 69 countries, with the USA having hosted the largest number (of 1042) events, then Canada comes with 124 events, then Italy, France, and Germany with 67, 67 and 64 events respectively. *Continent-level GD* refers to the frequency of occurrence of events among continents each year since the beginning. We computed the frequency of occurrence f_{ij} of all events belonging to community i in continent j . Then, we normalized these values to q_{ij} to ensure that the frequencies of occurrence of events in each community (C) sum up to one (Eq. 2).

$$f_{ij} = \sum_{k \in C} E_{ijk} , \quad q_{ij} = \frac{f_{ij}}{\sum_{k=1}^m f_{kj}} \quad (2)$$

Here, E_{ijk} is the number of events of an event series k in a community i taking place in continent j , and m is the number of event series in each community. As

shown in Table 3, Europe hosted IS events the most, followed by SEC events. North America has almost the same ratio for all communities. The remarkable observation emerging here is that Africa and South America host a significantly low number of events in all communities. For instance, South America hosted only four AI events and three IS events, while Africa hosted only one IS and one SE event. On the other hand, North America hosted the largest number of events (f_{ij}) in to all communities. *Country-level GD* refers to the change of the location of each event from year to year and denoted by ΔL_n (Eq. 3), where l_n is the location of an event in a year and l_{n+1} is the location of the same event in the next year.

$$\Delta L_n = \begin{cases} 1 & \text{if } l_n \neq l_{n-1} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

We computed the mean of these changes ($\bar{x} = (\sum_{i=0}^{n-2} (l_i - l_{i-1}))/n$) to measure the rate of the distribution of each event since the beginning. The higher this value is for an event, the more frequently the host country of an event changed. For instance, ICCV and ISMAR have $\bar{x} = 1$, which means that they moved to a different country every year, while SP and DCC have $\bar{x} = 0$, which means that they remained in the same country every year.

Table 3. Normalized frequency of occurrence (q_{ij}) of events by continent.

q_{ij}	Europe	N. America	Asia	Africa	S. America	Australia
AI	0.06	0.13	0.11	0.00	0.44	0.18
CSO	0.08	0.14	0.11	0.00	0.00	0.12
HCC	0.08	0.15	0.11	0.00	0.00	0.06
IS	0.22	0.11	0.15	0.50	0.33	0.24
SE	0.13	0.15	0.22	0.50	0.11	0.18
SEC	0.16	0.13	0.05	0.00	0.00	0.00
TOC	0.13	0.13	0.08	0.00	0.00	0.06
WWW	0.13	0.05	0.18	0.00	0.11	0.18

Progress ratio (PR): refers to the progress of an event in a given year within a fixed period of time. We define the PR of an event by the ratio of the number of publications of that event in a given year to the total number of publications in a given period of time. The progress ratio for an event e in a year y is defined in Eq. 4, where $P_y(e)$ is the number of publications of e in y and n is the number of years in the time span of the study. We computed the PR of the top-ranked events in each CS community in the period 1997–2017. As shown in Fig. 3, the PR of all events had a slight rise in the period 1997–2005; then, they all rose noticeably in the last decade. Overall, events of all CS communities have shown a drastic increase in PR since the beginning. We consider this to be an

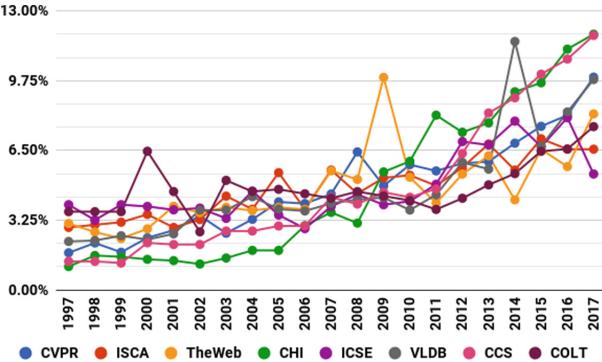


Fig. 3. PR of the top event in each CS sub-community in the last two decades.

effect of digitization, which has made event organization and paper submission considerably easier, thanks to conference management systems.

Publishers: It is observed that several events series organizers publish the proceedings of their events in their own digital library, e.g., AAAI, VLDB, or NIPS. On the other hand, ACM publishes the proceedings of 42% of the events in EVENTSKG, and IEEE comes next with 26%.

$$PR_y(e) = \frac{P_y(e)}{\sum_{i=1}^n P_i(e)} \quad (4)$$

8 Conclusions and Future Work

This paper presents a new release of the EVENTSKG dataset, a 5-star Linked Dataset, with dereferenceable IRIs, of all events of the 73 most renowned event series in computer science. The SEO ontology is used as the reference model for creating the dataset. We proposed a workflow of the curation process of EVENTSKG, starting from events identification until the publication and maintenance of the dataset. In addition, we present a new event ranking service (SER), which combines the rankings of CS events from four well-known ranking services. To the best of our knowledge, this is the first time a knowledge graph of metadata of top-ranked events in eight CS communities has been published as a linked open dataset. The dataset is coupled with an API for updating and maintaining the dataset, without going into the details of how this data is represented. We analyze EVENTSKG content over the last 50 years but found, during data acquisition, that there is not much information about events prior to 1990, in particular on the number of submissions and accepted papers. The most striking findings from the analysis of EVENTSKG's data are:

- The progress ratio of all events kept growing over the last two decades, most likely thanks to the digitization of scholarly communication,

- The USA have hosted most editions of events in all communities, followed by Canada, Italy, France, and Germany,
- The most of the events have a high distribution among countries to attract potential authors around the world,
- ACM publishes most of the proceedings of the events, and IEEE comes next,
- Europe hosted IS events the most, followed by SEC events, North America has almost the same ratio for all communities, and
- Africa and South America hosted a significantly low number of CS events.

These findings highlight the usefulness of EVENTSKG for events organizers, researchers interested in data publishing, as well as librarians. Finally, we believe that EVENTSKG can closes an important gap in analyzing the productivity and popularity of CS communities, i.e., publications and submissions, and it is of primary interest to steering committees, proceedings publishers and prospective authors.

To further our research, we are working in automating its subtasks; i.e., Data cleansing and completion, Data structure transformation and Event name unification. We are also planning to add more events from other fields of science, such as Physics, Mathematics, and Engineering, in addition to events from other CS communities such as Networks, Hardware and Applied computing. Furthermore, extending the OR-SEO ontology to cover authors, affiliations, titles and keywords in addition to adding a set of features to each event series that could be used to efficiently compare events in the same community, such as acceptance rate, h-index, and organizers' reputation, defined, e.g., in terms of their h-index and i10-index. Finally, we are planning to adopt a disambiguation mechanism for different events that have the same acronym, and to perform more complex semantic data analysis by querying EVENTSKG and automatically generating charts and figures from the obtained results.

Acknowledgments. This work has been supported by the European Union through the H2020 ERC ScienceGRAPH project (GA no. 819536). First author would like to thank the Ministry of Higher Education of Egypt for the financial support to conduct this work.

References

1. Agarwal, S., Mittal, N., Sureka, A.: A glance at seven ACM SIGWEB series of conferences. *SIGWEB Newslett.* (Summer), 5:1–5:10 (2016)
2. 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. *Scientometrics* **110**(1), 275–300 (2017)
3. Berners-Lee, T.: Is your linked open data 5 star. In: Berners-Lee, T. (ed.) *Linked Data*. Cambridge: W3C (2010)
4. Fathalla, S., Lange, C.: EVENTS: a dataset on the history of top-prestigious events in five computer science communities. In: González-Beltrán, A., Osborne, F., Peroni, S., Vahdati, S. (eds.) *SAVE-SD 2017-2018. LNCS*, vol. 10959, pp. 110–120. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01379-0_8

5. Fathalla, S., Lange, C.: EVENTSKG: a knowledge graph representation for top-prestigious computer science events metadata. In: Nguyen, N.T., Pimenidis, E., Khan, Z., Trawiński, B. (eds.) ICCC 2018. LNCS (LNAI), vol. 11055, pp. 53–63. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98443-8_6
6. Fathalla, S., Vahdati, S., Lange, C., Auer, S.: Analysing scholarly communication metadata of computer science events. In: Kamps, J., Tsakonas, G., Manolopoulos, Y., Iliadis, L., Karydis, I. (eds.) TPDL 2017. LNCS, vol. 10450, pp. 342–354. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67008-9_27
7. Fathalla, S., Vahdati, S., Auer, S., Lange, C.: Metadata analysis of scholarly events of computer science, physics, engineering, and mathematics. In: Méndez, E., Crestani, F., Ribeiro, C., David, G., Lopes, J.C. (eds.) TPDL 2018. LNCS, vol. 11057, pp. 116–128. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00066-0_10
8. Fathalla, S., Vahdati, S., Auer, S., Lange, C.: The scientific events ontology of the openresearch.org curation platform. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, pp. 2311–2313. ACM (2019)
9. Gyrard, A., Serrano, M., Atemezing, G.A.: Semantic web methodologies, best practices and ontology engineering applied to Internet of Things. In: 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT). IEEE (2015)
10. Heath, T., Bizer, C.: Linked data: evolving the web into a global data space. *Synth. Lect. Semant. Web: Theory Technol.* **1**(1), 1–136 (2011)
11. 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)
12. Luo, D., Lyons, K.: CASCONet: A Conference dataset. arXiv (2017)
13. Möller, K., Heath, T., Handschuh, S., Domingue, J.: Recipes for semantic web dog food — the ESWC and ISWC metadata projects. In: Aberer, K., et al. (eds.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 802–815. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_58
14. Osborne, F., Motta, E., Mulholland, P.: Exploring scholarly data with rexplore. In: Alani, H., et al. (eds.) ISWC 2013. LNCS, vol. 8218, pp. 460–477. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41335-3_29
15. Vasilescu, B., Serebrenik, A., Mens, T.: A historical dataset of software engineering conferences. In: 10th Working Conference on Mining Software Repositories. IEEE Press (2013)