Towards Facilitating Scholarly Communication using Semantic Technologies

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Said Mohamed Fathalla Abdelmaged
aus
Alexandria, Ägypten

Bonn 16.09.2020

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter:	Prof. Dr. Sören Auer
2. Gutachter:	Prof. Dr. Jens Lehmann

Tag der Promotion:	29.04.2021
Erscheinungsjahr:	2021

Abstract

Web technologies have substantially stimulated the submission of manuscripts, publishing scientific articles, as well as the organization of scholarly events, especially virtual events, when a global crisis occurs, which consequently restricts travels across the globe. Publication in scholarly events, such as conferences, workshops, and symposiums, is essential and pervasive in computer science, engineering, and natural sciences. The past years have witnessed significant growth in scholarly data published on the Web, mostly in unstructured formats, which immolate the embedded semantics and relationships between various entities. These formats restrict the reusability of the data, i.e., data analysis, retrieval, and mining. Therefore, managing, retrieving, and analyzing such data have become quite challenging. Consequently, there is a pressing need to represent this data in a semantic format, i.e., Linked Data, which significantly improves scholarly communication by supporting researchers concerning analyzing, retrieving, and exploring scholarly data. Notwithstanding the considerable advances in technology, publishing and exchanging scholarly data have not substantially changed (i.e., still follows the document-based scheme), thus restricting both developments of research applications in various industries as well as data preservation and exploration. This thesis tackles the problem of facilitating scholarly communication using semantic technologies. The ultimate aim is improving scholarly communication by facilitating the transformation from a document-based to knowledge-based scholarly communication, which helps researchers to examine science itself with a new perspective. Key steps towards the goal have been taken by proposing methodologies as well as a metrics suite for publishing and assessing the quality of scholarly events concerning several criteria, in particular, Computer Science as well as Physics, Mathematics, and Engineering. Within the framework of these criteria, steps towards assessing the quality of scholarly events and recommendations to various stakeholders have been taken. Furthermore, we engineered the Scientific Events Ontology in order to enable the enriched semantic representation of scholarly event metadata. Currently, this ontology is in use on thousands of *OpenResearch.org* events wiki pages. These steps will have far-reaching implications for the various stakeholders involved in the scholarly communication domain, including authors, sponsors, reviewers, publishers, and libraries. Most of the scholarly data publishers, such as Springer Nature, have taken serious steps towards publishing research data in a semantic form by publishing collated information from across the research landscape, such as research articles, scholarly events, persons, and grants, as knowledge graphs. Interlinking this data will significantly enable the provision of better and more intelligent services for the discovery of scientific work, which opens new opportunities for both scholarly data exploration and analysis. In the direction to this goal, we proposed the Science Knowledge Graph Ontologies suite, which comprises four OWL ontologies for representing the scientific knowledge in various fields of science, including Computer Science, Physics, and Pharmaceutical science. Besides, we developed an upper ontology on top of them for modeling modern science branches and related concepts, such as scientific discovery, instruments, and phenomena.

Contents

CHAPTER 1

Introduction

At the beginning of the 1990s, the World Wide Web (the Web) initiative, proposed by T. Berners-Lee [1], is designed to bring a comprehensive information space into existence comprising linked documents, i.e., “Web of Documents”. This initiative has been made by developing a network of documents using a Hypertext Markup Language (HTML)[1]. The Internet and the Web represent considerable progress for the retrieval and dissemination of scientific information across the world. Therefore, the amount of scientific data on the Internet and the number of publication venues, such as scholarly events and journals, continue to increase day by day or even hourly [2]. These data have been made available online by a diversity of sources, such as individuals and data publishers, through digital libraries. There is a significant potential benefit from the integration of the heterogeneous structure of the data (i.e., those data structures that contain a variety of dissimilar data types) available on the Web in a way that allows users to get the needed information efficiently. At that time, exploring these data was limited to keyword-based search. Search engines were not able to give satisfying results as they do not know the exact meaning of the keywords used. In order to obtain satisfying results, search engines should be able to understand the semantics and the interrelationships among the data being queried. To improve traditional search engines on the Web, semantic representation of the concepts and relationships of the data on the Web became a necessity.

The rapid development in the Web technologies reinforce the transition from Web of documents to the Web of data (i.e., the Semantic Web) [3]. The Semantic Web aims at representing information in a machine-readable way; hence the term “Linked Data” comes. Linked Data refers to the principles of publishing and interlinking structured data on the Web with the purpose of supporting semantic queries and knowledge inference [4]. These principles are developed upon standard Web technologies, such as Hypertext Transfer Protocol (HTTP) and Uniform Resource Identifiers (URIs). The primary objective of Linked Data is supporting sharing and reusing across the Web. A key factor in the reusability of data is the extent to which it is well structured [4]. The more well-defined the data, the more efficiently the machine can understand and reuse. When Linked Data is published as open data, then it is called “Linked Open Data” (LOD). LOD is published on the public Web and licensed under one of several open licenses permitting reuse. Publishing Linked Open Data enables distributed SPARQL queries to find information efficiently, as compared to a keyword-based search strategy.

The World Wide Web Consortium (W3C) proposed a collection of Semantic Web technologies that create and query Linked Data, such as 1) the RDF data model for knowledge representation,

[1] https://www.w3.org/html/

2) SPARQL for querying RDF stores, and 3) the Web Ontology Language (OWL) as a logical formalism to develop ontologies. These technologies together provide an environment where semantics-based systems can reason about that data. Ontologies are the formal representation of concepts and interlinking relations between them in a particular domain of discourse. In other words, they are a machine-readable description of concepts and relations among them in a specific domain. Recently, ontologies occupied a crucial role in capturing knowledge and providing a machine standard understanding the semantics of knowledge in many Computer Science areas, such as data mining [5], scholarly communication [6, 7] and bioinformatics [8, 9]. The real-world instantiation of ontology concepts forms a graph of knowledge called Knowledge Graph (KG), i.e., ontologies represent metadata for a knowledge graph. Knowledge graphs facilitate representation of facts about people, places, and things in a structured way and how these facts are all related. Besides, Knowledge graphs enable integrating data from heterogeneous sources and machine-driven semantic reasoning. Exploiting semantics encoded in knowledge graphs helps to solve a broad range of problems in various domains, such as health care [10], education [11], and scholarly communication [12].

In the context of scholarly communication, scientific results produced by researchers and scholarly data publishers have been made available on the Web with low marginal costs and easily accessible through digital libraries, e.g., ACM DL and IEEE explore, and scholarly services which indexes metadata of scholarly literature, such as Google Scholar. Using Semantic Web technologies in the process of creating a comprehensive scholarly knowledge graph can significantly facilitate the management of scholarly metadata dispersed over the Web.

This thesis is considered a step towards facilitating scholarly data management with the objective of improving the processes involved in the scholarly communication life cycle, including data collection and analysis, dissemination, and preservation. This step enables far-reaching implications for the various stakeholders involved in the scholarly communication domain, including authors, reviewers, publishers, and libraries.

1.1 Motivation

Recent years have witnessed a continual growth in scholarly information: at least 114 million English-language scholarly documents are accessible on the Web [13], thanks to digitization, in both academia and industry. Scholarly information, emanating from scientific events, publishing houses, and social networks (e.g., *ResearchGate*), is available online in an unstructured format (e.g., Call for Papers (CfP) emails) or semi-structured format (e.g., events web pages). These formats limit the visibility and hamper the discovery of interconnected relationships for humans as well as machines. This plethora of scientific literature and heterogeneity of the metadata makes it increasingly difficult to keep an overview of the current state of research. Therefore, the semantic representation of such data has motivated many researcher to develop of various data models for scholarly communication [14].

This thesis is considered as a step in a long-term research agenda to create a paradigm shift from document-based to knowledge-based scholarly communication (cf. Figure 1.1). This transition supports retrieving, exploring, and comparing research findings based on an explicit semantic representation of the knowledge contained in scientific publications [7]. Semantically enriched representation of such information makes it easier to efficiently query and process scholarly data published on the Web [14]. In this section, we present three examples that motivate the problem of exploring scientific data concerning three dimensions: scholarly data management, analysis, and retrieval.

Motivating Example 1: Scholarly events quality assessment. Due to the rapidly growing amount of scholarly information, exploration of research data still stays challenging for researchers, scholarly metadata managers, social scientists, and librarians. Aggregation of metadata from several data repositories, digital libraries, and scholarly metadata warehouses enables comprehensive analysis, derives insights, and supports concluding meaningful correlations between them. The absence of guiding principles for Linked Data curation has motivated us to propose a methodology for conducting the curation process of Linked Open Datasets for scholarly communication. Various stakeholders of scholarly communication need to assess the quality of scholarly events for different purposes: 1) *event organizers* – to trace their events' progress/impact, 2) *authors* – to identify renowned events to submit their research results, as well as 3) *proceedings publishers* – to know the impact of the events whose proceedings they are publishing. For instance, a publisher (e.g., Springer) desires to decide whether to publish the proceedings of a particular event and possibly become a sponsor. The decision mainly depends on several criteria concerning the quality of the events. Besides general criteria such as the impact factor of journals and the acceptance rate of events, there are community-defined criteria for the ranking of journals and events. Such criteria are not standardized nor maintained by a central instance, but are transferred from seniors to juniors. This thesis sheds light on these criteria across disciplines aiming at answering the following questions: 1) *How important are events for scholarly communication in the respective communities*? and 2) *What makes an event a high-ranked target in a community*? Statistical analysis of metadata of events, such as title, acronym, start date, end date, number of submissions, number of accepted papers, city, state, country, event type, field, and homepage, can give answers to such questions.

Motivating Example 2: Identifying appropriate scholarly events concerning certain criteria. Collecting, integrating, and analyzing metadata of scholarly events, important dates, locations, themes, or research area, is of paramount importance for pursuing scientific progress. An important topic in semantic publishing is the development of semantic models to describe the meaning and the relationships between scholarly data elements, thus enabling machines to interpret meaning, which is crucial for facilitating the information needs of stakeholders, including authors and publishers [15]. Given the heterogeneity of events metadata scattered across the Web, semantic representation of such information involves modeling event metadata covering different types of entities involved, such as persons, organizations, location, roles of persons before/during/after the event, etc. This problem has motivated us to tackle the problem of representing scientific events metadata using Semantic Web technologies, i.e., integrating existing vocabularies and making the relationships and interconnections between event data explicitly defined. Thus, supporting the transformation from a "Web of documents" to a "Web of data" in the scientific domain and enabling efficient information retrieval and analysis. Here, we show an example of the discovery of a potential list of scholarly events within a particular community for possible types of stakeholders among researchers, such as organizers, authors, reviewers, sponsors, speakers, and participants, etc. Finding the appropriate scholarly events is crucial for each of these roles and parties; however, this can only be developed over time by the researchers themselves, which requires time and experience, and it is prone to omissions. Consider the case, a researcher (e.g., Said Fathalla) wants to identify appropriate events to submit his work or accept a program committee member or organization role invitation. These events, according to his own rules, should satisfy specific criteria, such as the topics

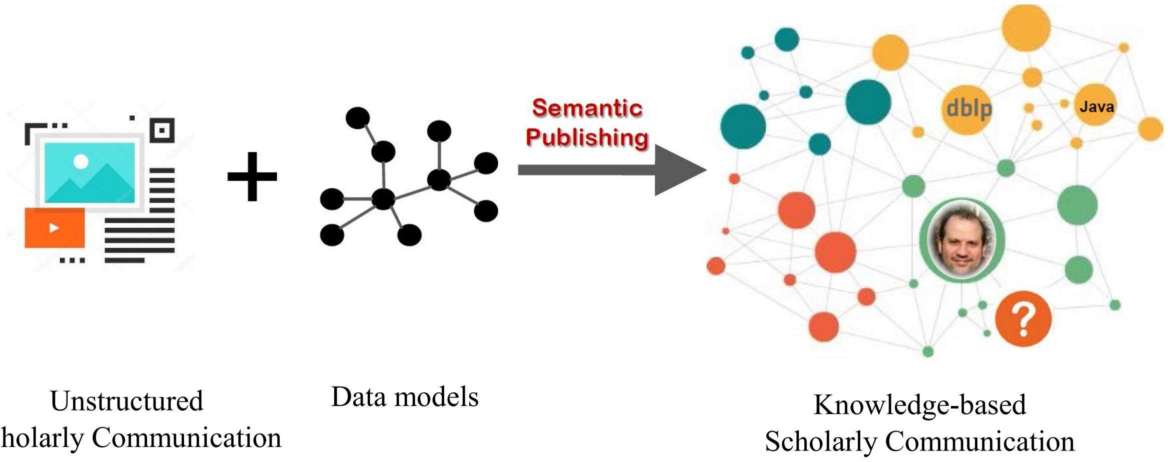


Figure 1.1: **The overarching objective**. Transition from document-based to knowledge-based scholarly communication.

covered by an event, proceedings publisher, geographical restrictions, and time. The trivial way is to ask colleagues and read the calls for papers published in conference management services (e.g., WikiCFP[2]), which is time-consuming and takes effort. For example, with these two sources, he can find events in Europe and related to his field of interest. However, the call for papers gives no clues about the quality of the event. Therefore, *Fathalla* has to make an extra effort by checking events Websites for more information, finding previous editions and related events, and possibly has to read the proceedings and explore the number of citations to get more information about these events since there is yet no common metrics for comparing them.

Motivating Example 3: Structuring, systematizing and efficiently comparing research results. Despite significant advances in technology in the last decades, the way how research work is accomplished and how researchers communicated has not changed much. Researchers still encode their findings in sequential text accompanied by illustrations and wrap these into articles, which are mostly published in printed form or as semi-structured PDF documents. The current approach for constructing, organizing, and evaluating research contributions is through writing survey or review articles. Typically, authors of such articles collect several articles about a particular research topic and (a) develop an organization scheme with feature comparisons, (b) provide a conceptualization of the research domain with mappings to the terminologies used in the individual articles, (c) compare and possibly benchmark the research approaches, implementations, and evaluations described in these articles and (d) identify directions for future research. As a result, survey and review articles significantly contribute to structuring a research domain and make its progress more transparent and accessible. However, such articles still share the same deficiencies as their original research counterparts. The content is not machine-understandable (i.e., not represented according to formal knowledge representation). Traditional representation prevents systematic identification of conceptualization problems, i.e., defining the intending meaning of entities, as well as supports the building of intelligent search, exploration, and browsing applications on top. The problem of exploring the information embedded in scientific articles has motivated many researchers to develop approaches to support

[2] http://www.wikicfp.com/

this process. Let us assume a group of researchers who want to generate an overview of relevant information about a particular research topic, e.g., *Ontology Matching.* Consider a researcher (e.g., *Fathalla*) who wants to get an overview of the topic "querying over federated SPARQL endpoints" in order to write a survey article or even a literature review section in a research paper. Using traditional scholarly search engines, such as Google scholar, too many results are retrieved (At the time of writing this thesis, 3,000 hits were retrieved by Google scholar for the query "Federated SPARQL Endpoints"). The most relevant one is a survey paper written by Rakhmawati et al. [16]. This one is the paper that *Fathalla* can read to get an overview of the state-of-the-art techniques/tools. There is a set of challenges waiting for him since he should find, read, analyze, and understand the articles included in this survey, which of course, will take a lot of time and effort. After overcoming these challenges, he yet did not acquire a comprehensive overview of that topic. On the other hand, Rakhmawati and her co-authors paid considerable effort and time to conduct this survey. Therefore, an approach to automatically generate an overview of, or compare, the state-of-the-art of a specific research topic will help researchers to identify the most relevant work, thus minimizing the effort and time spent by researchers to do it manually.

1.2 Problem Definition and Challenges

The research problem guiding the work of this thesis can be expressed by the question: How can scholarly data be understood by machines making data representation, metadata analytics, and information retrieval more efficient? On the way to achieve the goal, semantic data integration, analysis, and management problems have been encountered. Metadata Management implies a set of activities, which store data in a structured form for better usage. Metadata analytics are needed for different purposes, for example, solving challenges confronting scientists or interpreting data in a meaningful fashion aiming at consolidating decision-making. To enrich the infrastructure supporting the reuse of scholarly data, four foundational principles, i.e., FAIR principles [17], have been proposed to enhance scholarly data publishing through proper data management.

Despite all efforts in supporting scholarly communication by different means (see chapter 3), there are several challenges still remaining [18]. With the fast growth of digital publishing, managing and analyzing scholarly data (which contains information about millions of authors, papers, proposals, events, affiliations, and scholarly networks) has become quite challenging. Hence, the term *Big Scholarly Data* is coined. In the academic landscape, the use of big data analytics in the scholarly ecosystem has a significant influence on the ease of how scholarly data is managed, and research is performed [19]. In the following, we present the main challenges conducted by this thesis.

Challenge 1: Exploring the characteristics of the renowned scholarly events in Computer Science and other fields of science.

The past decade has witnessed increased attention to the analysis and representation of scholarly events and their related entities [20–26]. This plethora of scientific literature makes it increasingly difficult to explore the characteristics of renowned scholarly events. Researchers of different communities use different channels for publishing. The incorporation between these channels is based on a set of community-defined criteria for assessing the quality of them. For example, in some fields, such as biological science, journals are the leading channel for publication; however, some other fields, such as Computer Science, publishes mostly in events. Furthermore, community-defined criteria distinguish highly ranked instances of any particular class of channels as well as popular events and journals. Nevertheless, such criteria are not standardized nor centralized but generally, are transferred subjectively from seniors to juniors. However, a systematic and objective analysis of metadata supports researchers in better dissemination of results to the right communities. The challenge that arises here is proposing a metrics-suite of standard criteria in order to study the various characteristics of renowned scholarly events in different fields of science.

Challenge 2: Integrating heterogeneous scholarly events metadata.

Data about scholarly events is increasingly published on the Web, albeit often as raw dumps in unstructured formats, hence the heterogeneity comes. These formats immolate its semantics and interlinked relationships with other data, thus restricting the reusability of such data for, e.g., subsequent analyses. Therefore, there is a high demand to represent this data using semantic technologies, thus enabling implicit knowledge inference. The exponential growth of this data places excessive pressure on researchers who are working on scholarly communication to assess, analyze, and organize this massive amount of data generated day after day [27]. The existence of such data freely available online has motivated us to create a comprehensive dataset for renowned Computer Science events. It is a challenge to integrate heterogeneous scholarly events metadata distributed on the Web in order to create LOD that contains metadata of the renowned scholarly events belonging to various Computer Science communities. It is a good practice in the Semantic Web community is to publish datasets as Linked Data [28].

Challenge 3: Semantically representing knowledge about entities involved in the scholarly events domain.

Towards the development of an ontology for scholarly events, challenges started with identifying the pitfalls in the state-of-the-art models. In fact, the scholarly events domain itself relates entities from diverse aspects, including bibliographical information, and spatial and temporal data. Therefore, data models necessitate an effective integration of concepts and their semantics. After studying the domain and the state-of-the-art models, the diversity of information representation and large amounts of data pose high requirements to be addressed. The ontology should be maintained with respect to the evolution of Linked Data vocabularies and adaptable to other domains of science.

Challenge 4: Semantically representing research findings of different fields of science enabling intelligent services for the better discovery of the scientific work.

From one day to the next, researchers produce a considerable number of scholarly articles, mostly in PDF format, that needs to be explored, analyzed, interpreted, and understood by the community. The sheer amount of information being published in this way poses challenges for researchers since such a format does not make them efficiently accessible for comparative or other analyses and restricts knowledge discovery and the development of intelligent agents. To obtain an overview of the state-of-the-art in a particular research area, researchers tend to write a distinct type of scientific publications, called survey/review articles. However, exploring, analyzing, and comparing such articles require significant effort and time. Therefore, we faced the challenge of representing research contributions in a richer form (i.e., scholarly Knowledge Graphs), thus making them more comparable and accessible to semantic search engines. Besides, developing a knowledge graph-based approach for exploring scholarly Knowledge Graphs is required.

1.3 Research Questions

Based on the challenges mentioned earlier, we defined four research questions (RQs) as follows.

> **RQ1**: How can the characteristics of renowned scholarly events in different fields of science be utilized to assess their impact?

To address this question, the development of scholarly knowledge dissemination in four fields of science (Computer Science, Physics, Engineering, and Mathematics) has been analyzed. Particular attention has been given to the renowned events in eight Computer Science communities[3], including World Wide Web, Computer Vision, Software Engineering, Data Management, Computer Architecture, Knowledge Representation and Reasoning, Machine Learning, and Security and Privacy, targeting to answer: (a) *What is the orientation of submissions and acceptance rates of Computer Science events*? (b) *How did the number of publications of a Computer Science sub-community fluctuate*? (c) *Are high-impact events held around the same time slot each year*? and (d) *Which countries host the most events in each Computer Science community*? The diversity of the meaning of impact brings challenges for the development of a robust and widely accepted impact measures. This diversity limits the scope and quality of possible evaluations. To go beyond citation-based judgments, an extended list of metrics is required. Analyzing scholarly events metadata, such as event dates, the number of submitted and accepted articles, location, event type, and field, can help to answer such questions. In this thesis, exploratory and descriptive data analysis has been performed, aiming at exploring some facts and figures about Computer Science events over the last five decades. In order to systematize the evaluation, ten generic metrics that can be jointly applied for the selected communities have been defined and computed based on this data. As a result, the Scholarly Events Quality Assessment (SEQA) metrics suite has been proposed for assessing the quality of scholarly events metadata is proposed. SEQA aims at supporting scientometrics studies by helping to study the importance of scholarly events in different fields of science and to assess the comparative popularity and productivity of major Computer Science communities, in terms of submissions and publications. To be able to answer RQ1, we divide this question into three sub-questions: *RQ1.1) How important are events for scholarly communication in the respective communities*?, *RQ1.2) What makes an event a high-ranked target in a community*? and *RQ1.3) How can scholarly events be assessed using a mixture of metrics*?

> **RQ2**: How can we represent and integrate heterogeneous scholarly event metadata in knowledge graphs to facilitate scholarly data management and retrieval?

To address this question, aggregation of metadata from several data repositories, digital libraries, and scholarly social networking sites has been studied, which enables comprehensive data analysis intending to disclose valuable information, thus supporting decision-making. As a result, a conceptual framework, i.e., Scholarly Events Metadata Analysis (SEMA), for the data curation of scholarly events dataset is designed. The benefits of publishing these data as Linked Data are as follows: (a) *Data linking*: establish links between dataset elements so that

[3] Such a sub-community analysis was not possible for other fields due to the lack of data.

machines can explore related information, (b) *Semantic querying*: Linked Data can be queried using the SPARQL query language, (c) *Data enrichment*: inference engines could be used to infer implicit knowledge which does not explicitly exist, and (d) *Data validation*: semantically validate data against inconsistencies. The main goal is to facilitate the analysis of events metadata by enabling them to be queried using semantic query languages such as SPARQL. We believe that the resultant knowledge graph will bridge the gap between stakeholders involved in the scholarly events life cycle, starting from event establishment through paper submission till proceedings publishing, including event organizers, potential authors, publishers, and sponsors. By analyzing the data contained in this graph, we can answer RQ2.

RQ3: How can ontologies represent semantics encoded in entities involved in scholarly events and relationships among them?

Scholarly events have become a key factor in scholarly communication for many scientific domains. They are considered as the focal point for establishing scientific relations between scholarly objects, such as people (e.g., chairs and participants), places (e.g., location), actions (e.g., roles of participants), and artifacts (e.g., proceedings) in the scholarly communication domain. Metadata of scholarly events has been made available in an unstructured or semi-structured format, which hides the interconnected and complex relationships between them and prevents transparency. To answer this research question, we investigated the state-of-the-art models for describing scholarly events. Typically, these models differ by focus, i.e., event type, size, and level of abstraction, and they focus on the description of event metadata, including time, location, and topical classifications of events. As a result, an ontology (OR-SEO) for describing scholarly events is developed. OR-SEO enables a semantically enriched representation of scholarly event metadata, interlinked with other datasets and knowledge graphs. It does not only represent what happened, i.e., time and place of a scholarly event, but also the roles that each agent played, and the time at which a particular agent held this role in a specific event. OR-SEO is in use as the schema of the event pages of *OpenResearch.org*, a semantic media wiki platform for scientific events, research projects, publishers, journals, etc.

RQ4: How can published research in various fields of science be understood by machines, making information retrieval, analysis, and scholarly data management more efficient?

Researchers still encode their findings in sequential text accompanied by illustrations and wrap these into articles, which are mostly published in printed form or as semi-structured PDF documents online, which does not make them efficiently accessible for comparative or other analyses. Research work presented in such documents can be represented in a semantic format, resulting in a knowledge graph that describes the individual research problems, methodologies, approaches, experiments, and results in a structured and comparable format. For example, research addressing a specific problem can be automatically retrieved, approaches can be compared according to their features or concerning evaluation results in a particular defined setting. To respond to this question, ontologies representing research contributions in different fields of science, including Computer Science, Physics, and Pharmaceutical science, are developed. The objective of developing these ontologies is to support retrieving, exploring, and comparing research findings based on an explicit semantic representation of the knowledge contained in scientific publications. As a result of structuring and representing research contributions according these ontologies, scientific knowledge will become more comparable and accessible.

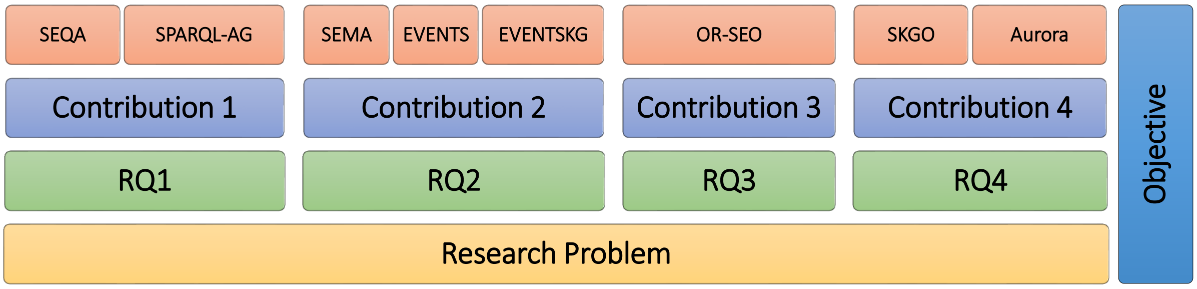


Figure 1.2: **Thesis Contributions**. The distribution of thesis contributions among the defined research questions.

1.4 Thesis Overview

With an eye towards preparing the reader to understand the purpose, nature, and direction of the thesis, we present an overview of the contributions of this thesis, research areas considered, references to scientific publications supporting this work, and an overview of the thesis structure.

1.4.1 Contributions

The contributions of this thesis have an impact on three main research areas:

- *Knowledge Management*: storing data in a structured form which makes it machine-readable and allowing it to be combined, thereby helping to maximize the added-value gained by scholarly digital publishing,
- *Information Retrieval*: utilizing Semantic Web technologies in order to exploit semantic knowledge within the data being queried for giving precise results and reasoning over data , and
- *Data Analysis and Visualization*: investigating, cleansing, transforming, modeling data, and visualization with the purpose of discovering useful information, which does not explicitly exist, thus supporting decision-making.

Each of these contributions aims to answer the research questions defined in section 1.3. Figure 1.2 depicts the distribution of thesis contributions among research questions and the relationship between individual contributions and the main thesis contributions. In the following, these contributions are described in detail:

Contribution 1: A metric suite for assessing the quality of scholarly events in various fields of science.

To answer RQ1, the characteristics of scholarly events in four fields of science, i.e., Computer Science, Physics, Engineering, and Mathematics, have been analyzed using descriptive as well as exploratory data analysis. The main objective is summarizing their main characteristics, often by plotting them visually. Out of this analysis, a metric suite, i.e., Scholarly Events Quality Assessment (SEQA) suite, based on the domain conceptualization has been proposed SEQA contains novel metrics, firstly defined by the author of this thesis, for scholarly events' impact assessment such as continuity, community popularity, field productivity, and progress ratio. This analysis provides a clear picture of the characteristics of high-quality scholarly events in the respective communities. After integrating and analyzing the results obtained from this study, the most noteworthy findings are highlighted. Based on these findings, a set of recommendations has been concluded to different stakeholders, involving event organizers, potential authors, and sponsors. Finally, this contribution is considered as a foundation for discovery, recommendation, and ranking services for scholarly events with well-defined measures.

Contribution 2: A methodology for scholarly events metadata curation and a Linked Open Dataset of renowned events in eight Computer Science communities. The lack of clear guidelines for data generation and maintenance has motivated us to propose the Scholarly Events Metadata Analysis methodology (SEMA) for the curation process of linked open scholarly events datasets to serve as a guideline for Linked Data generation and maintenance. A first attempt to create a dataset of renowned events in five Computer Science communities is represented by our EVENTS dataset [29]. It covers historical information about 25 top-renowned events, describing each of them with 15 metadata attributes. The main shortcoming of this dataset is that it is published as individual RDF dumps rather than one knowledge graph, by which it loses the potential links between dataset elements. As a result, a linked open dataset, i.e., EVENTSKG (the successor of EVENTS), of renowned events (together with events in EVENTS) in eight Computer Science communities has been created following best practices within the Semantic Web community (cf., e.g., [30]) for publishing linked data. EVENTSKG dataset[4] is a 5-star Linked Dataset, with dereferenceable IRIs, of all events of the most 75 renowned event series in eight Computer Science communities. In addition, we present a new event ranking service (SER)[5], which combines the rankings of Computer Science events from four well-known ranking services. To the best of our knowledge, this is the first time a knowledge graph of metadata of renowned events in eight Computer Science communities is published as a linked open dataset. EVENTSKG is coupled with an API (with a graphical user interface) that is used to update and maintain it, without going into the details of how the data is represented. EVENTSKG is generated from metadata collected from several data sources (e.g., DBLP, WikiCFP, and digital libraries). The curation of the EVENTSKG dataset is an incremental process starting from the identification of renowned events in each Computer Science community until the maintenance phase, which is performed continuously (see chapter 5 for more details). To widen the usability of EVENTSKG, a front-end, i.e., SPARQL-AG, that automatically generates and executes SPARQL queries for querying it has been developed. SPARQL-AG helps potential semantic data consumers, including non-experts and experts, by generating SPARQL queries, ranging from simple to complex ones, using an interactive web interface. This contribution covers RQ2.

Contribution 3: A scientific events data model representing metadata of scientific events and related entities. To answer the third research question, we present the Scientific Events Ontology (OR-SEO) [31, 32] to represent metadata of scientific events. We describe a systematic redesign of the information model that is used as a schema for the event pages of the *OpenResearch.org* community wiki and EVENTSKG dataset, reusing well-known vocabularies to make OR-SEO interoperable in different contexts. Standard methodologies and best practices have been considered when designing and publishing the ontology. OR-SEO was designed with a minimum of semantic commitment to guarantee maximum applicability for analyzing event metadata from diverse sources, and maximum reusability by datasets using the ontology for modeling different aspects of scientific

[4] `http://kddste.sda.tech/EVENTSKG-Dataset/`

[5] `http://kddste.sda.tech/SER-Service/`

events. Following best practices, OR-SEO emphasizes the reuse of events-related vocabularies and the alignment with concepts between them as well as the design and visualization patterns. The ontology is available using a persistent identifier (`https://w3id.org/seo#`); future versions can be collaboratively revised on a corresponding Git repository. To support knowledge discovery, a set of SWRL rules has been defined. The validation of the ontology is performed on syntactic and semantic levels. OR-SEO is now in use on thousands of *OpenResearch.org* events pages, which enables users to represent structured knowledge about events without having to deal with technical implementation challenges and ontology development themselves. This contribution covers RQ3.

Contribution 4: Data models representing research data in Computer Science, Physics and Pharmaceutical Science and an approach to generate overviews of research domains and their relevant artifacts based on these models.

Three generic vocabularies, part of the *Science Knowledge Graph ontologies* (SKGO), built upon an abstract level of scientific publications in Computer Science (SemSur[6]), Physics (PhySci[7]) and Pharmaceutical science (PharmSci[8]) domains as well as an ontology for describing various aspects of modern sciences (ModSci[9]) have been developed. These ontologies establish an infrastructure for the development of comprehensive machine-readable Knowledge Graphs for describing research data, including article metadata, document parts, authors, affiliations, evaluation, implementation, and funding projects. These ontologies aim at enabling linking of the domain-specific information extracted from scientific publications and to access data in a machine-understandable way. By this contribution, we initiate a step towards a paradigm shift away from purely document-based scholarly communication towards more knowledge-based methods of exchanging research results. Besides, an approach (Aurora) for the automatic unveiling of realm overviews for research artifacts has been proposed to provide structure to such information, thus enabling obtaining overviews about various research topics. Aurora is a knowledge-driven framework, which has been implemented on top of the crowd-sourcing platform *OpenResearch.org* and relies on extraction and curation methods for a scholarly knowledge graph. It facilitates the description of the scientific papers with the SemSur ontology [7] into a knowledge graph. Further, it enables the generation of surveys comprising comprehensive analytics that cover various research domain overviews by querying the knowledge graph. In addition, we provide domain overviews and suggestions such as which publication to read, which tools to use, where to publish similar results on sub-networks of the knowledge graph. A collaborative crowd-sourcing system can be employed to curate such information locked in the wisdom of community or represented in unstructured documents. Our evaluation confirms that Aurora, when compared to the current manual approach, reduces the effort for researchers to compile and read survey papers. This contribution covers RQ4.

1.4.2 List of Publications

At the beginning of each chapter, the publications on which the chapter is based on are referenced. The contributions of the publications in which Said Fathalla is the first author have been accomplished by himself and the role of the other authors is supervising and guiding the work. The papers co-authored by the following people are the result of their master theses closely

[6] `https://w3id.org/skgo/semsur`
[7] `https://w3id.org/skgo/physci`
[8] `https://w3id.org/skgo/pharmsci`
[9] `https://w3id.org/skgo/modsci`

supervised by the author of this dissertation: Aysegul Say, Zeynep Say, and Asha Arumugam. The entire list of publications completed during the Ph.D. studies can be found in Appendix A. In the following, the leading peer-reviewed publications building the basis of this thesis are outlined (see Figure 1.3). As shown in Figure 1.3, some publications, e.g., number (14), which published in ESWC 2019, contributes to more than one research discipline, i.e., Knowledge Management and Data Analysis in this case.

1. **Said Fathalla**, Sahar Vahdati, Sören Auer, and Christoph Lange. *Towards a Knowledge Graph Representing Research Findings by Semantifying Survey Articles.* In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 315-327. Springer, Cham, 2017. `DOI:10.1007/978-3-319-67008-9_25`

2. **Said Fathalla**, Sahar Vahdati, Christoph Lange, and Sören Auer. *Analysing Scholarly Communication Metadata of Computer Science Events.* In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 342-354. Springer, Cham, 2017. `DOI:10.1007/978-3-319-67008-9_27`

3. **Said Fathalla** and Christoph Lange. *EVENTS: A Dataset on The History of Top-Prestigious Events in Five Computer Science Communities.* In Proceedings of Semantics, Analytics, Visualization (SAVE-SD) at the World Wide Web conference, pp. 110-120. Springer, Cham, 2017. `DOI:10.1007/978-3-030-01379-0_8`

4. **Said Fathalla**, Sahar Vahdati, Sören Auer, and Christoph Lange. *SemSur: A Core Ontology for The Semantic Representation of Research Findings.* Procedia Computer Science 137 (SEMANTiCS), pp. 151-162, 2018. `DOI:10.1016/j.procs.2018.09.015`

5. **Said Fathalla**, and Christoph Lange. *EVENTSKG: A Knowledge Graph Representation for Top-Prestigious Computer Science Events Metadata.* In International Conference on Computational Collective Intelligence (ICCCI), pp. 53-63. Springer, Cham, 2018. `DOI:10.1007/978-3-319-98443-8_6`

6. **Said Fathalla**, Sahar Vahdati, Sören Auer, and Christoph Lange. *Metadata Analysis of Scholarly Events of Computer Science, Physics, Engineering, and Mathematics.* In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 116-128. Springer, Cham, 2018. `DOI:10.1007/978-3-030-00066-0_10`

7. **Said Fathalla**, Christoph Lange, and Sören Auer. *EVENTSKG: A 5-Star Dataset of Top-Ranked Events in Eight Computer Science Communities.* In the European Semantic Web Conference (ESWC), pp. 427-442. Springer, Cham, 2019. `DOI:10.1007/978-3-030-21348-0_28`

8. **Said Fathalla**, Sahar Vahdati, Christoph Lange and Sören Auer, *SEO: A Scientific Events Data Model.* In International Semantic Web Conference (ISWC), pp. 79-95, Springer, 2019. `DOI:10.1007/978-3-030-30796-7_6`

9. **Said Fathalla**, Christoph Lange, and Sören Auer. *A Human-friendly Query Generation Frontend for a Scientific Events Knowledge Graph.* In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 200-214. Springer, Cham, 2019. `DOI:10.1007/978-3-030-30760-8_18`

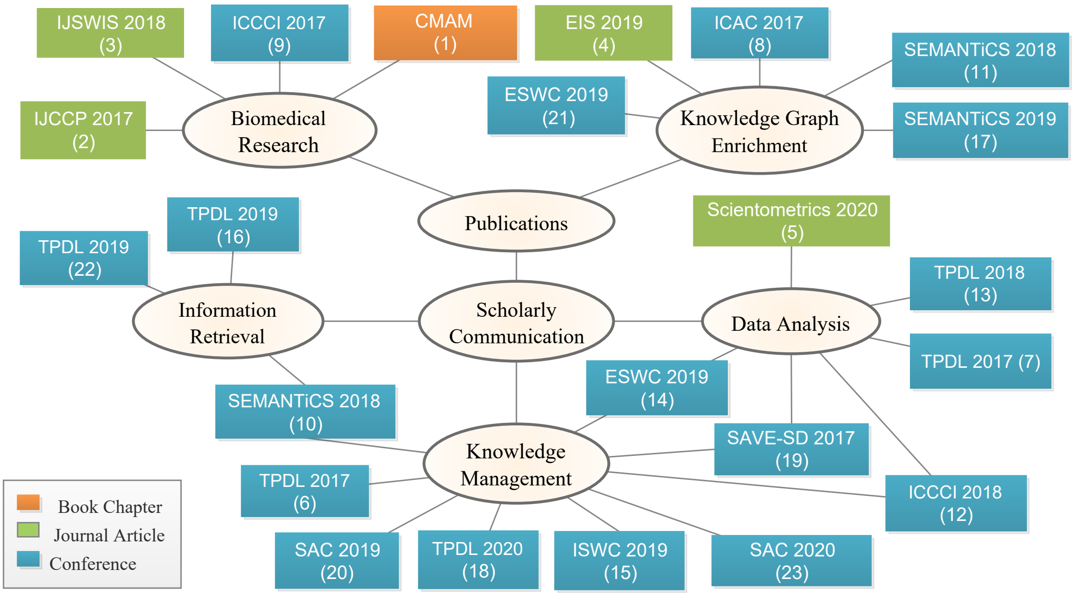


Figure 1.3: Distribution of publications across various research disciplines during PhD studies. Numbers between parentheses refer to that publication in the publications list in Appendix A.

10. **Said Fathalla**, Sahar Vahdati, Sören Auer, and Christoph Lange. *The Scientific Events Ontology of The Openresearch.org Curation Platform*. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (SAC), pp. 2311-2313. ACM, 2019. DOI:10.1145/3297280.3297631

11. Sahar Vahdati, **Said Fathalla**, Sören Auer, Christoph Lange, and Maria-Esther Vidal. *Semantic Representation of Scientific Publications*. In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 375-379. Springer, Cham, 2019. DOI: 10.1007/978-3-030-30760-8_37

12. **Said Fathalla**, Sahar Vahdati, Sören Auer, and Christoph Lange. *Scholarly Event Characteristics in Four Fields of Science: A Metrics-based Analysis*. Scientometrics 123 (2020): 677–705.DOI:10.1007/s11192-020-03391-y

13. **Said Fathalla**, Sören Auer, and Christoph Lange. *Towards The Semantic Formalization of Science: The Science Knowledge Graph Ontologies Suite*. In Proceedings of the 35th ACM/SIGAPP Symposium on Applied Computing (SAC), pp. 2057-2059, 2020. DOI: 10.1145/3341105.3374132

14. Zeynep Say, **Said Fathalla**, Sahar Vahdati, Jens Lehmann and Sören Auer. *Ontology Design for Pharmaceutical Research Outcomes*. In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 119-132. Springer, 2020. DOI:10.1007/978-3-030-54956-5_9

15. Aysegul Say, **Said Fathalla**, Sahar Vahdati, Jens Lehmann and Sören Auer. *Semantic Representation of Physics Research Data*. In 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, In press, SciTePress, 2020. DOI:10.5220/0010111000640075

1.5 Thesis Structure

In order to prepare the reader for the upcoming chapters, an overview of the thesis is presented. The thesis is structured in eight chapters, outlined as follows.

Chapter 1 is the introduction of the thesis in which the research problem, motivation, research questions, and challenges are presented. **Chapter 2** provides preliminary information about the terminologies, data models, tools, and technologies used in the work presented in this thesis. **Chapter 3** outlines the research efforts related to the research questions defined in this thesis. In **chapter 4**, we investigate the problem of study the various characteristics of scholarly events in different fields of science to assess their impact. **Chapter 5** presents our research efforts for publishing scholarly communication metadata as Linked Open Data. In **chapter 6**, we present SPARQL-AG, a front-end that automatically generates and executes SPARQL queries for querying EVENTSKG. **Chapter 7** presents a set of ontologies for modeling the research findings in various fields of science, resulting in a knowledge graph of the scientific findings in modern sciences. Finally, **chapter 8** concludes the thesis as well as provides directions for future research.

CHAPTER 2

Background

Scholarly communication is a system through which researchers create, peer-review, publish scholarly articles as well as preserve scholarly data for future use [33]. It mainly involves the publication of peer-reviewed articles through academic publication venues, involving scholarly events and journals. Digitization has supported the transition to digital publishing, which presents considerable new challenges and threats to access, explore, and analysis of scholarly information published on the Web. This digital environment also poses significant challenges for the long-term preservation of such information for further use as well as the rapid changes in technology platforms pose other serious preservation challenges, which is so-called *Digital Preservation* [34]. Recently, the digital preservation activities move forward to develop expeditiously, with fields of activities expanding, and best practices are still under debate [34]. Consequently, scholarly communication has encountered various challenges as a result of producing a large volume of heterogeneous scholarly artifacts from a day to another [35]. Currently, Semantic Web technologies are widely used to support scholarly communication, particularly, Semantic Publishing. Semantic Publishing involves developing semantic models for publishing data using Semantic Web technologies, such as RDF, thus improving information retrieval, and data analysis and integration [28].

This chapter provides preliminary information about the terminologies, data models, tools, and technologies used in the work presented in this thesis. Section 2.1 presents fundamental principles accompanied by scholarly communication as well as its life cycle. Section 2.2 gives a brief description of the primary standards s we used in this thesis. The underlying principles of scholarly communication are described later in this chapter.

2.1 Scholarly Communication

Scholarly communication involves the creation, dissemination, analyzing, and archiving scholarly articles aiming at making them available for the scientific community and preserving them for future use [36]. There are two ways of communication among scholars, formal and informal communication channels. Informal communication refers to scientific conversation and discussion, while formal communication refers to the publication in peer-reviewed journals and scholarly events [37]. In addition, scholarly communication encompasses activities, including conference presentations, informal seminar discussions, scientific meetings or telephone conversations, and grey literature [38]. Gradually these activities have been progressively boosted by new means of communications, such as emails, instant messaging spaces, mailing lists, and social media (e.g.,

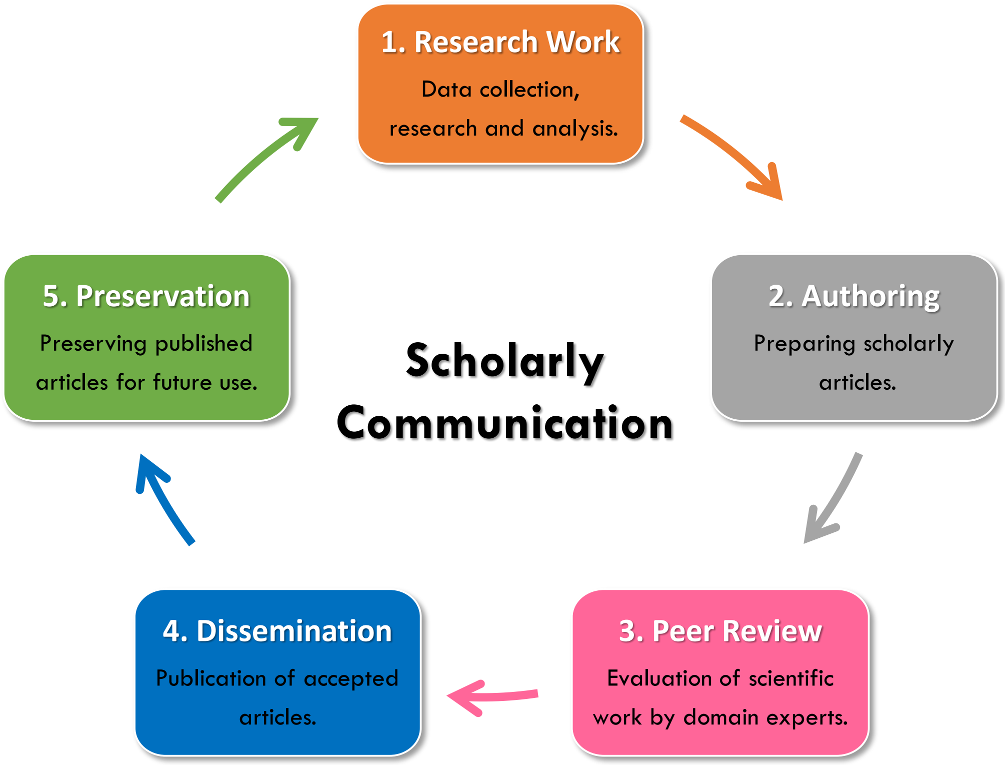

1. Research Work
Data collection, research and analysis.
2. Authoring
Preparing scholarly articles.
3. Peer Review
Evaluation of scientific work by domain experts.
4. Dissemination
Publication of accepted articles.
5. Preservation
Preserving published articles for future use.
Scholarly Communication

Figure 2.1: **Scholarly communication life cycle**. Scholarly communication involves several consequent stages, including research work, authoring, peer reviewing, dissemination, and preservation.

Twitter). In the past decades, how scholars communicate, write, and submit scholarly artifacts have changed a lot with the affordances of digitization. On the one hand, the Internet opens the way to new dissemination potentials, which in turn provide broader access to peer-reviewed publications, thus supporting collaborations between researchers across the world, i.e., it removes the physical barriers restrictions. On the other hand, scholarly knowledge publication gained new dissemination channels, such as digital journals, conferences, and workshops. Scholarly communication involves several consequent stages, frequently depicted as a life cycle documenting these stages, including research work, authoring, peer reviewing, dissemination, and preservation (see Figure 2.1).

Stakeholders. Various stakeholders appear at the numerous stages in the scholarly communication life cycle with different duties. According to [39], these stakeholders can be categorized into four major groups: 1) researchers (producers): who produce scholarly artifacts (i.e., *preprints*), 2) publishers: who publish scholarly artifacts (i.e., *publications*), 3) libraries: who preserve published artifacts for further uses, and 4) researchers (consumers): who explore libraries for interesting artifacts.

Scholarly Communication Crisis. Primarily, the crisis in scholarly communication has emerged by the growing expenses of publishing as well as offering open access to published scholarly articles. The following factors are the primary cause of the crisis in scholarly communication [40]: 1) the high subscription cost of scientific journals, 2) rising influence of scientific journals by the for-profit organizations, 3) libraries are facing economic challenges, which do not correspond to the rising costs of the journals, 4) the restrictions on the use of electronic publications due to its licenses, and 5) the orientation of open access publishing remains against the policies of commercial publishers.

2.1.1 Key Principles for Scholarly Communication

There are four fundamental principles accompanied by scholarly communication since the past century: registration (submission to scientific venues), peer-review (evaluation of the quality), dissemination (including distribution and access), and preservation (permanent archiving for further uses) [34].

Registration. The research work undertaken by individuals or groups of researchers is registered by submitting this work to any of the available publication channels. Many computer scientists prefer to submit their research papers to scholarly events for the reason that they 1) have a higher status, 2) provide higher visibility, 3) have a significant impact, 4) have short peer-review process compared to journals, and 5) have higher standards of novelty, whereas journals often only require 20%-40% of the research work to be new [41]. In some cases, journal publication is desirable because they offer more extended page limits, authors get more detailed reviews, the opportunity to revise and re-submit, and have, in most cases, slightly higher acceptance rates.

Peer-review. The peer-review process plays a vital role in scholarly publishing in which scientific work is evaluated and validated by research community experts (peers) [42]. It significantly improves the quality of the published papers as well as increase the effective communications within research communities, which is typically the key to professional success. Peer review can be either open or closed review. The former is not yet commonly used, but the latter is more commonly used which itself has three sub-types [43]:

- *Single blind review* – the identity of the reviewers are anonymous, i.e., concealed from the authors,
- *Double-blind review* – both reviewers and authors are anonymous, and
- *Triple-blind review* – both reviewers and authors are anonymous, as well as editors are also unknown to both.

In open review, both reviewers and authors know each other during the peer-review process, and, in some cases, the reviews are published after the review process.

Dissemination. Scholarly documents dissemination aims to distribute academic research, in the form of journal or conference article, book or thesis, making them accessible to research communities. Most of the researchers look for research papers published by high-quality journals and conferences, especially if they are free of charge [44]. The dissemination plan is usually based on the audience, which helps to determine the appropriate publication channel for submitting scholarly documents. For example, if the research is a pure research work, then the audience will likely be the academic community. Consequently, scientific journals and scholarly events are the most suitable channels for publication. While, if the research is community-based work, then social-media or blogs are more appropriate [44]. The ultimate objective of scholarly artifacts dissemination is to make it easier for researchers to explore scholarly work. Dissemination makes them easily reachable and openly accessible to various research communities, thus maximizing accessibility. Dissemination helps researchers to ensure that their work achieves broad dissemination, which helps to gain a good reputation and professional rewards. In the past decades, broad dissemination was achieved more convenient and cheaper via the Internet, thanks to digitization.

Preservation. Data preservation refers to the set of activities controlled by policies and guidelines required to guarantee persistent access to data for further uses [45]. The growing number of digital media available on the Web brings several preservation challenges, such as accessibility of files, and manageability of storage and backups effectively [46]. Nowadays, data preservation has become a vital factor in digital libraries due to the increasing number of digital objects produced in recent years. Data preservation systems should be able to address organizational issues, such as heterogeneity and complexity of data, technical aspects, such as dealing with large volumes of data and media monitoring and the data curation process [47]. The infrastructure of such systems should include policies, procedures, and practices to keep the information accessible to users over a long time.

2.1.2 Scholarly Publishing and Open Access

In the past decades, many changes have affected scholarly publishing, but the process itself has remained stable [34]. The number of scholarly documents is increasingly growing from day to day in the form of research papers, technical reports, and book reviews by researchers around the world [19]. In 2014, Khabsa and Giles [48] have estimated the number of scientific articles available on the Web through examining the coverage of two of the most prominent scholarly search engines, i.e., Google Scholar and Microsoft Academic search engines. They found that there are approximately 114 million scholarly documents written in English (of which Google Scholar has 88%) published on the Web in which 24% of them are available at no cost [48].

Semantic Publishing. The management of scholarly information published on the Web is an evolving field of research in the Semantic Web so-called Semantic Publishing [49]. Semantic Publishing paves the way for machines to understand the meaning of the data published on the Web, intending to make information retrieval, knowledge management, and data analysis more efficient. It overcomes the drawbacks of publishing scientific articles either as conventional HTML pages or PDF files, immolating its semantics and relationships to other data, thus restricting the reusability of the data, e.g., for knowledge discovery and data analysis. Although PDF documents are both accustomed and convenient for humans to read, they are negating the essence of the Semantic Web, which aims at enabling the machine to read and understand data, thus obstructing the development of intelligent services that can perform semantics-oriented tasks, such as knowledge discovery, linking and exploration [49]. Several publishers have already taken the lead in the use of semantic technologies in order to support scholarly publishing:

- *Public Library of Science (PLoS)*[1]: PLoS has taken serious steps towards providing semantic enhancements to journal articles. These enhancements involve the provision of live DOIs and interactive figures, and linking textual terms to related external resources [50].
- *Scientific, Technical and Medical publisher (STM)*[2]: Semantic enrichment of articles contents, e.g., the automatic extraction of the metadata of articles and linking of entities, is now widely used by STM for achieving high discoverability and reusability [51].
- *Springer Nature (SN)*[3]: SN has a published Linked Open Dataset aggregated from several data sources, such as Springer Nature. This dataset contains information about different entities involved in the scholarly domain, such as organizations, research projects, scholarly events, publishers, and publications.

[1] https://www.plos.org/

[2] https://www.stm-assoc.org/

[3] https://www.springernature.com/gp/researchers/scigraph

FAIR principles. The FAIR Data Principles (an acronym for Findability, Accessibility, Interoperability, and Reusability) are a concise set of principles developed to be a guideline for improving the reusability of scholarly data [17]. Scientific data is not only for human consumption, but it is also consumed by many systems and software agents to perform data retrieval and analysis on behalf of humans [17]. Therefore, FAIR does not focus only on human consumption, but it strengthens machine capabilities in order to automatically find, exchange, and use the data. To achieve this goal, FAIR proposes four main principles that modern data resources and infrastructures should expose to support the discovery and reuse of published data by third-parties [52]:

- *Findability* – metadata and data should be easily findable (i.e., registered in a searchable resource) by both humans and machines, thus supporting the automatic discovery of datasets and services,
- *Accessibility* – third-parties should be able to access required data; therefore there should be clarity and transparency around the conditions governing access,
- *Interoperability* – community agreed formats, language and vocabularies should be used to support data interoperability, and
- *Reusability* – metadata and data should be richly described and meet domain-relevant community standards so that it can be easily reused.

Open access. The significant increment in the publication fees intensely hindered the ability to obtain publications necessary for both research and industry. Open access (OA) publishing is a mechanism for tackling this dilemma by offering unrestricted online access to research articles, such as journal articles and books, with no access fees [53]. Therefore, research articles are available to anyone across the globe at any time. Therefore, it provides wider visibility of the published research, which results in increased readership and, thus, increasing the number of citations of this research. Open access is not just being able to access research articles without barrier, but it also allows researchers to *reuse* this work. Mainly, there are two roads for making research articles openly accessible, the green and gold roads [54]. *Gold open access* refers to publishing articles via the publisher's platform, which makes them instantly available upon publication. This type of open access is achieved after an article publishing charge (APC) is applied. *Green open access* refers to publishing articles either via non-commercial channels, such as university repository or the author's personal website, or self-archiving it in an open-access archive [54]. There is no access fee for authors who publish their work through green OA. The default license for the majority of scientific publishers, e.g., Springer Nature, is Creative Commons license (CC BY[4]), which allows for unrestricted reuse of the article, but the original source should be appropriately cited. Commercial publishers provide open access, such as BioMed Central (BMC)[5], who publish open access as well as subscription-based journals and nonprofit publisher, who aims at advancing the progress in science, such as the Public Library of Science (PLoS)[6].

Predatory publishing. The term "predatory journals" was coined by Jeffrey Beall, who created "Beall's list"[7] of predatory publishers [55]. There is no standard definition for predatory publishers, i.e., illegitimate publishers, but generally, they are charging authors for publishing their work for

[4] `https://creativecommons.org/licenses/by/4.0/`

[5] `https://www.biomedcentral.com/`

[6] `https://www.plos.org/`

[7] `https://en.wikipedia.org/wiki/Beall%27s_List`

gaining financial profit without providing publication services offered by legitimate publishers. These publishers fail to 1) meet scholarly publishing standards, such as providing a rigorous peer review process, 2) follow publication ethics, 3) plan a publishing schedule, and 4) ensure the long-term digital preservation of the published work. In contrast, legitimate publishers provide valuable services to protect published articles.

2.2 Semantic Web Technologies

Semantic Web Technologies are used to provide the meaning rather than the structure of the data by developing languages to express rich, linked, and machine-understandable data. That is what distinguishes Semantic Web Technologies from other data technologies, such as relational data models. Thus, machines are not only able to process data as a set of keywords but interconnected concepts. Semantic technologies do not refer to a single technology, but rather to a variety of tools and technologies used to create data stores on the Web, build vocabularies, and write rules for handling data [56]. The key Semantic Web Technologies are; the Resource Description Framework (RDF), the RDF Schema (RDFs), and Web Ontology Language (OWL) for data representation as well as SPARQL Protocol and RDF Query Language (SPARQL) for data querying. In the next subsections, a brief description of these technologies will be presented.

2.2.1 Linked Data

In 2006, the four design principles of Linked Data was introduced by Tim Berners-Lee [57]. The aim of publishing data on the Web as *Linked Data* is interlinking data from different sources [58]. Technically, Linked Data is a structured data, using vocabularies such as schema.org[8], published on the Web in a machine-readable format, i.e., can be easily interpreted by machines. Data can be linked using a Uniform Resource Identifier (URI), which points to other data in the same or other datasets. Linked data is empowered by Semantic Web technologies, such as RDF, RDFS, OWL, SPARQL, and SKOS [56]. In 2010, the term *Linked Open Data (LOD)* appears, which refers to linked data released under an open license, which does not inhibit its reuse [57]. A comprehensive overview of the datasets published as Linked Data can be found in the Linked Open Data Cloud[9].

Principles of Linked Data. Linked Data relies on two Internet standard protocols, which are the fundamentals of the Web: Uniform Resource Identifiers (URIs) [59] and the Hypertext Transfer Protocol (HTTP) [60]. The Resource Description Framework (RDF) data model is used for publishing structured data on the Web. The RDF data model is explained in more detail later in this chapter. According to [4], the main Linked Data principles are:

- *Naming entities with URIs* – Entities in Linked Data are identified by Uniform Resource Identifier (URI), which can be looked up by dereferencing the URI over the HTTP protocol. A URI can be a Uniform Resource Locator (URL), which is used to identify resources and properties, i.e., URIs are used to give a unique ID to each resource. URIs are very crucial for supporting data exchangeability across machines.
- *Making URIs defererenceable* – Entities can be looked up by dereferencing their URI over the HTTP protocol, which means that HTTP clients can look up the URIs (that are used

[8] https://schema.org/

[9] https://lod-cloud.net/

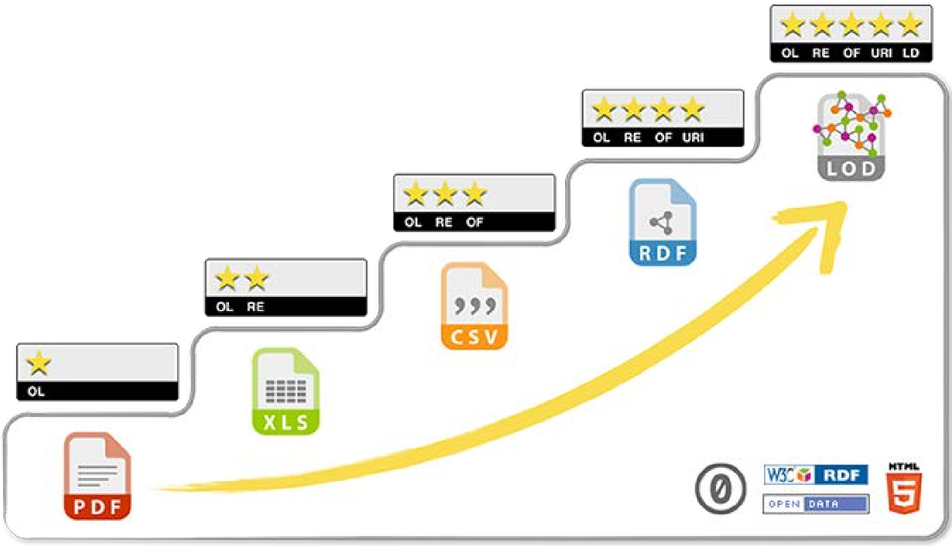


Figure 2.2: Five-star deployment scheme for Open Data proposed by Tim Berners-Lee [63].

in the Linked Data context) using the HTTP protocol and retrieve the description of entities identified by the URI. This can be achieved based on attributes of the retrieval request, an HTTP mechanism called *Content Negotiation* [60].

- *Searching using Semantic Web standards* – When looking up a URI, the client should get useful information, using the Semantic Web standards, such as RDF and SPARQL.
- *Pointing to other data* – One crucial principle of Linked Data is to set RDF links pointing to other data sources on the Web, thus allowing entities from different vocabularies to be connected to each other [61].

Five-star Deployment Scheme. Tim Berners-Lee introduced a 5-star deployment scheme for Open Data (illustrated in Figure 2.2), by which data publishers can award stars to their datasets according to the following criteria [62]:

- The *first star* can be awarded if the data is available on the Web (whatever format), but with an open license.
- The *second star* can be awarded if it is 1-star along with the data is available as machine-readable structured data (e.g., Excel).
- The *third star* can be awarded if it is 2-star along with the data is available in a non-proprietary open format (e.g., CSV instead of Excel).
- The *fourth star* can be awarded if it is 3-star along with the URIs are used to denote things so that people can point at this data.
- The *fifth star* can be awarded if it is 4-star along with the data is linked to other data to provide context.

2.2.2 The Resource Description Framework

The Resource Description Framework (RDF) is a data model developed by the World Wide Web Consortium (W3C)[10] for interchanging, encoding, and reusing of metadata on the Web

[10] https://www.w3.org/

in a structured format [64]. The fundamental concepts of RDF are resources, properties, and statements [65]. RDF is the minimum semantic representation of data on the Web, expressing the relationship among individuals via predicates.

Resources are objects in the domain of interest, which can be persons, books, events, publishers, locations, hotels, jobs, etc. *Properties* (known as predicates) are a particular type of resource; they also identified by URIs, which represent relationships between resources, e.g., "given name", "work in", "homepage", and so on. RDF can only represent binary predicates [65]. *Statements* (known as triples) describe metadata about a resource in RDF documents. In other words, an RDF statement comprises three parts; a resource (i.e., subject), property (i.e., predicate), and the value of this property (i.e., object). The URI of the resource *Book* can be either a complete URL, e.g., *http://www.SaidThesis.org/Book*, or a prefix name (a unique string associated for each namespace) along with the resource name, e.g., *ex:Book*, where *ex* refers to the base URI which is in this case "*http://www.SaidThesis.org/*". One of the prominent services that used by RDF developers to look up prefixes is `http://prefix.cc`. All prefixes in this thesis are consistent with prefix.cc. The purpose of the XML namespaces URIs in RDF is to organize terms of a particular vocabulary into logical groups and also used to avoid name conflicts between properties/resources of the same name, i.e., two resources with the same name should have different namespaces, enabling RDF parsers to differentiate between them even though they have the same name. An RDF triple consists of three parts: subjects, predicates, and objects, where the subject is the resource, the predicate represents the relationship between the subject and the object, and objects can be either resources or literals, i.e., atomic values, e.g., numbers. Literals appear only in the object position of an RDF statement. The combination of RDF triples forms a directed graph. Vertices in this graph can be either subjects or objects, while edges are the predicates, i.e., the relationships connecting resources.

As a logical formula, triples can be represented as $P(s, o)$, where the binary predicate P relates the subject s to the object o. When a resource has no URI or literal is not given RDF documents can have *blank nodes*, which represent resources that their URIs are not given. An RDF triple is formally defined as follows:

Definition 2.1: RDF Triple [66]

Let $\mathbf{I}$, $\mathbf{B}$, $\mathbf{L}$ be disjoint infinite sets of URIs, blank nodes, and literals, respectively. A tuple $(s, p, o) \in (\mathbf{I} \cup \mathbf{B}) \times \mathbf{I} \times (\mathbf{I} \cup \mathbf{B} \cup \mathbf{L})$ is denominated an RDF triple, where s is called the subject, p the predicate, and o the object.

In 2014, W3C proposed the RDF 1.1 concepts and abstract syntax [67]. In RDF 1.1, resources are identified by IRIs rather than URIs. Besides, RDF 1.1 introduces the concept of RDF Datasets [68], which can be a collection of RDF Graphs. RDF data models can be validated using W3C's RDF Validation Service[11]. The purpose of the online RDF Validator is to parse RDF documents to detect syntax errors and visualize the document either by tabular or graph-based views. To better understand how RDF can represent information, let us consider a simple example. Consider the sentence: "*The Egyptian student Said Fathalla is studying Computer Science at the University of Bonn*". First, resources, properties, and statements should be identified. In this case, the resources are *Said Fathalla*, *Computer Science*, and the *University of Bonn*, while the statements are:

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

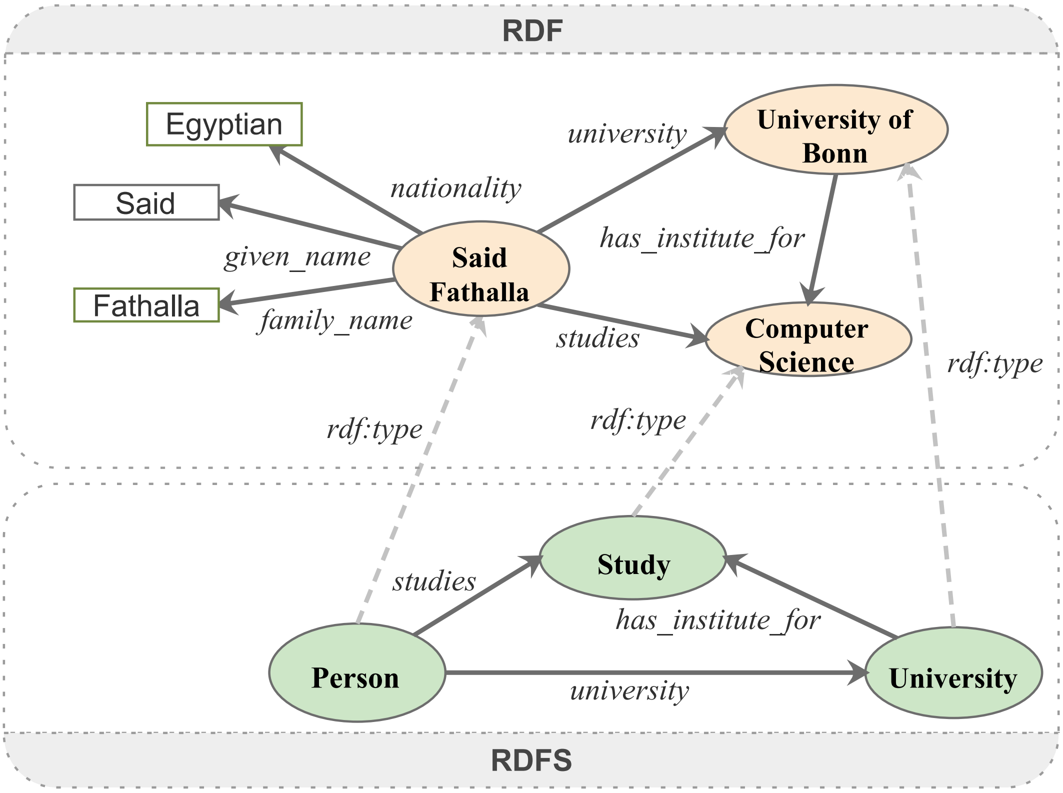


Figure 2.3: An RDF graph representing statements S1–S6 and its RDF Schema.

(S1) the given name of Said Fathalla is Said,
(S2) the family name of Said Fathalla is Fathalla,
(S3) the nationality of Said Fathalla is Egyptian,
(S4) the university of Said Fathalla is the University of Bonn,
(S5) Said studies Computer Science, and
(S6) the University of Bonn has an institute for Computer Science.

A single RDF triple can represent each of these statements. The upper part of Figure 2.3 depicts the RDF graph representing this statement in which ovals represent resources, arrows represent predicates, and rectangles represent literals.

RDF Serializations

RDF has no single syntax, but it can be expressed in various formats, called "Serializations". There are many serializations available for representing RDF documents, such as RDF/XML, Turtle, N-Triples, JSON-LD, TriG, and more. In the following, we list some of the most well-known serializations:

RDF/XML Serialization. The most popular format is the XML-based format, known as "RDF/XML syntax". This kind of representation of data models eases the information exchange between various applications running on different types of operating systems. The RDF/XML document is a tree-like document, in which the root node (Root Tag) of an RDF/XML document is `rdf:RDF`, which is followed by a set of namespaces as attributes (cf. Listing 2.1). Each resource is identified by an `rdf:Description` element with the `rdf:about` attribute, which can contain several statements about the same subject. As shown in Listing 2.1, the resource “*http://www.SaidThesis.org/SaidFathalla*” contains several statements describing the resource involving the given name (`foaf:givenName`), the family name (`foaf:familyName`) and the nationality (`st:nationality`). The properties `foaf:familyName`, and `foaf:givenName` are defined in the foaf (http://xmlns.com/foaf/0.1/) namespace. Properties of resources can be defined either as attributes (e.g., *st:nationality="Egyptian"*) or as resource. The latter can be represented in RDF/XML as follows:

```
<st:university rdf:resource="http://www.SaidThesis.org/UniversityOfBonn"/>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#Literal#"
xmlns:st="http://www.SaidThesis.org/"
xmlns:foaf="http://xmlns.com/foaf/0.1/">

<rdf:Description rdf:about="http://www.SaidThesis.org/ComputerScience">
  <rdf:type rdf:resource="http://www.SaidThesis.org/FieldOfStudy"/>
</rdf:Description>

<rdf:Description rdf:about="http://www.SaidThesis.org/SaidFathalla">
  <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
  <st:studies rdf:resource="http://www.SaidThesis.org/ComputerScience"/>
  <st:university rdf:resource="http://www.SaidThesis.org/UniversityOfBonn"/>
  <st:nationality rdf:datatype="xsd:string">Egyptian</st:nationality>
  <foaf:familyName rdf:datatype="xsd:string">Fathalla</foaf:familyName>
  <foaf:givenName rdf:datatype="xsd:string">Said</foaf:givenName>
</rdf:Description>

<rdf:Description rdf:about="http://www.SaidThesis.org/UniversityOfBonn">
  <rdf:type rdf:resource="http://www.SaidThesis.org/University"/>
  <st:has_institute_for rdf:resource="http://www.SaidThesis.org/ComputerScience"/>
</rdf:Description>
</rdf:RDF>
```

Listing 2.1: Representation of the example depicted in Figure 2.3 using the RDF/XML serialization.

N-Triples Serialization. RDF N-Triples [69] is a line-based syntax for an RDF graph in which each statement (i.e., RDF triple) appears in a separate line in the RDF document, with a “.nt” file extension. Every simple triple (or triple) is a series of subject, predicate, and object separated by white-space and terminated by ”.. An N-Triples document is a sequence of RDF triples terminated by a ’ . ’ and contains no parsing directives. Considering the example, in Listing 2.1), the following triples represent two RDF triples about the subject *http://www.SaidThesis.org/SaidFathalla*:

```
<http://www.SaidThesis.org/SaidFathalla> <http://www.SaidThesis.org/studies>
<http://www.SaidThesis.org/ComputerScience> .
<http://www.SaidThesis.org/SaidFathalla> <http://www.SaidThesis.org/university>
<http://www.SaidThesis.org/UniversityOfBonn> .
```

Turtle Serialization. Turtle (Terse RDF Triple Language)[70] is an RDF serialization that allows for representing RDF graphs in a compact and natural text form. Turtle format (with the file extension “.ttl”) is much easier to understand for humans, whereas multiple triples with the same subject can be grouped under the same URI, i.e., the URI of the subject is not repeated for each triple. This feature is called *syntactic sugar* [71]. In this case, a series of predicates and objects (separated by ’;’) is repeated beyond the subject of the triple and terminated by ”.. Listing 2.2 shows the Turtle serialization of the example depicted in Figure 2.3.

```
@prefix st: <http://www.SaidThesis.org/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<http://www.SaidThesis.org/ComputerScience> a <http://www.SaidThesis.org/FieldOfStudy> .
<http://www.SaidThesis.org/SaidFathalla>
    a foaf:Person ;
    st:studies st:ComputerScience ;
    st:university st:UniversityOfBonn ;
    st:nationality "Egyptian"^^rdfs:Literal ;
    foaf:familyName "Fathalla"^^rdfs:Literal ;
    foaf:givenName "Said"^^rdfs:Literal .

st:UniversityOfBonn
    a st:University ;
    st:asInstituteFor st:ComputerScience .
```

Listing 2.2: Representation of the example depicted in Figure 2.3 using the turtle serialization.

JSON-LD Serialization. JSON-LD (JavaScript Object Notation for Linked Data) [72] is a lightweight serialization for representing Linked Data. It is based on JSON format[12], a data format that uses human-readable text for *attribute–value* pairs data objects transmission, which allows for a large number of JSON parsers to be reused for parsing also JSON-LD documents. The primary intention behind developing JSON-LD is to facilitate the use of Linked Data in Web-based applications for the purpose of creating interoperable Web services. In JSON-LD documents, the JSON object is represented as a sequence of attribute-value pairs (separated by a single comma ',') surrounded by a couple of curly brackets. A single colon separates the attribute from its value. Listing 2.3 shows the JSON-LD serialization of the example depicted in Figure 2.1.

2.2.3 Ontologies

In the past, machines were carrying out tasks instead of humans, with non-sort of intelligence. Therefore, to increase the benefits from them to the maximum, there is a pressing demand to make computers as intelligent as humans [73]. The shortcomings of non-semantic applications are sketched by the statement "*lack of semantics*", especially when the talk is regarding information retrieval and data management [9]. The notion of the ontology originally comes from Philosophy, which is the branch of metaphysics that studies the nature of being [74]. In the field of Computer Science, an ontology is defined by Gruber as "*the explicit specification of conceptualization*" [75]. Formally, an ontology can be defined as follows:

Definition 2.2: Ontology [76]

Let C be a conceptualization, and L a logical language with vocabulary V and ontological commitment K. An ontology O_K for C with vocabulary V and ontological commitment K is a logical theory consisting of a set of formulas of L, designed so that the set of its models approximates as well as possible the set of intended models of L according to K.

[12] https://json.org/

Ontologies define the concepts and the interrelationships between them in a particular domain of discourse. Ontologies store data in hierarchies and help to infer new information based on data and relationships among them. In the past decade, ontologies played a crucial role in data representation, whereas their success lies in defining vocabularies in a semantic format, thus enabling explicit specification of the semantics of the defined concepts and the relationships between them. Ontologies benefit enterprise applications in different areas of computer science, such as information retrieval [77], text classification [78], scholarly communication [79, 80] and bioinformatics [81]. For instance, in the field of health informatics, ontologies are used for representing and organizing medical vocabularies.

Since RDF capabilities are limited to the creation of statements about resources, i.e., a method of conceptual data modeling, therefore it is not capable of describing groups of related resources and the relations among them, i.e., classes, sub-classes, and properties. Consequently, another data modeling schema should be used to describe the formal semantics of RDF resources and interlinking relationships, enabling intelligent agents to understand and process this information in an efficient manner. In the following, the two most popular schema languages for creating ontologies will be described.

```
{
  "@context": {
    "foaf": "http://xmlns.com/foaf/0.1/",
    "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
    "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
    "st": "http://www.SaidThesis.org/",
    "xsd": "http://www.w3.org/2001/XMLSchema#"
  },
  "@graph": [
  {
    "@id": "st:SaidFathalla",
    "@type": "foaf:Person",
    "foaf:familyName": {
      "@type": "rdfs:Literal",
      "@value": "Fathalla"
    },
    "foaf:givenName": {
      "@type": "rdfs:Literal",
      "@value": "Said"
    },
    "st:nationality": {
      "@type": "rdfs:Literal",
      "@value": "Egyptian"
    },
    "st:studies": { "@id": "st:ComputerScience" },
    "st:university": {  "@id": "st:UniversityOfBonn" }
  },
  {
    "@id": "st:UniversityOfBonn",
    "@type": "st:University",
    "hasInstituteFor": { "@id": "st:ComputerScience"  }
  },
  {
    "@id": "st:ComputerScience",
    "@type": "st:FieldOfStudy"
  }
  ]
}
```

Listing 2.3: Representation of the example depicted in Figure 2.3 using the JSON-LD serialization.

RDF Schema

RDF Schema (RDFS)[82] is an extension of RDF, which provides a framework to define the vocabulary for describing RDF data. RDFS is a data-modeling vocabulary for RDF data, i.e., it is a semantic extension of RDF. RDFS offers taxonomic (subclass) relations, which build the class hierarchy of the ontology, and the properties as well. In RDFS, resources are grouped into groups (type or category) called *classes*, which are defined using `rdfs:Class`. For example, a class A can be a specialization of another more general class B, in this case, A is said to be a subclass of B and can be defined using `rdfs:subClassOf` property. The members of a class, known as instances/individuals of the class, are defined using `rdf:type`. Instances of a class are also instances of all its superclasses. For example, consider the case: A is a subclass of B, and C is a subclass of A, if c is an instance of C, then it is also an instance of A and B. Properties of classes are defined using `rdf:Property`. The possible values of a property are restricted by defining the *range* of that property using `rdfs:range`. The range(s) of a property can be one or more classes. In other words, when defining the range of a property to be a particular class, that means that the values of that property are instances of that class. The *domain(s)* of a property (defined using `rdfs:domain`) asserts that any resource that has a given property is an instance of one or more classes. A single property can have several domains or range statements. RDF Schema offers a way to specialize properties, i.e., defining sub-properties. This can be defined using `rdfs:subPropertyOf`. For example, *ex:hasFather* can be defined as a sub-property of *ex:hasParent*. Actually, `rdfs:subPropertyOf` property is both reflexive and transitive. A human-readable version of the resource name is given via `rdfs:label`, therefore the `rdfs:range` of `rdfs:label` is `rdfs:Literal` [73]. Multilingual labels can be given to the same resource using the language tag. For example, the English and German translations of the resource *Policeman* can be reprsented as `:policeman rdfs:label "Policeman"@en ; rdfs:label "Polizist"@de`. A human-readable description of a resource can be given via `rdfs:comment`. The lower part of Figure 2.3 depicts the RDFS for the RDF document of the example mentioned above in subsection 2.2.2.

RDFS Reasoning Capabilities. The purpose of performing logical inference is to discover hidden knowledge that can be made explicit, i.e., automatically derived instead of being made by hand. One deduction that can be made using RDFS is the deduction of a class membership from the range or domain of one of its properties. For instance, consider the following triples:

```
ex:studies          rdfs:domain         ex:Person .
ex:studies          rdfs:range          ex:Study .
ex:Said             ex:studies          ex:ComputerScience .
```

It can be inferred that

```
ex:Said             rdf:type            ex:Person .
ex:ComputerScience  rdf:type            ex:Study .
```

Another deduction that can be made using RDFS is the deduction of entity-superclass membership from a class hierarchy. For instance, consider the following triples:

```
ex:PhDStudent       rdfs:subClassOf     ex:Student .
ex:Student          rdfs:subClassOf     ex:Person .
ex:Said             rdf:type            ex:PhDStudent .
```

It can be inferred that

```
ex:Said             rdf:type            ex:Person.
ex:Said             rdf:type            ex:Student.
```

Similarly, an entity super property membership from a property hierarchy can be deduced. For instance, consider the following triples:

```
ex:studyComputerScience    rdfs:subPropertyOf       ex:studies .
ex:Said                    ex:studyComputerScience  ex:Databases .
```

It can be inferred that

```
ex:Said         ex:studies           ex:Databases .
```

The Web Ontology Language

The Web Ontology Language (OWL)[83] is a language that is used to formalize a domain by defining classes and relations between them. Mainly, it is a language for defining and instantiating Web ontologies. OWL offers several use cases for the semantic representation of the data on the Web that would require much more expressiveness than RDF and RDF Schema, i.e., it is richer and more expressive than RDFS. Using OWL, it is possible to specify properties cardinalities as well as define logical operators in the domain and range restrictions, e.g., the union of classes can be used, for example, when defining the range of a property. These capabilities give OWL its richness. The OWL formal semantics defines how logical consequences from the data exist in OWL ontologies can be derived, i.e., facts are not explicitly provided in the ontology but entailed by the semantics. This derivation can be achieved within a single ontology or multiple distributed ontologies that have been combined using defined OWL mechanisms [84]. Formal semantics enables reasoning about the knowledge provided in the ontology. The fundamental modeling primitives of RDFS provides only the ability to organize the concepts in typed hierarchies (for classes and properties), domain and range restrictions, and instances of classes. On the other hand, OWL provides more capabilities than RDFS, which provides a way to support reasoning about classes and properties. Some of these capabilities is listed below [84]:

- *Disjointness of classes*: two or more classes can be disjoint (using `owl:disjointWith`), for example, *Father* and *Mother* are disjoint classes,
- *Boolean combinations of classes*: classes can be combined using set operations, such as union (`owl:unionOf`), intersection (`owl:intersectionOf`), and complement (`owl:complementOf`),
- *Express equivalences*: two classes can be identified as equivalent, i.e., precisely have the same instances, using `owl:equivalentClass` and to specify that two individuals are identical, `owl:sameAs` is used,
- *Cardinality restrictions*: to place restrictions on the values of a property, several property restrictions are used, such as `owl:someValuesFrom`, `owl:allValuesFrom`, `owl:cardinality`, and `owl:hasValue`, and
- *Characteristics of properties*: OWL permits specifying several property characteristics, including transitive, symmetric, functional, and inverse.

Some of the possible reasoning outcomes supported by OWL are listed as follows [73]:

- *Class membership*: it can be inferred that the instance of a particular class is also an instance of all its superclasses,
- *Equivalence of classes*: assume that there are three classes C1, C2, and C3. Given that $(C1 \equiv C2) \land (C2 \equiv C3)$, then it can be inferred that $C1 \equiv C3$,

- *Inconsistency detection*: a possible inconsistency that can occur in the ontology is that two disjoint classes share an instance, and
- *Classification*: if an individual x satisfies a certain property-value pair for a class A, then it can be inferred that x is an instance of A.

OWL Reasoning Capabilities. OWL offers the RDF/RDFS reasoning capabilities in addition to the following capabilities.
Transitive properties: Formally, if a property, P, is specified as transitive then:

$$\forall x, y, z,\ P(x, y) \land P(y, z) \to P(x, z)$$

For example, the property *locatedIn* is transitive and can be represented using OWL as follows:

```
<owl:ObjectProperty rdf:ID="locatedIn">
    <rdf:type rdf:resource="&owl;TransitiveProperty" />
    <rdfs:domain rdf:resource="#Place" />
    <rdfs:range rdf:resource="#Region" />
</owl:ObjectProperty>
<Region rdf:ID="Poppelsdorf">
    <locatedIn rdf:resource="#Bonn" />
</Region>
<Region rdf:ID="Bonn">
    <locatedIn rdf:resource="#Germany" />
</Region>
```

Given the previous data, it can be inferred that

```
<Region rdf:ID="PopplesdorfPlatz">
    <locatedIn rdf:resource="#Germany" />
</Region>
```

Inverse Property: if a property P is tagged as an inverse property of Q, then:

$$\forall x, y,\ P(x, y) \iff Q(y, x)$$

Symmetric Properties. The property has itself as an inverse. Formally, if a property P is tagged as symmetric then:

$$\forall x, y,\ P(x, y) \iff P(y, x)$$

For example, the property *brotherOf* is symmetric, where if x is *brotherOf* y, then it can be inferred that y is *brotherOf* x.
Functional Property: if P is a functional property, then for any given individual, the property can have at most one value. Formally, if a property P is characterized as functional then:

$$\forall x, y, z,\ P(x, y) \land P(x, z) \to y = z$$

Inheritance of disjointness constraints: if two classes A and B are tagged a disjoint and a class C is a subclass of B, then it can be inferred that C is also disjoint with A. Formally, this can be represented as:

$$\forall A, B, C,\ disjoint(A, B) \land subClassOf(B, C) \to disjoint(A, c)$$

2.2.4 Knowledge Graphs

A knowledge graph (KG) can be defined as a network of a set of interlinked entities, including persons, events, objects, or abstract concepts that cover various topical domains [85]. This representation of data allows both humans and machines to process information efficiently and unambiguously. Graph-based representation of data, via knowledge graphs, plays a significant role in modern knowledge management. The prominent feature of knowledge-based representation of data is the support of knowledge sharing, reuse, and retention. In knowledge graphs, entities, e.g., Student, are the nodes of the graph and connected by relations, which are the edges of the graph, e.g., study. The sets of possible types and relations are organized in a schema or ontology, i.e., the terminological component (TBox), which defines their interrelations and restrictions of their usage [86], while the main focus is on the actual instances, i.e., the assertion component (ABox). The key characteristics of knowledge graphs are:

- data semantics – the meanings of the data are encoded so that all information is self-descriptive,
- logic formalization – allows to derive new information, enforce consistency and automatic analysis, such as generic entity disambiguation,
- scalability – the data can continuously grow by extending its scope, allows dynamic updates, and data governance.

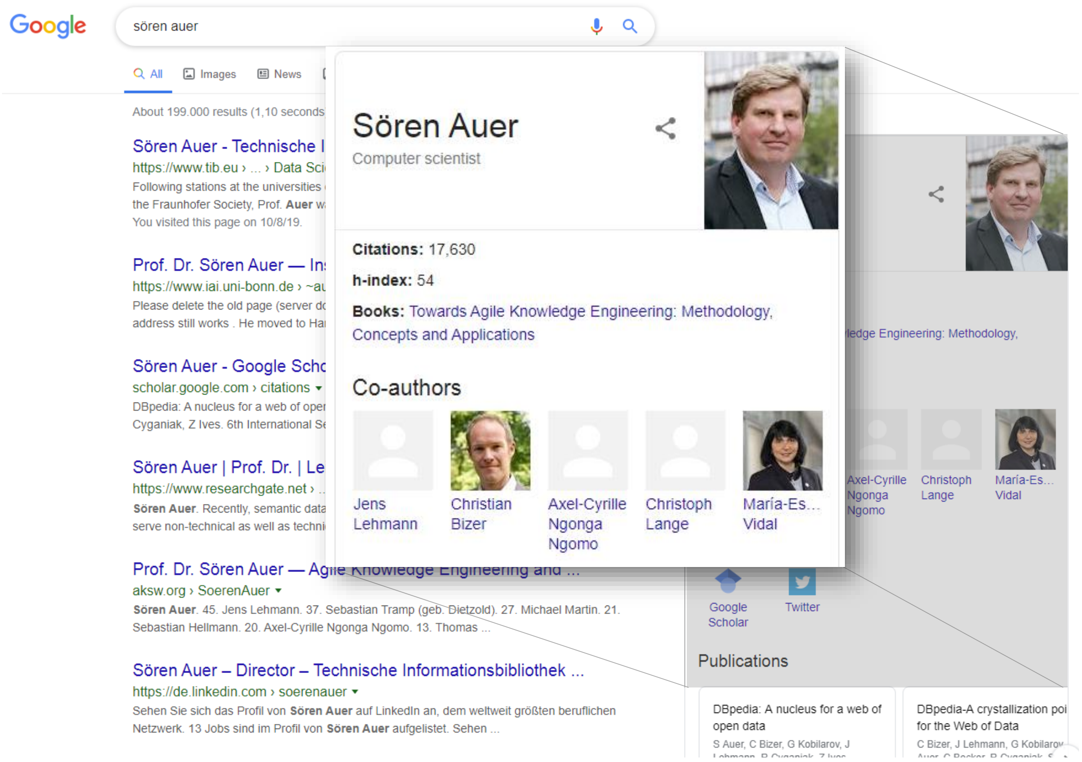
Well-known Knowledge graphs. Knowledge graphs are widely used in industry (e.g., Facebook); on the Web (e.g., Freebase [87], DBpedia [88], Wikidata [89]); and academia (e.g., YAGO2 [90] and Bio2RDF [91]) for several purposes.

Google's Knowledge Graph. In 2012, Google introduced Google's Knowledge Graph [92] to improve its search engine by allowing it to recognize facts about people, places, and things and interconnected relations. The primary objective is to provide popular facts alongside Google's traditional results, which opens the way of searching not only for pages that match query keywords, i.e., keyword-based search but rather for concepts, i.e., Semantic Search. This kind of search strategy brings new smarts to search capabilities. With this Knowledge Graph, Google can better understand user-submitted queries and can pick out the key facts for each object that most related to that object, then summarizes relevant content around that object. For example, if the query contains the computer scientist "Sören Auer", the results will include some key facts about him, such as citations, h-index, books, and co-authors (see Figure 2.4).

YAGO. Yago is a multilingual knowledge base created based information combined from Wikipedia, WordNet [93], and Geonames [13]. The initial version of YAGO extracted facts primarily from the category names of the English Wikipedia [94]. Currently, YAGO3 contains about 4.5 million entities and +8 million multilingual facts from 10 languages [95]. A Wikidata extractor has been developed in order to extract Wikidata from Wikipedia article, mainly from infoboxes, and produce a theme Dictionary of facts [95].

Wikidata. It is a free and open knowledge base of Wikipedia that enables Wikipedia to manage its data on a global scale. Wikidata contains multilingual data, where labels, such as *Book cover (Q1125338)* has the German label *Bucheinband*, are translated into several languages, while Wikipedia has independent editions for each language [89]. The content of Wikidata is licensed under CC0 1.0 Universal[14] and can be exported using standard formats, such as RDF, JSON and XML.[15]

[13] http://geonames.org/
[14] https://creativecommons.org/publicdomain/zero/1.0/
[15] https://www.wikidata.org



About 199.000 results (1,10 seconds

Sören Auer - Technische I
https://www.tib.eu › ... › Data Sci
Following stations at the universities
the Fraunhofer Society, Prof. **Auer** w
You visited this page on 10/8/19.

Prof. Dr. Sören Auer — In
https://www.iai.uni-bonn.de › ~au
Please delete the old page (server do
address still works . He moved to Ha

Sören Auer - Google Scho
scholar.google.com › citations ▾
DBpedia: A nucleus for a web of ope
Cyganiak, Z Ives. 6th International Se

Sören Auer | Prof. Dr. | Le
https://www.researchgate.net › ..
Sören Auer. Recently, semantic data
serve non-technical as well as techni

Prof. Dr. Sören Auer — Agile Knowledge Engineering and ...
aksw.org › SoerenAuer ▾
Sören Auer. 45. Jens Lehmann. 37. Sebastian Tramp (geb. Dietzold). 27. Michael Martin. 21. Sebastian Hellmann. 20. Axel-Cyrille Ngonga Ngomo. 13. Thomas ...

Sören Auer – Director – Technische Informationsbibliothek ...
https://de.linkedin.com › soerenauer ▾
Sehen Sie sich das Profil von **Sören Auer** auf LinkedIn an, dem weltweit größten beruflichen Netzwerk. 13 Jobs sind im Profil von **Sören Auer** aufgelistet. Sehen ...

Sören Auer
Computer scientist

Citations: 17,630

h-index: 54

Books: Towards Agile Knowledge Engineering: Methodology, Concepts and Applications

Co-authors

Jens Lehmann
Christian Bizer
Axel-Cyrille Ngonga Ngomo
Christoph Lange
María-Es... Vidal

Publications

DBpedia: A nucleus for a web of open data
S Auer, C Bizer, G Kobilarov, J

DBpedia-A crystallization poi for the Web of Data
C Bizer, J Lehmann, G Kobilarov

Figure 2.4: The outcome of the use of knowledge graphs in Google search, key facts about the computer scientist “Sören Auer”.

2.2.5 Querying Semantic Data

As Structured Query Language (SQL) is used to manipulate relational databases; the SPARQL Query Language [96] is used to query RDF data. SPARQL is the standard language to query data that follow the RDF specification of the W3C. Formally, a SPARQL query is defined as a tuple (E, DD, F), where E is a SPARQL algebra expression [97], DD is the dataset definition of the RDF dataset being queried, and F is a query form [96]. SPARQL use graph pattern matching to build result sets or RDF graphs [96]. Depending on the objective, SPARQL queries can be written in four forms [96]:

- *SELECT*: is used to query RDF graphs and return the results (after variable bindings) that match of the query graph pattern in a tabular format,
- *CONSTRUCT*: is used to construct an RDF graph specified by a graph template by combining the matched triples into a single RDF graph using union operator,
- *ASK*: is used to provide a Boolean result which indicates whether the query graph pattern matches or not, and
- *DESCRIBE*: is used to describe the resources found by whether they are identified by IRI or by a blank node.

Each of these query forms can have a WHERE clause (it is optional only for DESCRIBE), which contains restrictions on the returned results (also called *solutions*). The main components that formulate the SPARQL query are:

1. *Prefix declarations*: specify the set of namespace prefixes used in the query,
2. *Dataset definition*: specifies the dataset being queried,
3. *Result clause*: identifies what information to return from the query,
4. *Query patterns*: specify the query's graph pattern that matches the data, such as `UNION`, `MINUS`, `FILTER`, and `OPTIONAL`, and
5. *Query modifiers*: specify the set of modifiers for the query results, such as order, projection, distinct, offset, and limit.

Most of the forms of the SPARQL query contain a set of patterns called *basic graph pattern*, which is a set of triple patterns that query results must match. Formally, a triple pattern and a basic graph Pattern can be defined as shown in Definition 2.3.

Definition 2.3: Triple Pattern and Basic Graph Pattern [98]

Let $U, B, andL$ be disjoint infinite sets of URIs, blank nodes, and literals, respectively. Let V be a set of variables such that $V \cap (U \cup B \cup L) = \theta$. A triple pattern *tp* is a member of the set $(U \cup V) \times (U \cup V) \times (U \cup L \cup V)$. Let $tp_1, tp_2, \ldots, tp_n$ be triple patterns. A Basic Graph Pattern (BGP) B is the conjunction of triple patterns, i.e., $B = tp_1 AND tp_2 AND \ldots AND tp_n$

Listing 2.4 illustrates a set of graph patterns of one basic graph pattern of a simple SPARQL query. The basic graph pattern in this example consists of two triple patterns with two variables (`?name` and `?homepage`) in the object position of each triple. The result of this query is a solution sequence (can be zero, one, or multiple solutions) based on the match between the query's graph pattern and the data. The DISTINCT modifier is used to eliminate duplicate solutions.

```
PREFIX foaf:    <http://xmlns.com/foaf/0.1/>
SELECT DISTINCT ?name ?homepage
WHERE  {
  ?x foaf:name      ?name .
  ?x foaf:homepage  ?homepage .
}
```

Listing 2.4: A SPARQL query contains a group graph pattern of one basic graph pattern for getting the home page of all persons in the dataset who has a name.

Matching RDF Literals. RDF literals, such as strings and numeric types, are used in the triple pattern in the value position in the triple. Listing 2.5 depicts how query results can be restricted using RDF literals. A more expressive way to match RDF literals is the use of `FILTER`. For example, the values of strings can be restricted using `FILTER` functions, such as *regex*. The job of *regex* is to match the lexical forms of literals by using the *str* function. For example, to perform made case-insensitive string match, the "*i*" flag is used (restricting *foaf:familyName* in Listing 2.5. SPARQL FILTER can also be used to restrict on arithmetic expressions (restricting *foaf:age* in Listing 2.5).

```
PREFIX foaf:<http://xmlns.com/foaf/0.1/>
SELECT  ?name ?age
WHERE {
  ?x   foaf:familyName   ?name .  FILTER regex(?name, "fathalla", "i" )
  ?x   foaf:age          ?age  .  FILTER (?ag > 30)
  ?x   foaf:gender       "male".
  ?x   foaf:name         ?name .  OPTIONAL { ?x  foaf:mbox  ?mbox }
}
```

Listing 2.5: A SPARQL query that shows how results can be restricted FILTER.

Optional Pattern Matching. In an optional match, whether the optional graph pattern matches a triple or not, the triple is retrieved. For example, the query in Listing 2.5 retrieves all triples, whether containing the predicate *mbox* with the same subject or not. A graph pattern may have zero or more optional graph patterns [96].

Combining graph patterns. Several alternative graph patterns can be combined so that one of them may match. When several alternatives match, then all matched solutions are retrieved. This can be achieved using the `UNION` keyword. For example, the query in Listing 2.6 is used to retrieve all persons, whether the name is recorded using *vCard:FN* or *foaf:name*.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX vCard: <http://www.w3.org/2001/vcard-rdf/3.0#>
SELECT ?name
WHERE
{
  { ?x foaf:name ?name } UNION { ?x vCard:FN ?name }
}
```

Listing 2.6: A SPARQL query shows how graph patterns can be combined using the `UNION` keyword.

Query modifiers. Since query patterns generate an unordered collection of solutions, these solutions can be modified to create another more informative sequence. A solution sequence modifier can be the following [96]:

1. `ORDER BY`: for ordering the solution sequence in a specific order,
2. `DISTINCT`: for eliminating duplicate solutions,
3. `LIMIT`: for specifying an upper bound on the number of solutions returned,
4. `PROJECTION`: for choosing certain variables,
5. `REDUCED`: for permitting the elimination of some, or all, of the duplicate solutions, and
6. `OFFSET`: for specifying where the solutions start after the specified number of solutions.

CHAPTER 3

Related Work

In recent years, there is a considerable amount of literature on the metadata analysis of scholarly data. The mega-trend of digitization has made the preparation, submission, and publication of manuscripts, as well as the organization of scholarly events substantially easier and efficient. Few researchers have addressed the problem of identifying the characteristics of renowned scholarly events in Computer Science overall or within a particular community. However, none of them provides services to ease the process of gaining an overview of a field, which is one of the contributions of this thesis. Overall, we found that the characteristics of these events have not been comprehensively studied. This chapter presents a literature review of the research efforts related to the research questions defined in this thesis. We have divided the literature on this topic into three areas: scientometric analyses of scholarly events metadata, datasets representing scholarly events metadata, and data models supporting scholarly communication. Relevant approaches, data models, and tools related to this work are considered.

This chapter is organized as follows: section 3.1 examines research studies that track the evolution of a particular research community by analyzing the metadata of its renowned scholarly events. Here, we elaborate on the methodologies that have been followed as well as the outcomes of these analyses. A brief description of scholarly events metadata available on the Web, as Linked Data, is given in section 3.2. We begin section 3.3 with a brief description of how ontologies can support scholarly communication. Then, we look at how different parts of scholarly artifacts can be represented using ontologies by presenting state-of-the-art data models.

3.1 Scientometric Analysis of Scholarly Events Metadata

The next decade is likely to witness a considerable rise in metadata analysis of scholarly events as the metadata of scholarly events have become increasingly available on the Web. Most of them are available under an open license. In our recent review of the literature [6, 22, 25, 99, 100], we found that most studies tended to focus on grabbing information about scholarly communication through analyzing Computer Science communities publications. In the literature, there are several attempts to develop tools for analyzing scholarly data. For instance, Osborne, Motta and Mulholland [101] developed the *Rexplore* tool for exploring and making sense of scholarly data through integrating visual and statistical analytics.

3.1.1 Ranking Services

There are several online services for ranking scholarly events available to help researchers to assess the scientific impact of these events. Table 3.1 lists a set of popular such services. This section gives a short overview of these services.

CORE[1] (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*` (i.e., flagship), `A` (i.e., excellent), `B` (i.e., good), `C` (i.e., meet minimum standards), `Australian` (i.e., the audience is primarily Australians and New Zealanders), `National` (i.e., run primarily in a single country), `Regional` (i.e., similar to National but may cover a region crossing national borders), and `un-ranked` (i.e., no ranking decision has been made yet). The portal shows international event series in the first four classes.

Guide2Research[2] is an academic portal for ranking of top scientists and events in Computer Science and Electronics. Conferences ranking by Guide2Research is based on the h5-index of the conference provided by Google Scholar Metrics (GSM)[3]. Only conferences with H5-index greater than or equal to 12 are considered. H5-index, as defined by Google, is "the largest number h such that h articles published in 2015-2019 have at least h citations each". H-index is computed for articles published in the last five complete years. Information about CfP of the upcoming conferences is also provided, such as submission deadline, conference dates, conference address, and conference website.

AMiner conference ranking[4] is a ranking service for ranking renowned scholarly events in ten Computer Science sub-fields. Conferences ranking is also based on the H5-index of the conference. In addition to ranking events, it provides some useful information about them, including keywords frequency, top-cited authors in the last five years, and authors' distribution among countries, gender, and language.

QUALIS: The Brazilian Federal Agency sponsors the QUALIS conference ranking with the aim of improving Higher Education. It uses the h-index as a performance measure for conferences. Based on the h-index percentiles, the conferences are ranked into seven classes – in decreasing order: `A1`, `A2`, `B1`, `B2`, `B3`, `B4`, and `B5`.

ERA[5] (Excellence in Research for Australia) ranking is created by the Australian Research Council. Evaluations are conducted by committees of distinguished researchers from around the world. The classes, in decreasing order, are: `A` , `B`, and `C`.

MSAR[6] (Microsoft Academic's conference field ratings) ranking is created by Microsoft Academic, which extracts information about publication venues, such as conferences, creating Microsoft Academic Graph (MAG). MAG is used to provide useful information about conferences, such as upcoming deadlines, top authors, number of papers, and number of citations. Conferences are ranked using their h-index.

GGS[7] ranking is sponsored by Group of Italian Professors of Computer Engineering, Group of Italian Professors of Computer Science, and Spanish Computer-Science Society. The ratings

[1] `http://www.core.edu.au/`

[2] `http://www.guide2research.com/`

[3] `https://scholar.google.com/intl/en/scholar/metrics.html`

[4] `https://aminer.org/ranks/conf`

[5] `https://www.arc.gov.au/excellence-research-australia/era-2018`

[6] `https://academic.microsoft.com/conferences`

[7] `http://gii-grin-scie-rating.scie.es/ratingSearch.jsf`

Table 3.1: **Comparison of scholarly events ranking services**. Popular ranking services that are widely used to identify renowned events in various Computer Science communities.

Service	**Ranks**	**Criteria**	**Provide CfP**	**Metadata analysis**	**Standalone service**
CORE	A*, A, B, C, Australian, National, Regional, and Unranked	Community-defined criteria	✗	✗	✓
Guide2Research	Numeric	Google Scholar Metrics	✓	✗	✓
AMiner	Numeric	Google Scholar Metrics	✗	✓	✓
QUALIS	A1, A2, B1, B2, B3, B4, and B5	Google Scholar Metrics	✗	✗	✗
ERA	A, B, and C	Community-defined criteria	✗	✗	✗
GGS	A++, A+, A, A-, B, B-, and C	Automatic algorithm well-known, existing international classification	✗	✗	✓
MSAR	Numeric	Google Scholar Metrics	✓	✓	✓
ConferenceRanks	uses ranks of ERA, QUALIS, and MSAR		✗	✗	✓

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`.

SCImago Journal Rank (SJR indicator) is a measure of the scientific influence of scholarly journals and events based on both the number of citations received by a journal and the prestige of the journals where such citations come from [102]. It is publicly available via an online portal[8]. This rank is calculated based on the information contained in the Scopus[9] database starting from the year 1996.

3.1.2 Meta-analytic Research Activities

Analyzing metadata of scholarly events has received much attention in the past decade [24–26, 103, 104]. In addition, several studies [20–23, 25, 104] assessed the evolution of a particular

[8] `http://www.scimagojr.com`

[9] `https://www.scopus.com/`

scientific Computer Science community by analyzing the metadata of publications of a specific event series. These studies highlight the peculiarities of the individual domains in terms of publications and collaboration trends and compare them to each other.

Preliminary work on meta-analytic methods in the 1980s focused on synthesizing the results of statistical methods and analyzing statistical variables for the purpose of making sense of them [105, 106]. Guilera et al. [107] presented a meta-analysis for publications in psychology in order to provide an overview of meta-analytic research activity and to show its evolution over time. Different bibliometric indicators were used, such as the number of authors per article, productivity by country and national, and international collaborations between authors. El-Din, Sharaf Eldin, and Hanora [108] presented a descriptive analysis of Egyptian publications on the Hepatitis C virus using several indicators, such as the total number of citations, authors and their affiliations, publication types, and the Google Scholar citation index. Bakare and Lewison [109] investigated the Over-Citation Ratio (i.e., researchers tending to over-cite researchers from the same country) for publications from six different scientific fields based on data from Web of Science (WoS) Clarivate Analytics. This analysis was performed in 2010 on publications from 20 countries in seven different years between 1980 and 2010. The authors concluded that chemistry and ornithology had had the highest, while astronomy and diabetes research papers had had the lowest over citation rate.

Biryukov and Dong [100] provided an in-depth exploration of Computer Science communities according to publications of several event series based on the DBLP database (as in August 2009), in XML format. They typically examined the authors in the co-author network through their contributions to the research community, i.e., their publications. One of the outstanding observations they found is that there are 2,623 authors in the *top-Conferences* dataset whose career is more than ten years, they called them "*experienced scientists*". Besides, around 29% of them are working in one area only. Regarding collaboration trends, the authors identified that the *Data Mining* community exhibits less collaboration than *Information Retrieval*, the latter focuses on a much smaller number of topics and thus facilitates the collaboration.

Similarly, Aumüller and Rahm [21] analyzed affiliations of *Database* publications (more than 2,700 papers) in the renowned Database conferences SIGMOD and VLDB and in the VLDBJ and TODS journals in the period 2000–2009 using author information from DBLP. The outcomes of the affiliation analysis indicated that the number of papers per year is almost doubled during the decade (i.e., from 188 to 352). Besides, about 60% of the papers and nearly 25% of the research papers have authors from two or more countries.

Nascimento, Sander and Pound [23] analyzed the co-authorship graph of all papers published at the SIGMOD conference in the period 1975–2002. The most striking observation is that around 70% of the authors have only one paper in that period. Overwhelmingly, over 90% of the authors in the co-authorship graph have less than or equal to three papers, but single individuals have a large number of papers, i.e., there are 32 single-author papers.

Yan and Lee [110] proposed a list of alternative measures for ranking events based on the "*goodness*" of the articles published in these events. The *goodness*, of both events and journals, is defined as the goodness of the articles published in them. Then, a list of new measures is proposed for *goodness* of articles considering bibliographic properties, author information, and readers' evaluation. In contrast, our comparisons are based on metrics derived from the characteristics of the events themselves, such as event continuity, geographic distributions, etc.

Agarwal, Mittal and Sureka [99] presented a bibliometric analysis of the metadata of seven ACM conferences covering different Computer Science sub-fields, including Knowledge Management (KM), Data Mining (DM), Digital Libraries (DL), and Information Retrieval (IR). The authors presented an exploratory and a scientometric analysis of authors and publications at SIGWEB[10] conferences in 10 years (from 2006 to 2015). This analysis involves the number of articles, h5-index, and overall citation rate of papers in the DBLP database (released on August 8, 2016).

Despite the low growth rate in the number of publications and authors' participation over the past decade, some conferences, e.g., the ACM Conference on Recommender Systems (RecSys), had received a large number of citations every year, e.g., RecSys had got up to 150 citations for the papers published in 2008.

3.1.3 Meta Analysis of Event Series

To better assess progress and quality of a particular event series, exploratory and descriptive data analyses are performed by steering committee members or other members of the community. They often include the analysis of bibliographic data of each edition and rarely comprise comparisons with other events or editions of the same event series. The ultimate goal of metadata analysis of scholarly events metadata is to uncover information hidden and explore the characteristics of the renowned events within the community. In this section, a brief description of the research work related to analyzing metadata of a particular event series will be presented. Table 3.2 reports on the state-of-the-art approaches for meta-analysis of event series metadata in various Computer Science communities, involving Computer-Supported Collaborative Learning (CSCL), Computer Supported Cooperative Work (CSCW), Digital Libraries (DL), Human-Robot Interaction (HRI), and Knowledge Management (KM), along with the analytical dimensions used.

The analysis presented in [111–113] showed the growth of the Conference on Human Factors in Computing Systems (CHI), most of the people in the organization committees are from the USA, Canada, and the UK. The number of published papers has drastically doubled, and the length of the papers has also increased. In addition to these studies, Liu, Goncalves, Ferreira, Xiao, Hosio and Kostakos [114] analyzed publications of CHI conference editions from 1994 to 2013 through co-word analysis. They utilized strategic diagrams, granular networks, and hierarchical cluster analysis in order to identify the major research themes and trends within the Human-computer interaction community. They constructed Keywords networking map of keywords in published papers within two different periods: 1994–2003 and 2004–2013. A core-periphery analysis, to determine which nodes are part of a densely connected core [115], was performed to determine the core research topics in HCI based on the structure of these networks.

A comprehensive analysis of the Principles of Database Systems (PODS) conference series includes detailed author analyses such as the distribution of the number of papers per author, which, for example, showed that two-thirds of the authors are only involved in a single PODS publication (e.g., Ph.D. students) but 10% are involved in five or more (e.g., active supervisors) [20]. It also includes a relatively short analysis of submission and acceptance rates for ten years (2002–2011), which shows an increasing number of submissions at the beginning of the period, while they reduced in the last four years.

Barbosa, Silveira and Gasparini [25] analyzed full papers published in the Brazilian Symposium on Human Factors in Computing Systems (IHC) conference series in the period 1998–2015 intending to investigate the evolution of the Brazilian HCI community. They found that one of the main characteristics of the IHC conference is to be a peripatetic conference, in distinct regions of Brazil. This finding opens the way for community members to participate, either as a listener or presenting a paper or poster.

[10] `https://www.sigweb.org/`

Table 3.2: **Meta analysis of event series**. Approaches for meta analysis of event series metadata in various Computer Science communities along with the analytical dimensions used.

Approach	**Series**	**Analytical dimensions**	**Field**	**Temporal coverage**
Liu et al. [114]	CHI	Co-word analysis	HCI	1994–2013
Bartneck et al.[111], Barkhuus et al.[112], Greenberg et al.[113]	CHI	Number papers, length of the papers, affiliation analysis, bibliometric analysis	HCI	1983–2008
Ameloot et al. [20]	PODS	Distribution of the number of papers per author, submission, acceptance rates	DB	2002–2011
Barbosa et al. [25], Gasparini et al. [116]	IHC	Geographical distribution, co-authorship, distribution over time, number of papers, bibliographic analysis	HCI	1998–2015
Correia et al. [117]	ACM, CSCW, ECSCW, JCSCW	Publication patterns, citation analysis, author analysis, institutional and geographical distribution, keyword analysis	CSCW	2001–2015
Wainer et al. [118]	ACM, CSCW	Bibliographic analysis, accepted papers, type of empirical research	CSCW	1998–2004
Kienle et al. [119]	CSCL	Geographical distribution, co-authorship, policy for the selection of conference location and program committees	CSCL	1995–2005
Bartneck [120]	HRI	Keyword analysis, citation count, ranking of countries, top organizations, top authors	HRI	2006–2010
Nichols et al. [121]	CHINZ	Citation analysis, type of papers, h-indices, keyword analysis	HCI	2001–2012
Hiemstra et al. [22]	ACM SIGIR	Authors distribution, the number of papers per year for each country, co-authorship	IR	1978–2002
Smeaton et al. [122]	ACM SIGIR	Co-authorship network analysis	IR	1978–2002
Nascimento et al. [23]	SIGMOD	Co-authorship network analysis	DB	1975–2002
Aumüller et al. [21]	SIGMOD, VLDB	Affiliation analysis, frequency distribution of publications, average of participants per paper, intra- and cross-country co-operations	DB	2000–2009
Agarwal et al. [99]	SIGWEB	Bibliometric analysis, scientometric analysis of authors	KM, DM, DL, IR	2006–2015

Correia, Paredes and Fonseca [117] analyzed a corpus of 1713 Computer Supported Cooperative Work (CSCW) publications, including 985 papers published at ACM CSCW, 298 papers from the International Conference on Supporting Group Work (GROUP), 165 papers from the European Conference on Computer-Supported Cooperative Work (ECSCW), and 265 articles published in the Journal of Collaborative Computing and Work Practices (JCSCW). They performed a 15-year period (2001–2015) scientometrics analysis using several dimensions, involving publication patterns, citation analysis, author analysis, institutional and geographical distribution, and keyword analysis.

Hiemstra, Hauff, De Jong and Kraaij [22] analyzed the International ACM SIGIR Conference on Research and Development in Information Retrieval publications in the period 1978–2007 in terms of authors' countries, the number of papers per year for each country and co-authorship. This study described the geographical origin, i.e., the country, of the papers published in this period. These papers are distributed among 41 countries, but the divergence of those papers over the various countries was very high. In terms of productivity, the USA comes first, i.e., the most productive country, with 595 papers published in this time span, then followed by the UK, Canada, and Germany.

Similarly, Smeaton et al. [122] performed a content analysis of papers at the SIGIR Conference since its beginning until 2002. This analysis focuses on the co-authorship network analysis, which is based on a set of 853 papers assembled from the 25 years of SIGIR Conference. The objective is to measure the centrality of the author in terms of a co-authorship graph.

Despite these continuous efforts, a comprehensive comparative analysis of the characteristics of scholarly events in multiple scientific communities has not been performed to the best of our knowledge. We observed that the characteristics of these events had not been comprehensively studied. The research mentioned above work has led us to conduct in-depth analysis (presented in chapter 4), which is based on a comprehensive list of metrics, considering quality in terms of event-related metadata in eight communities within Computer Science and also events from other communities, such as Physics, Mathematics, and Engineering.

3.2 Linked Datasets Representing Scholarly Information

Recently, publishing scholarly events metadata as Linked Data has become of prime interest to several publishers, such as Springer and Elsevier. There are several sources of scholarly events metadata on the Web available as Linked Data, including DBLP, DBWorld, WikiCFP, EVENTKG, and SWDF. A brief description of these datasets is given in the remaining part of this section. Table 3.3 outlines the main characteristics of such datasets.

DBLP[11] and *DBWorld*[12], the most widely known bibliographic databases in Computer Science, are widely used as a source of the data being analyzed, which provide information mainly about publications and events, but also consider related entities, such as authors, editors, conference proceedings and journals. DBLP allows event organizers to upload XML data with bibliographic data for ingestion.

WikiCFP[13] is a Semantic Wiki for Calls for Papers (CfP) of scholarly events in Computer Science with the goal of helping researchers to organize and share academic information in an efficient way. Computer Science events are categorized into several categories, based on the scope

[11] http://dblp.uni-trier.de/

[12] https://research.cs.wisc.edu/dbworld/

[13] http://www.wikicfp.com/

of the events, including image processing, data mining, information systems, and computational intelligence. It employs web crawlers to track high-profile conferences.

The first considerable work to provide a comprehensive semantic description of scientific events metadata is the Semantic Web Conference (SWC) ontology [123] and Semantic Web Dog Food (SWDF) corpus [124]. The Semantic Web Dog Food (SWDF) dataset and its successor *ScholarlyData* are among the pioneers of datasets of comprehensive scholarly communication metadata. ScholarlyData contains Linked Data about papers, people, organizations, and events related to academic conferences. SWDF data is available for download as single RDF dumps[14] for each event in various formats, such as RDF/XML, Turtle, N-triples, and JSON-LD, and can be accessed via URI dereferencing or queried via a Virtuoso SPARQL endpoint[15]. The Semantic Web Conference (SWC) ontology[16] is used as a reference ontology for modeling data about scholarly events contained in SWDF [123].

Vasilescu et al. [125] presented a dataset of eleven well-established Software Engineering Conferences (SEConf), such as ICSE and ASE, containing accepted papers along with their authors, program committee (PC) members and the number of submissions each year. The purpose of creating such a dataset is to help conference steering committee chairs to assess their selection process (e.g., the change in PC members) and study the number of newcomers year by year. The data about accepted papers and the corresponding authors were extracted from the DBLP records, accessed February 2013, while the PC members and the number of submissions were retrieved from events' websites and online proceedings.

Luo and Lyons [126] presented a dataset (CASCONet[17]) containing 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)[18] in the period 1993–2017. CASCONet data has been collected from 25 editions of CASCON conferences (around 800 CASCON papers). It contains not only information about papers, but also data about all aspects of the CASCON conference, including papers' metadata, demos, panels, workshops, and keynote presentations. The core entity of the CASCONet dataset is "*Person*" and the associated role in each conference. The outstanding findings the authors found (after analyzing the dataset) is that 24% of the authors have published more than one paper, and the maximum number of papers published by one person is 20 [126]. Furthermore, they found that the number of accepted papers every year since its establishment is relatively stable.

Springer LOD is a LOD dataset containing metadata about 8,965 conferences belonging to 1,646 conference series in which Springer is the publisher of their proceedings. Anyone can access Springer's LOD platform free of charge. Springer LOD enables users to find published papers by ISBN, DOI, conference acronym, or the volume number of the book series. The dataset is openly available for download at DataHub[19]. In 2017, Springer Nature (SN) published the first release of the Linked Open Dataset (SN SciGraph) aggregated from several data sources involving journals, events, institutions, funders, research grants, patents, clinical trials, and conference series. SN SciGraph contains more than a billion triples of scholarly information divided into several separated chunks (accessible at `https://sn-scigraph.figshare.com/`) for Persons,

[14] `http://www.scholarlydata.org/dumps/`

[15] `http://www.scholarlydata.org/sparql/`

[16] `http://www.scholarlydata.org/ontology/`

[17] `https://github.com/iDBKMTI/CASCONet`

[18] `http://cascon.ca/`

[19] `https://old.datahub.io/dataset/lod-for-conferences-in-computer-science`

Table 3.3: **Linked Dataset Representing Scholarly Information.** State-of-the-art published Linked datasets representing different scholarly entities, including Publications, events, persons, and organizations, along with their size, license, dump format(s) and whether a search facility is provided or not. *Abbreviations:* CC stands for Creative Commons, ODbL for Open Database License, ODC for Open Data Attributions license, and ARFF for Attribute-Relation File Format.

Dataset	**Size (B)**	**Entity types**	**License**	**Format**	**Search facility**
DBLP	55M	Publications, persons	CC0 1.0	NT	Yes[11]
DBWorld	153K	Events	N/A	ARFF	No
WikiCFP	N/A	Events	CC BY-SA 3.0	N/A	Yes[13]
SNSciGraph	12G	Publications	CC BY 4.0	JSON-LD	No
SWDF	242K	Publications, persons, organizations	CC BY 3.0	RDF/XML, TTL, NT, JSON-LD	Yes[15]
SEConf	541K	Publications, persons	ODbL v1.0	SQL dump	No
CASCONet	1.32M	Publications, persons, events	N/A	SQL dump	No
EVENTKG	4.7G	Events	CC BY 4.0	NT, TTL	Yes[21]
MAKG	1.2T	Publications, persons, organizations	ODC-By v1.0	NT	Yes[22]
BibBase	200K	Publications, persons, organizations	CC BY-SA 3.0	RDF/XML	No
OpenCitations	3M	Publications, persons, organizations	CC BY 4.0	JSON-LD	No
BNB-LOD	109M	Publications, persons	CC-Zero	NT, RDF/XML	Yes[23]

Grants, Books, etc. The datasets have been made available under CC-BY licensing.[20]

In 2018, Gottschalk and Demidova [127] published a multilingual RDF knowledge graph (EVENTKG) about events and temporal relations aiming at describing them in various languages. Currently, EventKG covers five natural languages, including English, German, French, Russian, and Portuguese. It describes general events at a high level of abstraction. The main objective is to paint a global view of events and temporal relations spread across entity-centric knowledge graphs, such as Wikidata and DBpedia. EVENTKG reuses several existing data models, including Simple Event Model [128], thus enabling efficient reuse in real-world applications [127]. Furthermore, the authors developed an open-source extraction framework to extract and maintain the subsequent versions of the dataset.

In 2019, Färber [129] presented the Microsoft Academic Knowledge Graph (MAKG)[24]. It is a

[20] `https://www.springernature.com/gp/researchers/scigraph`

[21] `http://eventkginterface.l3s.uni-hannover.de/sparql`

[22] `http://ma-graph.org/sparql`

[23] `https://bnb.data.bl.uk/flint-sparql`

[24] `http://ma-graph.org/rdf-dumps/`

large RDF dataset, based on the Microsoft Academic Service (MAG[25]) [130], with over eight billion triples with information about scientific publications and related entities. It covers ten entity types, including authors, institutions, authors, and fields of study. Both the initial MAG and the MAKG are licensed under the Open Data Commons Attribution License[26]. Compared to Wikidata, MAKG contains significantly more bibliographic information (about 13 times more).

British National Bibliography (BNB-LOD)[27] is Linked Open Dataset, created by the British National Bibliography (BNB)[28], linked to external sources involving LexVo[29] and Library of Congress Linked Data Service[30]. The dataset contains descriptions of books and serials published in the UK since 1950. Entities and relationships were described using well-known ontologies, such as Dublin Core and FOAF, while new classes and properties were defined in the British Library RDF schema[31] [131].

3.3 Data Models Supporting Scholarly Communication

In the past years, several ontologies for supporting scholarly communication have been published. Mainly, these ontologies have been created, for example, for describing how surveys for research fields can be represented in a semantic format, modeling knowledge about conferences, and characterizing the publication status of a document, etc. Nevertheless, it is also essential to describe the content of the published article in order to enhance the search services across a large number of research articles published each year. In the next subsections, we categorized data models supporting scholarly communication into three categories: 1) Data models for describing scholarly articles, 2) Data models for describing bibliographic citations, 3) Data models for describing semantic publishing, and 4) Data models for scientific events. These models are summarized in Table 3.4.

3.3.1 Data Models for Describing Scholarly Articles

Despite significant advances in technology in the last decades, the way how research is accomplished and especially communicated has not changed much. Researchers still present their findings in the form of text accompanied by illustrations and wrap these into articles, which are mostly published in printed form or as semi-structured PDF documents. It is better for researchers to work on a common knowledge base comprising comprehensive descriptions of their research, thus making research contributions transparent and directly comparable. Semantic technologies provide enormous support to scholarly communication in sharing, disseminating, and publishing research findings. Therefore, scientific data analysis, information search, and data integration have become more efficient for global academic communities. This process is called *Semantic Publishing*. Semantic publishing provides machine-comprehensible representations of scientific methods, models, experiments, and research data [132]. Several ontologies have been created in the last years for the semantic annotation of scholarly publications and scientific documents. A growing body of literature has addressed the problem of describing various parts

[25] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

[26] https://opendatacommons.org/licenses/by/1-0/index.html

[27] https://bnb.data.bl.uk/doc/data/BNB

[28] http://bnb.data.bl.uk/

[29] http://www.lexvo.org/

[30] http://id.loc.gov/

[31] http://www.bl.uk/schemas/bibliographic/blterms

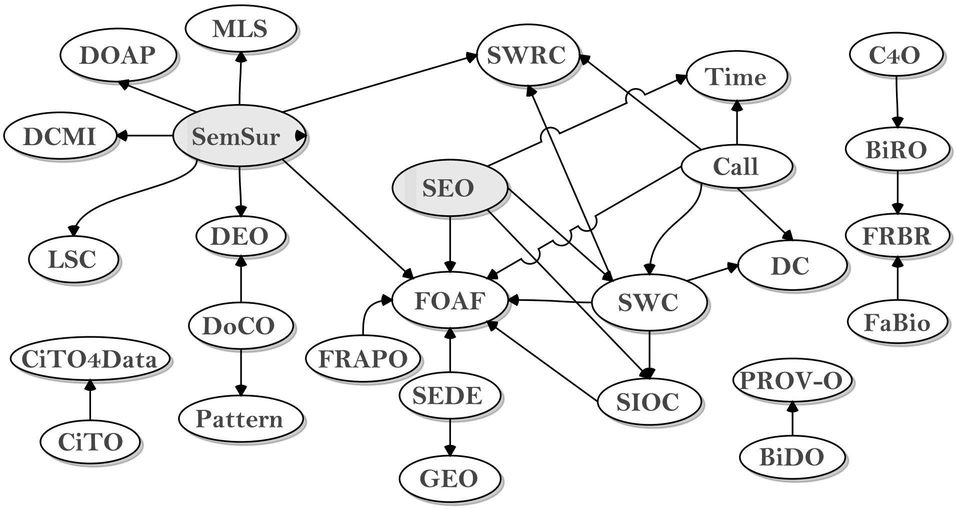


Figure 3.1: The dependency graph of the data models supporting scholarly Communication. More information about highlighted ontologies (part of the contributions of this thesis) will be presented in the next chapters.

of scholarly articles using ontologies. In recent years, several ontologies have been developed for describing scholarly communication metadata. For modeling the top-level metadata, such as authors, affiliations, abstract, etc, of a scientific article, seven well-known ontologies are used: DCMI[32], SWRC[33], DEO[34], LSC[35], DOAP[36], MLS[37], and FOAF[38].

In the remainder of this section, we focus on the state-of-the-art ontologies for describing the content and the structure of scholarly articles. Figure 3.1 shows a part of the dependency graph of data models supporting scholarly communication. A part of these efforts is described below.

Ontology of Rhetorical Blocks (ORB)[39] is a formalization capturing the coarse-grained rhetorical structure of scientific publications. In Computer Science, there are blocks carrying rhetorical roles such as *Scenario*, *Related Work* and *Evaluation*, while in the Biomedical science, there are others (in the majority of the papers), such as *Background*, *Method, Experimental Results* or Discussion. In principle, most of these blocks can be found in the ample majority of the domains, however, with slightly different names. Consequently, ORB ontology represents a publication through three components: the Header, the Body, and the Tail. The header `orb:Header` describes meta-information about the publication, such as title, authors, affiliations, publishing venue, and abstract, as a rhetorical summary. The body is composed of four rhetorical blocks adapted from the IMRAD (Introduction, Methods, Results, and Discussion)[133], the mostly model used for structuring scientific articles.

The core entities in ORB ontology are: 1) The methods block `orb:Methods`, which describes when, where, and how was the study done, 2) The introduction block `orb:Introduction`

[32] http://purl.org/dc/terms/
[33] http://swrc.ontoware.org/ontology
[34] http://purl.org/spar/deo/
[35] http://linkedscience.org/lsc/ns
[36] http://usefulinc.com/ns/doap
[37] https://www.w3.org/ns/mls
[38] http://xmlns.com/foaf/0.1/
[39] https://www.w3.org/TR/hcls-orb/

describes what the tested hypothesis is and what is the purpose of the research, 3) The results block `orb:Results`, which describes the results of the study presented in the paper, and 4) The discussion block `orb:Discussion`, which specifies whether the tested hypothesis was confirmed, interprets the results to understand their consequences and importance, and presents possible perspectives of future research. Finally, the Tail provides additional meta-information about the paper, such as Acknowledgments (`orb:Acknowledgements`) and References (`orb:References`). After the first development, ORB got an extension in which some rhetorical blocks have been decomposed into finer-grained entities. For example, the Methods block is decomposed into several entities, such as *Purpose*, *Objects of Study*, *Tools* and *Procedures*. The ORB is available in RDF format. For more illustration, a Turtle serialization for an article [6] described using ORB ontology is listed in Listing 3.1. In this example, three ontologies are reused: ORB, Dublin Core, and BIBO.

```
@prefix bibo: <http://purl.org/ontology/bibo/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix orb: <http://purl.org/orb/> .
[ a bibo:Article ;
    dc:creator "Said Fathalla" ;
    dc:creator "Sahar Vahdati" ;
    dc:creator "Christoph Lange" ;
    dc:creator "Soeren Auer" ;
    dc:title "Towards a Knowledge Graph Representing Research Findings by Semantifying Survey
    Articles" ;
    dcterms:abstract " Despite significant advances in ..."
    dcterms:hasPart
    [ a orb:Introduction ;
    dc:description "Introduction content ..."
    ]
    ...
].
```

Listing 3.1: A Turtle serialization for a part of the metadata of the article [6] described using ORB ontology.

SPAR ontologies. One of the first efforts to develop data models for describing the various parts of published articles is the SPAR (Semantic Publishing and Referencing) ontologies[40] [15, 134]. The SPAR ontologies are a set of modular OWL 2 DL ontologies for describing the main aspects of the publishing domain, such as document components, bibliographic references, publishing roles, and workflow processes. SPAR Ontologies are used for creating and publishing machine-readable RDF metadata concerning the publishing domain. The original suite of SPAR ontologies comprises eight distinct modules which are based on two other more general models: FRBR (Functional Requirements for Bibliographic Record)[41] and DEO (the Discourse Elements Ontology)[42]. The current suite[43] contains 17 ontologies [134]:

- *The FRBR-aligned Bibliographic Ontology* (FaBiO) is an ontology for describing published articles, such as journal articles, conference papers and book chapters, that contain or are referred to by bibliographic references [135],

[40] `http://www.sparontologies.net`
[41] `http://www.sparontologies.net/ontologies/frbr`
[42] `http://purl.org/spar/deo`
[43] `http://www.sparontologies.net/ontologies`

- *The Citation Typing Ontology* (CiTO) is an ontology that enables characterization of citations (either direct or indirect) in scientific articles, enabling citations descriptions to be published on the Web [135],
- *The Bibliographic Reference Ontology* (BiRO) is an ontology for describing individual bibliographic references found in scientific articles and link them to references lists [136],
- *The Citation Counting and Context Characterization Ontology* (C4O) is an ontology that permits the number of in-text citations of a cited source to be recorded along with their textual citation contexts [136],
- *The Document Components Ontology* (DoCO) is an ontology for describing document components (reuses ORB) from different perspectives: structural, e.g., paragraph and section, and rhetorical, e.g., discussion, appendix, and acknowledgments [137],
- *The Publishing Status Ontology* (PSO) is an ontology for describing the publication status of documents within each phase of the publishing process, such as accepted, submitted or under review [138],
- *The Publishing Roles Ontology* (PRO) is an ontology for describing the roles of agents (i.e., people, corporate bodies and computational agents) in the publication process, such as authors, editors, reviewers, publishers or librarians [138], and
- *The Publishing Workflow Ontology* (PWO) is an ontology for describing the logical steps in the workflow in the publication process of a document, e.g., writing the article [139].

SPAR ontologies[44] have been extended with five complementary ontologies that extend the coverage of the possible description of the publishing domain:

- *The Scholarly Contributions and Roles Ontology* (SCoRO) is an ontology for describing the contributions and roles of scholars, i.e., authors, in a published article, e.g. intellectual contributions and authorship roles respectively,
- *The Funding, Research Administration and Projects Ontology* (FRAPO) is an ontology for describing the administrative information associated with published articles, such as funding bodies and research projects,
- *The DataCite Ontology* (DataCite) is an ontology that enables the metadata properties of the consistent identification of a resource for citation and retrieval purposes by defining identifiers for bibliographic resources, e.g., papers and datasets,
- *The Bibliometric Data Ontology* (BiDO) is an ontology describing numerical and categorical bibliometric data, such as impact factor of journals and h-index of authors [140],
- *The Five Stars of Online Research Articles Ontology* (FiveStars) is an ontology for describing the five attributes of a published journal article: peer review, open access, enriched content, available datasets, and machine-readable metadata [141].

Semantic Web for Research Communities ontology (SWRC)[45] is an ontology for modeling different entities involved in the scholarly domain, such as persons, organizations, bibliographic metadata of publications, and the relationships among them[142]. SWRC ontology is written in OWL. The core concepts in the SWRC ontology are Person, Publication, Event, Organization, Topic, and Project. These concepts have been specialized to more specific ones. For instance, the *Person* concept is specialized by several sub-concepts, e.g., *Employee* and *Student*, and the *Thesis* concept is specialized by *MasterThesis* and *PhDThesis*. SWRC is modularized into a core

[44] `http://www.sparontologies.net/ontologies`

[45] `http://ontoware.org/swrc`

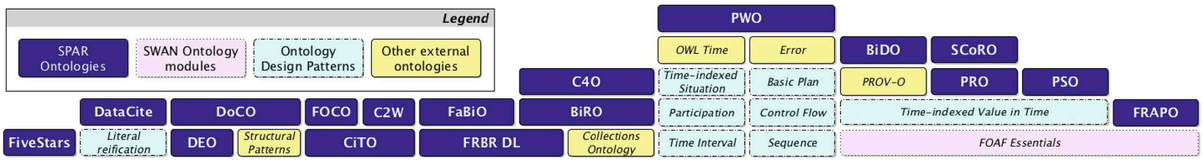


Figure 3.2: The SPAR Ontologies and their relations with other models [134].

ontology and several extension modules. Several Web portals use the SWRC ontology to annotate staff, publications, projects such as the portal of the Institute "*Angewandte Informatik und Formale Beschreibungsverfahren*" (AIFB)[46], OntoWare[47] (a software development community platform for Semantic Web related software projects), and the European Semantic Knowledge Technologies project (SEKT)[48].

Scholarly Article (SA)[49] ontology compromises a set of concepts related to published articles, such as *Article*, *Keywords*, *Contributor* and *Citation*. Moreover, it comprises a set of properties, such as *isStyleOf*, *roleAffiliation*, and *dateRejected*. SA is formally written in RDFS and uses the semantic of schema.org Role. Roles are predominantly useful for concepts of scholarly communication, such as author affiliations, correspondence, and sources of financial backing. These concepts are often specific to a role, e.g., a person as an author of a particular article. For instance, a person can be affiliated with different institutions but has contributed to, at a given work, only a subset of the institutions he/she is affiliated with.

Scientific EXPeriments Ontology (EXPO)[50], is an ontology, written in OWL-DL, proposed by the University of Wales as a core ontology about scientific experiments [143]. EXPO ontology aims to provide a structured framework for a consistent and shareable description of scientific experiments by formalizing the generic concepts involved, such as experimental design, methodology, and results representation. Furthermore, it aims to provide a formal description of experiments for efficient analysis, annotation, and sharing of results. For this purpose, EXPO defines over 200 concepts. EXPO can describe computational and physical experiments, experiments with an explicit and implicit hypothesis. EXPO defines a class taxonomy, including `ScientificExperiment`, the most general class, and a set of related classes, including `ExperimentGoal`, `ExperimentTechnology`, `ExperimentResult`, and `ExperimentHypothesis`.

3.3.2 Data Models for Bibliographic Citations

Much work on the potential of modeling bibliographic citations has been carried out. A brief description of some of this work is outlined below.

Citation Typing Ontology (CiTO)[51] is an ontology written in OWL 2 DL to provide a set of object properties related to citing published articles, e.g., the object property `cito:cites` and its inverse property `cito:isCitedBy` [135]. Upon the creation of version 2.0 of CiTO, several new sub-properties of `cito:cites` have been added, and the inverse properties of all its sub-properties were created, all of which are sub-properties of `cito:isCitedBy`. The ontology

[46] `http://www.aifb.uni-karlsruhe.de`
[47] `http://www.ontoware.org`
[48] `http://www.sekt-project.com/`
[49] `http://ns.science.ai/`
[50] `http://expo.sourceforge.net/`
[51] `http://purl.org/spar/cito`

has also been integrated with the SWAN[52] (Semantic Web Applications in Neuromedicine) Ontology by making `cito:cites` a sub-property of `swan:refersTo`. The central concept in CiTO is the `cito:CitationAct`, which is a performative act of making a citation from a citing entity to a cited entity, typically instantiated by the inclusion of a bibliographic reference or a data reference in the reference list of the citing entity.

Bibliographic Ontology (BIBO)[53] is an ontology for describing the main concepts and properties related to citations and bibliographic references on the Semantic Web [144]. It had the latest update in 2009. The prominent feature of BIBO is that it links the article directly to the original published paper on the Web through its Digital Object Identifier (DOI), as shown in Listing 3.2.

```
<info:https://doi.org/10.1016/j.websem.2009.07.002>
    a bibo:Article ;
    dc:title "DBpedia - A crystallization point for the Web of Data"@en ;
    dc:date "2009-09-01";
    dc:isPartOf <urn:issn:1570-8268>;
    bibo:volume "7" ;
    bibo:issue "3" ;
    bibo:pageStart "154" ;
    bibo:pageEnd "165" ;
    bibo:authorList(<http://ex.net/author/1> <http://ex.net/author/2>...).

<urn:issn:1570-8268> a bibo:Journal ;
    dc:title "Web Semantics"@en ;
    bibo:shortTitle "Web Semantics"@en .
```

Listing 3.2: Representation of an article using BIBO ontology.

Bibliographic Reference Ontology (BiRO)[54] is an ontology for describing individual bibliographic references found in scientific articles and link them to references lists, each of which, i.e., a reference item, references a publication. It is structured according to the FRBR model to define bibliographic records and bibliographic references. The most prominent feature of BiRO is that it links an individual bibliographic reference, in the reference list (`biro:ReferenceList`) of a published article, to the full bibliographic record (`biro:BibliographicRecord`) for that cited article.

Citation Counting and Context Characterization Ontology (C4O)[55] is an ontology for characterizing the number and contexts of bibliographic citations. C4O imports FRBR, BiRO, and RDFS ontologies. A source of information about bibliographic citations, such as Google Scholar, is represented by a class named `BibliographicInformationSource`, and the number of times a work has been cited is represented by a class called `GlobalCitationCount`. C4O has an object property called `denotes`, which is used to assert the connection between an in-text reference pointer and the bibliographic reference it points to.

3.3.3 Data Models for Scholarly Events

In recent years, several data models have been developed for describing events, such as the Event Ontology (EO) [145], Linking Open Descriptions of Events (LODE) [146], and the Simple Event Model (SEM) [128]. Typically, these models differ by focus, i.e., event type, size, and

[52] https://www.w3.org/TR/hcls-swan/
[53] http://bibliontology.com/
[54] http://www.sparontologies.net/ontologies/biro
[55] http://www.sparontologies.net/ontologies/c4o

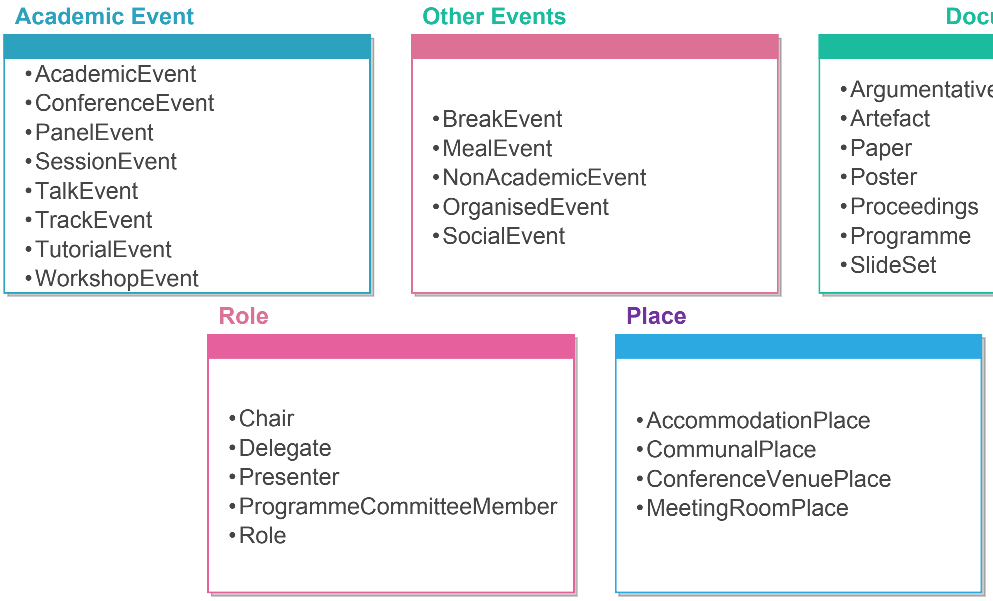


Figure 3.3: SWC terms grouped in broad categories.

level of abstraction. They focus on the description of event metadata, including time, location, and topical classifications of events. Early efforts towards events metadata modeling include the metadata projects of the ESWC 2006 and ISWC 2006 conferences [124], but they did not provide detailed descriptions of the events. Further trends are crowdsourcing and collaboration, open data, as well as big data analytics. These developments have profound effects on scholarly communication in all areas of science. In this section, we mainly focus on the well-known ontologies for describing scholarly events.

Semantic Web Conference Ontology (SWC)[56] is an ontology for modeling metadata of scientific conferences, such as a person affiliation during the event, meals, and social events. It was initially designed to support the ESWC conference (European Semantic Web Conference) in 2007 and later extended for both ESWC and ISWC (International Semantic Web conference) conferences. SWC comprises 59 classes and 19 properties which are categorized into two categories: 1) *Describing Papers*, which describe the paper itself, the authors and their affiliations, and the talk where the paper was presented, and 2) *Roles*, which uses role modeling for describing the different roles at a conference, such as Program chair. Figure 3.3 shows SWC terms grouped in broad categories. SWC imports several ontologies, including 1) *FOAF* for describing people, 2) *SWRC* for describing various elements of the papers, 3) *SIOC*[57] for describing the online community information, and 4) *Dublin Core* for describing paper's metadata.

Call for Papers ontology (Call) is a comprehensive ontology for CfP which takes DERI's CfP ontology[58] as a core ontology [147]. It defines the `call:Call` class as a subclass of the `cfp:Call`, the `call:Submission` class to represent different types of submissions, and the `call:Publication` class to represent published papers. The main objective for the Call ontology is to emphasize the logical consistency of concepts in the CfP ontology [147]. Call imports what SWC and SWRC import.

[56] `http://www.scholarlydata.org/ontology/doc/`

[57] `http://sioc-project.org/`

[58] `http://sw.deri.org/2005/08/conf/cfp`

Scholarly Event Description Ontology (SEDE) is a comprehensive ontology for describing scholarly events information, thus enabling software agents to process scholarly event data [79]. SEDE can be used to represent, collect, and share scholarly event data [79]. SEDE contains a set of properties used to express the content and structure of a scholarly event as an RDF graph. SEDE contains 39 classes with its own namespace, six FOAF classes representing agents (e.g., `foaf:Agent`, `foaf:Person`, and `foaf:Document`), two GEO [148] classes representing the place (`geo:SpatialThing` and `geo:Point`) and two SKOS [149] classes representing the topic (`skos:Concept` and `skos:ConceptScheme`). A distinguishing feature of SEDE is that it represents some concepts as logic rules to support the inference process. For example, consider the event *"'Said Fathalla' presented on 'EVENTSKG dataset' on 2019-06-06 in ESWC2019"*, it can be stated as:

$$(\exists Agent(SaidFathalla)) \wedge (\exists Action(present)) \wedge (\exists Entity(EVENTSKG)) \\ \wedge(\exists Event(ESWC2019)) \wedge \ (\exists Time(2019-06-06))$$

Event Ontology (EO)[59] is a simple ontology centered around four classes (Event, Agent, Factor, and Product) and 17 properties [145]. EO has been designed as a general ontology, and therefore it does not cover the domain knowledge specific to scientific events. EO ontology reflects the domain of events but does not cover more aspects related to scientific events and related entities, such as the roles of participants, sponsors, and publishers.

Linking Open Descriptions of Events (LODE)[60] is an ontology for describing historical events and mapping between other event-related vocabularies and ontologies, such as Time, EO and SKOS. In other words, it links people, places, or things to an event. Compared with EO, it has some restrictions and follows a higher level of abstraction. In the latest version of LODE (as of 2010), it contains only one class (`Event`) and only seven properties: `illustrates`, `inSpace`, `circa`, `atPlace`, `involved`, `involvedAgent` and `atTime`. Furthermore, LODE does not model the connection of agents to events through roles. It also does not cover entities related to scientific events, such as sponsors, publishers, and hosting organizations.

Simple Event Model ontology (SEM)[61] is an ontology for modeling event-related concepts, which is relatively close to EO and LODE, but still far from completeness, in terms of describing aspects related to scientific events, which do not exist in regular events, such as publishers. SEM is formalized purely in RDFS, describing the fundamental constituents of an event, including their types, roles, temporary validity, and the view according to which these constraints hold. SEM has four core classes: `Event`, `Actor`, `Place` and `Time` in addition to three types of constraints: `Role` (the role of an individual in a specific event), `Temporary` (defines the temporal boundary within which a property holds, e.g., the type of the place) and `View` (defines points of view).

3.4 Query Builders for Semantic Data

The origins of query builders go back to preliminary research works in the 1990s [150]. Natural language interfaces (NLIs) are widely used to ease the process of querying semantic data [151–154]. Many contributions have been made for this purpose. Below, we present some of the state-of-the-art efforts in SPARQL query building using NLIs. Most of these contributions use NLIs in two different ways: generating SPARQL queries based on User Interactions (UI) and answering

[59] `http://motools.sourceforge.net/event/event.html`

[60] `http://linkedevents.org/ontology/`

[61] `https://semanticweb.cs.vu.nl/2009/11/sem/`

Table 3.4: **Data Models Supporting Scholarly Communication.** State-of-the-art ontologies representing various types of scholarly data, including bibliographic data, scholarly articles components, and articles and scholarly events metadata.

Group	Ontology	Description	Lang.
Bibliographic	CiTO	Characterization of citations, both factually and rhetorically.	RDF/XML Turtle JSON-LD
	BiRO	Describe bibliographic records and references, and their compilation into bibliographic collections and reference lists.	
	C4O	Characterizing the number and contexts of bibliographic citations.	
	DataCite	Enables the metadata properties of the DataCite Metadata Schema Specification.	
	BIBO	Describe the main concepts and properties for describing citations and bibliographic references.	RDF/XML
Scholarly Article	FaBiO	Describe entities that are published or potentially publishable (e.g., journal articles, conference papers, books), and that contain or are referred to by bibliographic references.	RDF/XML Turtle JSON-LD
	DoCO	Describe the component parts of a bibliographic document.	
	SA	Describe the production process, content and preservation of scholarly article.	RDFS JSON-LD
	SWRC	Model entities of research communities.	RDF/RDFS OWL
	EXPO	Describe scientific experiments for both humans and computer systems.	OWL
	FRBR	Describe documents and their evolution.	
Articles Metadata	SCoRO	Describe the contributions that may be made, and the roles that may be held by a person with respect to a journal article or other publication.	RDF/XML Turtle JSON-LD
	FRAPO	Describe the administrative information of research projects.	
	BiDO	Provide classification of authors and journals according to bibliometric data.	
	FiveStar	Enable characterization of the attributes of an online journal article.	
	PRO	Describe roles in the publication process, or in other scholarly activities or situations, held by a particular agent.	
	PSO	Describe the publication status of documents at each stage of the publishing process.	
	PWO	Describe the workflow associated with the publication of a document.	
ScholarlyEvents	SWC	Model knowledge about conferences.	RDF/XML
	Call	Describe call for papers.	
	SEDE	Describe scholarly event information.	
	LODE	Describe historical events.	
	SEM	Describe aspects related to scholarly events, which do not exist in regular events, such as publishers.	

user queries using a Question-Answering system (QA). The latter completely hides SPARQL queries from the user, allowing them to directly submit their question, e.g., AquaLog [155], NLP-Reduce [156] and PowerAqua [153], whereas the former focuses on generating SPARQL queries using a visual interface, e.g., Semantic Crystal [152], Querix [151], and SPARQL Views [154].

UI-based systems. NLI-based systems are often tailored to a specific application and require exceptional design and implementation efforts. Semantic Crystal [152] is a graphical-based query tool that can be used for querying OWL knowledge bases by generating SPARQL queries. The generated query is composed by clicking on ontology elements from the ontology graph displayed on a screen and selecting items from menus. In the end, the SPARQL query is generated and executed (using Jena[62]), and the result is displayed to the user in a new tab. Querix [151] is an NLI-based tool that translates natural language questions, written in English, to SPARQL queries with little user interactions. WordNet is used to extract synonyms of the words in the input query to improve the matching between them and the elements in a knowledge base. One drawback of Querix is that it does not resolve ambiguities in the input text but asks the user for clarification. Live SPARQL auto-completion [157] is a SPARQL auto-complete library that suggests recommendations of possible RDF terms, such as predicates, classes, or named graphs, with regard to the current state of the query. SPARQL Views [154] is an NLI-based tool that supports visual query building via drag and drop over RDF data in a Drupal CMS[63]. Via an auto-complete search box, users can filter predicates, which can be used in the query pattern. QUaTRO2 [158] provides a graphical user interface to formulate complex queries based on an abstract domain-driven query formulation. QUaTRO2 tool has been used to query the UniProt[64] protein database. QueryVOWL [159] is a visual query language tool for creating SPARQL queries using GUI controls. This tool is developed based upon the VOWL [160] ontology visualization.

QA-based systems. AquaLog [155] is an ontology-driven QA system, which takes a query expressed in natural language and the ontology being queried and returns results after making sense of the terms and relations in the input query using the input ontology and lexical databases, such as WordNet. PowerAqua [153] is a QA-based system that takes a query in natural language, translates it into a set of logical queries, and then it retrieves the results by aggregating information derived from multiple heterogeneous semantic sources. PowerAqua is the extension of AquaLog, which overrides the main limitation of AquaLog, since AquaLog can only be used for only one particular ontology, by providing the ability to retrieve the results from multiple heterogeneous data sources. NLP-Reduce [156] is a domain-independent NLI for querying semantic data. It takes an OWL knowledge base and a natural language query as a bag of words and applies several conventional NLP techniques; then, it identifies triple structures in the rest of the query words. Afterward, these triples are merged and translated into SPARQL statements. Jena is used for executing the SPARQL query, and the results are displayed to the user with a desktop interface.

Much of the current work on building SPARQL queries, which tends to focus on translating natural language queries to SPARQL, is still far from efficient. Therefore, more work is needed regarding the use of NLIs for facilitating the process of querying distributed semantic data, for both end-users and SPARQL experts, and further comprehensive usability studies to investigate the end-users' perspective are required.

[62] `http://jena.sourceforge.net/`

[63] `https://www.drupal.org/`

[64] `https://www.uniprot.org/`

CHAPTER 4

Quality Assessment of Scholarly Events

Scholarly events have numerous characteristics that can be used to assess their quality. Nevertheless, there are no standard metrics to assess the quality of scholarly events. Researchers identify high-quality events in different ways, depending on their perspective or experience. On the one hand, some researchers found events interesting based on the reputation of the publisher and the organizers. On the other hand, other researchers have another perspective (financial view), which is the proximity of the location in order to reduce travel expenditures. In some cases, researchers can decide whether it is worth to submit their work to particular events by asking experts in the community. Therefore, there is a need to develop a robust metrics suite to assess the quality of the event, which is essential for both researchers and publishers. This suite helps the former to decide whether to submit their research work, give a keynote speech, or accept an invitation to be a program committee member, while it helps the latter to decide whether to publish the proceedings of the event.

In chapter 3, we reviewed the research efforts that have been carried out to study the characteristics of renowned scholarly events in Computer Science as well as ranking services. We observed that the characteristics of these events had not been comprehensively studied. Hence, this shortcoming has led us to conduct an in-depth analysis (presented in this chapter), which is based on a comprehensive list of metrics, considering quality in terms of event-related metadata in eight communities within Computer Science and also events from other communities, such as Physics, Mathematics, and Engineering.

In this chapter, we investigate the problem of study the various characteristics of scholarly events in different fields of science to assess their impact. Section 4.1 presents the Scholarly Events Quality Assessment (SEQA) metrics suite, which contains ten generic metrics that can be jointly applied for assessing the quality of scholarly events. Section 4.2 presents the Scholarly Events Metadata Analysis methodology (SEMA). Section 4.3 presents an analytical study of the evolution of key characteristics of the renowned Computer Science scholarly events over the last five decades. Section 4.4 presents a study of the various characteristics of scholarly events in four fields of science, namely Computer Science, Physics, Engineering, and Mathematics, using the SEQA (cf. section 4.1). In Particular, we analyzed renowned scholarly events of five communities within computer science, namely the World Wide Web (WEB), Computer Vision (CV), Software Engineering (SE), Data Management (DM) as well as Security and Privacy (SEC). This analysis is based on a systematic approach using descriptive statistics as well as exploratory data analysis. The following research question is investigated in this chapter:

RQ1: How can the characteristics of the renowned scholarly events in different fields of science be utilized to assess their impact?

The work presented in this chapter is based on the following publications:

- **Said Fathalla**, Sahar Vahdati, Christoph Lange, and Sören Auer. *Analysing Scholarly Communication Metadata of Computer Science Events.* In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 342-354. Springer, Cham, 2017.
- **Said Fathalla**, Sahar Vahdati, Sören Auer, and Christoph Lange. *Metadata Analysis of Scholarly Events of Computer Science, Physics, Engineering, and Mathematics.* In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 116-128. Springer, Cham, 2018.
- **Said Fathalla** and Christoph Lange. *EVENTS: A Dataset on The History of Top-Prestigious Events in Five Computer Science Communities.* In Proceedings of Semantics, Analytics, Visualization (SAVE-SD) at the World Wide Web conference. Springer, Cham, pp. 110-120, 2017.
- **Said Fathalla**, Sahar Vahdati, Sören Auer, and Christoph Lange. *Scholarly Event Characteristics in Four Fields of Science: A Metrics-based Analysis.* Scientometrics 123, 677–705 (2020).

4.1 Scholarly Events Quality Assessment Suite

Providing statistics about the impact of scholarly objects and measurements of the quality of research has been recently accelerated [161]. Community-defined criteria, such as citation-related measurements, distinguish highly-ranked instances of publication channels, e.g., scholarly events and journals. Nevertheless, such criteria are not standardized nor centralized but generally transferred subjectively from seniors to juniors. To go beyond citation-related measurements, an extended list of metrics is required. Analyzing scholarly events metadata, such as event dates, the number of submitted and accepted articles, location, event type, and field, can help to answer such questions. In order to systematize the evaluation, we define the Scholarly Events Quality Assessment Suite metrics suite (SEQA), which contains ten generic metrics that can be jointly applied for assessing the quality and identifying the characteristics of high-quality scholarly events within scientific communities. SEQA contains numeric values metrics, such as average acceptance rate, and complex data types metrics, such as geographical distribution with two different levels of granularity. The aim is to study various characteristics of high-quality events in different fields of science, which in turn can be used to assess the quality of those events.

The proposed metrics suite can be used in further events" metadata analysis and for multi-criteria events ranking. These metrics can also be used to compare scholarly events within their communities. Figure 4.1 depicts the ten metrics of the SEQA suite. In the rest of this section, the formalization of these metrics is presented.

Definition 1 (Average Acceptance Rate) *The acceptance rate of an event is defined as the ratio between the number of accepted articles and the number of submitted ones. The Average Acceptance Rate (AAR) for an event series is the average of the acceptance rate of all editions of this series.*

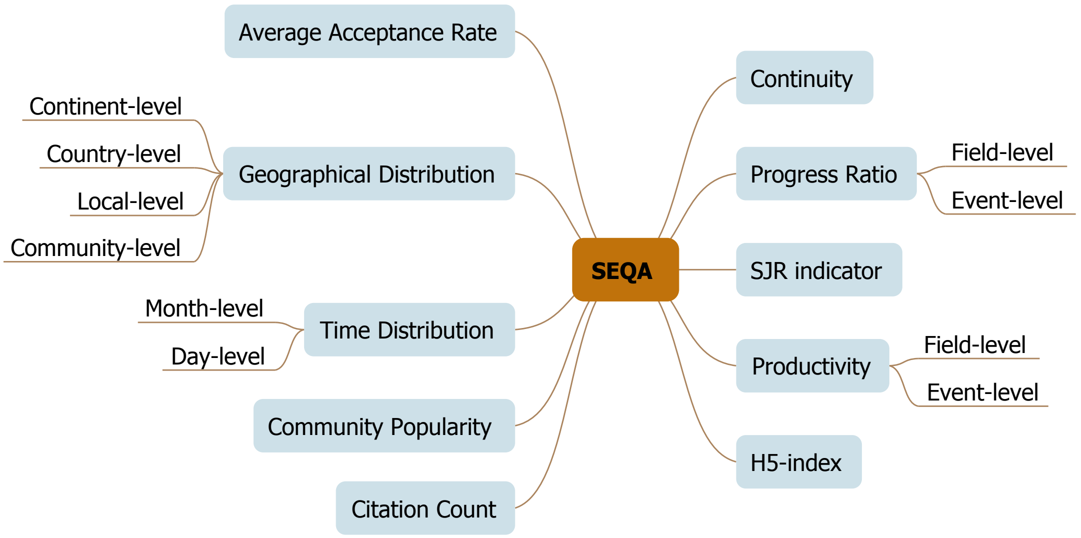


Figure 4.1: **SEQA Suite.** The SEQA suite containing ten metrics (with different granular levels) for scholarly events' impact assessment.

AAR can be used to get an overview of the overall acceptance rate of a particular event series since its beginning. Many researchers believe that a low acceptance rate, for example, less than 25%, indicates a good quality of the accepted submissions when the number of submissions is large [162]. For example, the acceptance rate of the CHI conference (Conference on Human Factors in Computing Systems), the most renowned conference in the HCI community, in 2017 is 25%, out of 2400 submissions. In most cases, information about the acceptance rate of events is not available. The acceptance rate of an event can vary, but most of the renowned events try to keep it stable, with a slight change either with an increase or decrease from a year to the next.

Definition 2 (Continuity) *The continuity (Cont) of an event reflects the continuation of an event series since its beginning.*

A new formula (shown in Equation 4.1) is proposed to compute the percentage of the continuity for a specific event, where *Cont* is the continuity, E is the number of editions of the event, R is the regularity of the event editions, e.g., $R = 2$ for events which take place every two years, and A is the age, i.e., the number of years since the event took place first. For events that changed their regularity after specific periods, e.g., from 2 to 1, we computed the continuity of both periods, and the overall continuity is the direct average of the continuities calculated in these periods. The year is the granularity for this metric.

$$Cont = \min\left\{100\%, \frac{E * R}{A}\right\} \tag{4.1}$$

The high continuity of an event is an indicator of the success of that event, which means that this event attracts submissions and participants in every edition.

Definition 3 (Geographical Distribution) *The Geographical Distribution (GD) metric measures the number of distinct locations visited by an event each year since its beginning.*

For the finer granularity GD, one can derive the number of distinct locations, either city (local-level GD), country (country-level GD), or continent (continent-level GD), visited by a particular event.

Definition 3a (Local-level GD) *refers to the change of the location of local events from year to year in the same country and denoted by ΔL_n (Equation 4.2), where l_n is the location of an event in a year and l_{n+1} is the location of the next edition.*

$$\Delta L_n = \begin{cases} 1 & \textit{if } l_n \neq l_{n+1} \\ 0 & \textit{otherwise} \end{cases} \tag{4.2}$$

Then, the mean of these changes ($\bar{x}$) is computed to measure the rate of the distribution of each event since the beginning (Equation 4.3), where N is the number of editions. The higher this value is for an event, the more frequently the location of an event changed.

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N-1} \Delta L_i \tag{4.3}$$

Definition 3b (Country-level GD) *refers to the change of the country of international events from year to year (analogously to Definition 3a). Then, the mean of these changes ($\bar{x}$) is computed to measure the rate of the distribution of each event since the beginning (Equation 4.3).*

Definition 3c (Continent-level GD) *refers to the change of the continent of international events (i.e., the country) from year to year (analogously to Definition 3a). Then, the mean of these changes ($\bar{x}$) is computed to measure the rate of the distribution of each event since the beginning (Equation 4.3).*

Definition 3d (Community-level GD) *refers to the frequency of occurrence of the editions of a particular event series belonging to a particular community among continents since the beginning.*

First, the frequency of occurrence f_{ij} of all events belonging to community i took place in continent j is computed. Then, these values are normalized to q_{ij} to ensure that the frequencies of occurrence of events in each community (C) sum up to one (Equation 4.4).

$$f_{ij} = \sum_{k \in C} E_{ijk} \quad , \quad Community - level\ GD = q_{ij} = \frac{f_{ij}}{\sum_{x \in C} f_{xj}} \tag{4.4}$$

Here, E_{ijk} is the number of events of an event series k in a community i took place in continent j.

The key questions that can be answered by this metric are 1) which countries hosted most of the events of a series in the dataset and 2) how frequently a country has hosted an event during a given period. We map every distinct location to the number of times the event has taken place there (by city, country, or continent). We can thus classify event series by their most frequent location, e.g., as a “German” or “European” series. The Geographical Distribution of an event series increases the awareness of researchers about the existence of the event and its covered topics; on the other hand, holding events in an expensive or very distant place, e.g., Auckland, New Zealand for ISWC 2019, discourages researchers with a low budget from submitting their work due to the high costs of travel expenses and accommodations. Thus, we can determine which country pays more attention to a particular type of event in terms of the field of research.

Definition 4 (Time Distribution) *The Time Distribution (TD) of an event refers to the time of the year in which the event takes place.*

For the finer granularity of this metric, one can derive the distribution of events either among months of the year (Month-level TD), or the change of the start date of the event (Day-level TD).

Definition 4a (Month-level TD) *Month-level TD of an event refers to the month of the start date of the event in which the event takes place. The standard deviation (σ) is computed in order to quantify the variation or dispersion of the month in which an event takes place each year. The standard deviation is computed using Equation 4.5, where N is the number of editions, x_i is the month (the numeric value) in which the event has been held, and $\bar{x}$ is the mean value.*

$$TD_{month} = \sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{4.5}$$

It is crucial for researchers who are interested in a particular event to be able to estimate when a particular event will be held next time, which helps to devise a submission plan. Namely, EMC (International Symposium on Electromagnetic Compatibility) has been held 17 times in August, and only three times in July, therefore, it has a low standard deviation of $\sigma = 0.4$, while NDSS (Annual Network and Distributed System Security Symposium) has been held in February every year since the beginning, therefore $\sigma = 0$.

Definition 4b (Day-level TD) *Day-level TD of an event refers to the change of the starting day of an event to the subsequent edition. The absolute differences between subsequent editions are computed (Equation 4.6). The standard deviation (σ) is computed in order to quantify the variation or dispersion of the starting day in which the event takes place each year. The standard deviation is computed using Equation 4.5, where N is the number of editions, x_i is the absolute difference between subsequent editions, and $\bar{x}$ is the mean value.*

$$\Delta D_n = (|d_n - d_{n+1}|) \tag{4.6}$$

Definition 5 (Community Popularity) *The Community Popularity (CP) of an event reveals how popular an event is in a research community, in terms of the number of submissions.*

The more submissions to events of a particular field, the more popular is this field. Several renowned events, such as VLDB, publish statistics about the number of submitted and accepted papers after each edition either on their Web pages or in the preface of the published proceedings. This metric also gives an indication about which community attracts a large number of researchers. Therefore, it helps in assessing the comparative popularity of scientific communities.

Definition 6 (Productivity) *Productivity measures how productive is an event or a research field in terms of the number of publications.*

For the finer granularity of this metric, the productivity of an event compared to competitive events in the same field (Event Productivity) and the productivity of a research field (Field Productivity) compared to other subfields in the same community can be computed as follows.+

Definition 6a (Event Productivity) *Event Productivity (EP) of a particular event is the ratio between the publications of that event since the beginning and the total number of publications of all editions of event series in the same research field.*

Event Productivity reveals how productive, in terms of the number of publications, an event compared to other events in the same fields. This metric helps to position an event as a productive event, among other events in the same field. The event productivity ($EP(e)$) of an event e belonging to a research field f is defined in Equation 4.7, where N is the number of editions of an event series e, and $P_i(e)$ is the number of publications of the i^{th} edition of e.

$$EP(e) = (\sum_{i=1}^{N} P_i(e))(\sum_{e \in f} \sum_{i=1}^{N} P_i(e))^{-1} \tag{4.7}$$

Definition 6b (Field Productivity) *The Field Productivity (FP) of a given research field is the ratio of the publications of all events in this field to the total number of publications of all events in the dataset.*

Field Productivity reveals how productive, in terms of the number of publications, a research field in a given year within a particular period is compared to other fields. The field productivity for a research field (f) is defined in Equation 4.8, where N is the number of editions of an event series e, F is modeled as the set of all research fields exist in the dataset, and $P_i(e)$ is the number of publications of the i^{th} edition of e. For instance, in the period 2007–2016, the FP of computer vision community is 22%, where the total number of computer vision publications of all events is 3,022, and the total number of publications of all events from all fields is 13,701.

$$FP(f) = (\sum_{e \in f} \sum_{i=1}^{N} P_i(e))(\sum_{f \in F} \sum_{e \in f} \sum_{i=1}^{N} P_i(e))^{-1} \tag{4.8}$$

Definition 7 (Progress Ratio) *The Progress Ratio (PR) refers to the relative amount of publications of an event or field in a fixed time period.*

For the finer granularity of this metric, the progress of an event compared to its previous years (Event-level PR) and the progress of a research field compared to other subfields in the same community can be computed as follows.

Definition 7a (Event-level PR) *Event-level PR (EPR) of an event is the ratio of the publications of an event in a given year to the total number of publications by all editions of that event since the beginning.*

This metric sketches the progress of an event each year within a particular time span. Thus, it gives events' organizers an overview of the progress of their event compared to other competitive events in the same field. The EPR of an event e in a year y is defined in Equation 4.9, where $P_y(e)$ is the number of publications of e in y and N is the number of editions of e.

For instance, in 2017 ECCV (European Conference on Computer Vision) has a pretty large progress ratio of 19% (compared to other events in the dataset), while VLDB (International Conference on Very Large Data Bases) and EDBT (International Conference on Extending Database Technology) have the lowest PR of around 4%.

$$EPR_y(e) = P_y(e)(\sum_{i=1}^{N} P_i(e))^{-1} \tag{4.9}$$

Definition 7b (Field-level PR) *The Field-level PR (FPR) of a research field is the ratio between the total number of publications of events belonging to this field in a given year and the total number of publications of this field since the beginning.*

This metric sketches the scientific progress of a research field, e.g., Computer Security, among other fields within the same community, i.e., Computer Science. Thus, it gives an indication of the research development in these fields, i.e., how active are the researchers in these fields.

The FPR of a field f in a given year y is defined in Equation 4.10, where $P_y(e)$ is the number of publications of e in y, and N is the number of editions of e.

$$FPR_y(f) = (\sum_{e \in f} P_y(e))(\sum_{e \in f} \sum_{i}^{N} P_i(e))^{-1} \tag{4.10}$$

Definition 8 (Citation Count) *The Citation Count (CC) of an event is the number of citations that papers published by that event receive by other published articles.*

For example, the citation count of ICML (International Conference on Machine Learning) is 1,583, which means that the number of citations in 2018 received by articles published in 2015, 2016, and 2017 is 1,583.

Definition 9 (SCImago Journal Rank indicator) *The SCImago Journal Rank (SJR indicator) of an event is the average number of citations received in a particular year by the papers published by an event in the three previous years [102].*

SJR is a measure of the scientific influence of scholarly journals and events based both on the number of citations received by a journal or event proceedings and the prestige of the journals/events where such citations come from [102]. For instance, the SJR of JCDL (ACM/IEEE Joint Conference on Digital Libraries) in 2018 is 0.26.

Definition 10 (H5-index) *It is the highest number h such that h articles published in the past five complete years have at least h citations each.*

Usually, high-ranked events have a high h-index. For instance, the h5-index of CVPR (Conference on Computer Vision and Pattern Recognition) is 240 and of NIPS (Neural Information Processing Systems conference) is 169.

By using these metrics, it is possible to provide a flexible and broad study on various characteristic dimensions of scholarly events in different fields of science.

4.2 Scholarly Events Metadata Analysis Methodology (SEMA)

This section presents the Scholarly Events Metadata Analysis methodology (SEMA). SEMA is a novel methodology for building knowledge graphs of scholarly events with the purpose of identifying the characteristics of renowned events and providing a recommendation to various stakeholders in the scholarly communication domain. Figure 4.2 depicts the initial version of SEMA (i.e., SEMA 1.0), which comprises five phases: (1) Data gathering, (2) Identification of prestigious events, (3) Data preprocessing, (4) Data analysis and visualization, and finally, (5) Conclusions.

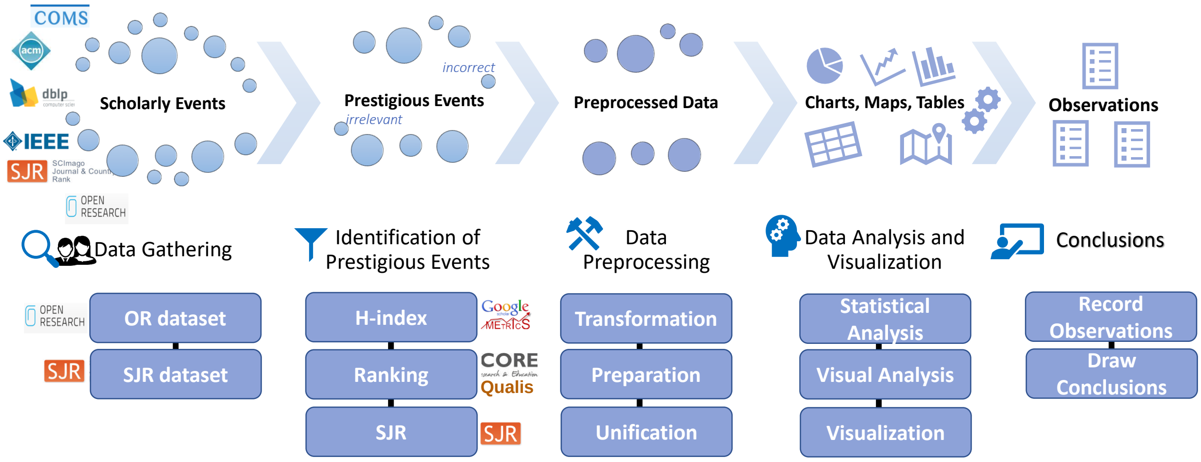


Figure 4.2: **SEMA 1.0**. The initial version of Scholarly Events Metadata Analysis Methodology, beginning by data gathering until recording observations and drawing conclusions.

Actually, SEMA 1.0 has been followed at the early beginning of the analysis of scholarly events in Computer Science, Physics, Mathematics, and Engineering (see section 4.4). Afterward, we realized that the initial version of SEMA has some deficiencies; therefore, we realized that it should be adapted for the following reasons. First, some steps should be overlapped, i.e., executed in parallel. For example, while executing the methodology, data gathering and the identification of relevant events were performed simultaneously, whereas while collecting data, we found some renowned events that had not been considered from the beginning; therefore, we began to collect the metadata of these events as well. Second, new sources for events metadata, such as the number of submissions and accepted papers, have been added, including Digital Libraries and the SCImago dataset. Third, the outcomes of the analysis can be used to derive a set of recommendations, which does not exist in SEMA 1.0, for event organizers, potential authors, proceedings publishers, and sponsors. Consequently, we proposed a new version of SEMA, i.e., SEMA 2.0, which is adapted from the initial version. SEMA 2.0 comprises five phases (see Figure 4.3): data gathering, identification of renowned events, data preprocessing, data analysis and visualization, and recording observations and drawing conclusions.

A series of challenges have been encountered in these steps, such as data duplication, incomplete data, incorrect data, and the change of event titles over time. Therefore, *Data Curation* methods as a set of activities related to organization, integration, publication, and annotation of the data [163] have been applied to ensure that the data is fit for the intended purpose, and reuse. Data analysis and recording observations were also executed in parallel. Further details about each step are given in the following sections.

4.2.1 Data Gathering

Data gathering is the process of collecting data from a variety of online sources in an objective manner. In this phase, relevant metadata of various scholarly events are collected, involving conferences, workshops, symposiums, and meetings. Scholarly events metadata can be available as Linked Data in the case of DBLP, and other structured forms, or semi-structured and unstructured in the case of WikiCFP, or conference.city[1]. This metadata can be found in several

[1] `http://www.conference.city/`

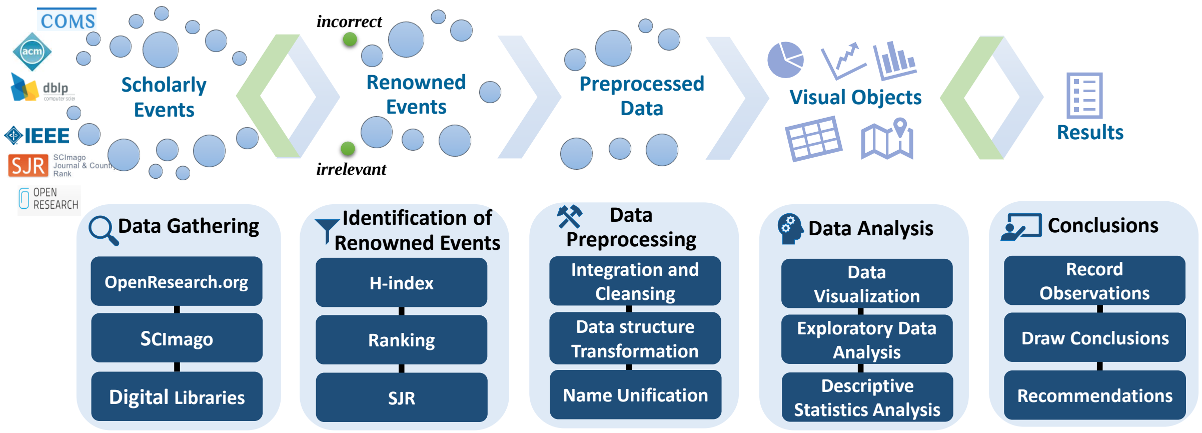


Figure 4.3: **SEMA 2.0**. The adapted version of SEMA 1.0. The data gathering and identification of renowned events have become overlapped and recommendations for scholarly events stakeholders are provided.

data sources, such as

- The *OpenResearch.org* wiki[2] serves us both as an additional source of semantically structured data, and as a tool to support data analysis. In OpenResearch.org, information is represented in wiki pages with semantic annotations, with the possibility to be queried via a SPARQL endpoint as well as inline queries embedded into wiki pages Currently, OpenResearch.org contains crowd-sourced metadata about more than 6,100 conferences, 1,000 workshops, and 379 event series (as in November 2019).
- *SCImago* is an online database that contains information about journals and event proceedings since 1996 for 27 different research fields, including computer science, physics, engineering, and mathematics. Journals or event proceedings can be grouped by research field, sub-fields, or by country [164].
- Digital Libraries, such as the ACM digital library[3], is one of the primary sources of scholarly metadata, where they publish the proceedings of the events.

4.2.2 Identification of Renowned Events

To identify renowned events from a large number of scholarly events available nowadays, we used the following metrics (see subsection 3.1.1 for more details), which are commonly used to identify high-quality events in each field of science.

4.2.3 Data Preprocessing

The main objective of the data preprocessing phase is to 1) fill in missing data, 2) identify incorrect data, 3) eliminate irrelevant data, and 4) resolve inconsistencies. In order to prepare the raw data for analysis, we carried out four preprocessing tasks: *data integration*, *data cleansing*, *data transformation* and *name unification*.

[2] http://openresearch.org

[3] http://dl.acm.org/

- **Data integration:** involves combining data from multiple sources into meaningful and valuable information. Also, this process involves eliminating redundant data, which might occur during the integration process.
- **Data cleansing:** focuses on getting rid of incorrect or inaccurate records. For instance, some websites provide incorrect information about events' submissions and accepted papers. We verify this information against the official websites of the events or proceedings published in Digital Libraries.
- **Data structure transformation:** involves converting cleaned data values from unstructured formats into a structured one. For instance, data collected from websites of the events as text (i.e., unstructured format) is manually transformed to CSV (i.e., structured format) and subsequently to RDF.
- **Name unification**: involves integrating all individual events of a series with multiple names under its most recent name because it is important for the researchers who want to submit their work to know the recent name rather than the name that had been in use for the longest time. The rationale for name unification is that we observed that some events had changed their names once or several times since they had been established. The change sometimes happens because of changing the scale of the event to a larger scale, e.g., from Symposium to Conference or from Workshop to Symposium, such as ISWC and ISMAR, respectively. Also, the change sometimes happens because of adding a new scope or topic, such as SPLASH. Besides, conferences such as SPLASH keep both names, the old and the new one. In this case, we also keep the most recent name. More examples are listed in Table 4.2

4.2.4 Data analysis

Data analysis is defined as the process of studying and modeling data with the purpose of discovering useful information from the data in order to support decision-making effectively. Usually, data is collected and analyzed to answer questions, confirm/falsify assumptions, or test hypotheses [165]. Generally, the data analysis process can be categorized into three categories: Descriptive Statistics Analysis (DSA), Exploratory Data Analysis (EDA) and Confirmatory Data Analysis (CDA).

- *Descriptive Statistics Analysis* describes the main aspects of the data being analyzed by providing a summary statistic that quantitatively summarizes the features of this data [166].
- *Exploratory Data Analysis* is an approach for data analysis that explores the new features and unknown relationships in the data to provide future recommendations [167]. Besides, it summarizes the main characteristics of the data being analyzed often by producing visual objects, such as charts, graphs, and maps.
- *Confirmatory Data Analysis* aims to understand data through hypothesis testing to confirm or falsify existing hypotheses [167].

4.2.5 Conclusions

At the end of SEMA, the Conclusions phase comes in which outstanding observations are recorded, conclusions based on the analysis results are reported, and a set of recommendations out of these conclusions is provided.

4.3 Metadata Analysis of Scholarly Events in Computer Science

The mega-trend of digitization affects all areas of society, including business and science. Digitization is accelerated by ubiquitous access to the Internet, the global, distributed information network. Data exchange and services are becoming increasingly interconnected, semantics-aware, and personalized. Further trends are crowdsourcing and collaboration, open data, as well as big data analytics. These developments have profound effects on scholarly communication in all areas of science. In this section, the evolution of key characteristics of scientific events over time has been analyzed, including frequency, geographic distribution, and submission and acceptance numbers, with a particular focus on Computer science events, where conferences and workshops are of paramount importance and a primary means of scholarly communication. The method of choice for this study is a meta-analysis that refers to the statistical methods used in research synthesis for drawing conclusions and providing recommendations from the results obtained from multiple individual studies.

The preliminary work of analyzing the metadata of Computer Science events was conducted in 2017, in which we analyzed 40 conference series in computer science with regard to these indicators over a period of 30 years. Our analysis methodology was based on descriptive statistics analysis, exploratory data analysis, and confirmatory data analysis. A key question is: What were the measurable effects of digitization on scholarly communication? For instance, due to the Coronavirus disease (COVID-19) crisis raised at the beginning of 2020, digitization played a vital role in the switching of scholarly conferences from physical to virtual conferences in which participants can attend and present their work remotely, helping to restrain the virus prevalence. Of particular interest, we analyzed the effect of digitization in computer science events by answering the following questions:

- Did the number of submissions increase?
- Is there a proliferation of publications?
- Can we observe popularity drifts?
- Which events are more geographically diverse than others?

We shed light on these questions by analyzing comprehensive metadata of Computer Science scholarly events in the last 30 years. Extensive collections of such data are nowadays publicly available on the Web. Research has recently been conducted to browse and query such data [168, 169], with a focus on authors, publications, and research topics [170].

4.3.1 Method

The preliminary work of analyzing Computer Science scholarly events has been carried out following the methodology shown in Figure 4.4, which comprises four steps: (1) Identification of relevant events, (2) Data gathering, (3) Ingestion into the OpenResearch.org semantic scholarly communication data curation platform, and (4) Data analysis and visualization.

Identification of Relevant Events

A sample of the identified renowned Computer Science events, based on CORE, H5-index of both GSM and AMiner, and Qualis, is listed in Table 4.1.

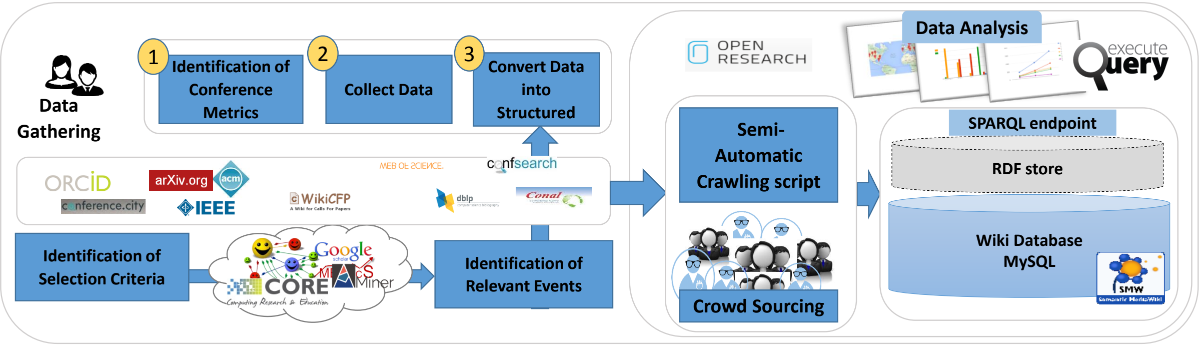

Data Gathering
1
Identification of Conference Metrics
2
Collect Data
3
Convert Data into Structured
Identification of Selection Criteria
Identification of Relevant Events
OPEN RESEARCH
Data Analysis
execute Query
Semi-Automatic Crawling script
Crowd Sourcing
SPARQL endpoint
RDF store
Wiki Database MySQL

Figure 4.4: **Scholarly events analysis methodology.** The overall workflow of analyzing Computer Science scholarly events starting from identification of relevant events till visualizing the results.

Data Gathering

We collected metadata about 40 conference series in different Computer Science communities from various sources of metadata, including title, event series, field, start date, end date, homepage, country, and Twitter account.

Data Preprocessing

In this step, we carried out several preprocessing tasks in order to prepare the collected data for subsequent analysis (see subsection 4.2.3) After preparing the data, it has been ingested into OpenResearch.org in several ways using either single or bulk import. For a single import, one can use semantic forms[4]. The required steps for bulk import are: (1) Create a spreadsheet with important information, (2) Export the spreadsheet to CSV, and (3) import CSV file using OpenResearch's *ImportCSV* service[5].

Data analysis

The heart of this analysis is an exploratory analysis of the metadata of selected Computer Science events (see Table 4.3) over the past 30 years. Statistical metrics over numeric values have been defined, as well as metrics having other complex data types, focusing on scholarly events, such as conferences, because of their high impact on research communities rather than smaller-scale ones, such as workshops. We chose spreadsheets as the primary tool to compute statistical metrics over numeric values; charts support the evaluation of the results. OpenResearch.org provides further components for visual analytics, in particular for displaying non-numeric results (e.g., the conferences with the highest number of submissions). Even though spreadsheets are, in principle, based on the relational data model, they practically lack support for joins across sheets. Joins may be required for connecting information about events to information about related entities, such as persons participating in events. The SPARQL query language for RDF, which is supported by OpenResearch, facilitates such join computations. However, while SPARQL also supports fundamental statistical analysis via aggregate functions, this type of analysis is better supported by spreadsheets. Table 4.3 shows research communities and corresponding conferences investigated.

[4] https://www.openresearch.org/wiki/Special:FormEdit/Event

[5] https://www.openresearch.org/wiki/Openresearch:HowTo

Table 4.1: **Renowned Computer Science Events Rankings.** A sample of the renowned Computer Science events based on CORE, H5-index of both GSM and Aminer, and Qualis.

Conference	**Core**	**H5 Index**		**Qualis**
	Rank	**GSM**	**Aminer**	
CAV	A*	39	35	A1
CHI	A*	83	71	A1
COLT	A*	22	20	A2
CRYPTO	A*	51	24	A1
DSN	A*	16	21	A1
EuroCrypt	A*	50	30	A1
FOCS	A*	47	31	A1
ICCV	A*	92	58	A1
ICFP	A*	28	21	A2
ISCA	A*	50	36	A1
KR	A*	23	24	A2
NIPS	A*	83	51	A1
OOPSLA	A*	36	28	A1
PERCOM	A*	28	24	A2
PLDI	A*	45	36	A1
PODC	A*	25	20	A1
POPL	A*	48	36	A1
SIGGRAPH	A*	21	–	–
WWW	A*	74	66	A1

Entity-centric Visual Analytics. In contrast to spreadsheets and their charting facilities, OpenResearch makes it easy to generate visualizations that focus on entities rather than numbers. Besides geographical maps and ranked tables or lists, timelines are a prominent example of entity-centric visualizations. The input for a timeline is provided by a query in the Semantic MediaWiki inline queries[6]. Semantic MediaWiki is an extension of MediaWiki[7], an open-source wiki engine that powers Wikipedia. It utilizes semantic technologies to improve the usage of MediaWiki by addressing core its problems, involving consistency of content, finding and comparing information from different pages, and reusing knowledge [171]. Listing 4.1 defines a timeline of events with upcoming submission deadlines.

```
{{#ask: [[Category:Event]] [[submission deadline::>{{CURRENTYEAR}}...]][[Category:{{#urlget:field}}]]
 | ?title = Name                          | ?abstract deadline
 | ?submission deadline                   | ?notification
 | ?Category:Conference = Conference      | ?Category:Workshop   = Workshop
 | format=timeline                        | sort=submission deadline
 }}
```

Listing 4.1: **ASK query** for displaying a timeline of events with upcoming submission deadlines.

[6] https://www.semantic-mediawiki.org/wiki/Help:Inline_queries

[7] https://www.mediawiki.org/wiki/MediaWiki

Table 4.2: **Event Name Unification.** Conference title and acronym evolution for a sample of renowned events.

Unification	Acro.	Full title	Time Span
IEEE VR	IEEE VR	IEEE Virtual Reality	1999–2017
	VRAIS	Virtual Reality Annual International Symposium	1993–1998
ASE	ASE	Automated Software Engineering	1997–2017
	KBSE	Knowledge-Based Software Engineering Conference	1990–1996
ISWC	ISWC	International Semantic Web Conference	2002–2017
	SWWS	Semantic Web Working Symposium	2001
FOCS	FOCS	Annual Symposium on Foundations of Computer Science	1975–2017
	SWAT	Annual Symposium on Switching and Automata Theory	1966–1974
	SWCT	Annual Symposium on Switching Circuit Theory and Logical Design	1960–1965
ISMAR	ISMAR	International Symposium on Mixed and Augmented Reality	2002–2017
	ISAR	International Symposium on Augmented Reality	2000–2001
	IWAR	International Workshop on Augmented Reality	1999
ISSAC	ISSAC	International Symposium on Symbolic and Algebraic Computation	1988–2017
	SYMSAC	Symposium on Symbolic and Algebraic Manipulation	1966, 1971, 1976, 1981, 1986
	EUROSAM	International Symposium on Symbolic and Algebraic Computation	1974, 1979, 1982 and 1984
SPLASH	SPLASH	Systems, Programming, Languages and Applications: Software for Humanity	2010–2017
	OOPSLA	Conference on Object-Oriented Programming, Systems, Languages, and Applications	1986–2009

Similar types of queries that we have implemented in OpenResearch.org include:

- event series in a given field and their average acceptance rates,
- countries with a high number of events in a given field,
- fields with decreasing numbers of accepted papers over the years,

Joins Across Entity Types. Currently, *OpenResearch.org* focuses on the semantic representation of CfPs as one wiki page per event, but including semantic relations to related entities, e.g., to document the role that a person had in the organization of an event. A concrete use case for querying this data is supporting the research community in taking decisions on what conference to submit one's results to, or whether to accept invitations for assuming certain roles in the organization of a particular event. Such queries often require joins across multiple entity types.

Table 4.3: **Research Communities.** Distribution of top events involved in the study among five computer science research communities.

Acronym	Community	Events
GRA	Computer Graphics	ACMMM, EuroGraphics, IEEE VR, SIGGRAPH
SEC	Computer Security	CCS, CRYPTO, EuroCRYPT, ASIACRYPT
PROG	Programming Languages	ICFP, PLDI, POPL, SPLASH
SE	Software Engineering	ICSE, FSE, ASE, FASE
DB	Database Systems	PODS, SIGMOD, ICDT, VLDB

Simple queries of this kind can be implemented in the MediaWiki expression language (an example is shown in Listing 4.1); more complex ones require SPARQL. The output of both kinds of queries can be a table, list, map, timeline, etc. For example, finding all roles that a person had in events requires joins between person and event entities. The results can be achieved by the SPARQL query shown in Listing 4.2.

```
SELECT ?event ?person ?hasRole
WHERE {
  ?e            rdfs:label              ?event .
  ?e            ?hasRole                ?person .
  ?hasRole      rdfs:subPropertyOf      property:Has_person .
  ?person       rdfs:label              "PERSON NAME" .
  }
```

Listing 4.2: **SPARQL query** for finding all roles that a person has ever had in events. This query requires joins between person and event entities.

Geographical distribution and affiliation changes of persons in the role of general chairs of events related to a certain field over the last ten years can be shown on a map or graph by embedding a SPARQL query (see Listing 4.3) into the wiki page representing a certain field (i.e., in MediaWiki, a *category* page).

4.3.2 Results and Discussion

In this section, we report the outcome of comprehensively analyzing the metadata of 40 renowned Computer Science conference series over the past 30 years.

Acceptance Rate. Figure 4.5 shows the average acceptance rate for a sample of 10 conferences from different Computer Science communities in five consecutive 5-year periods from 1992 (due to the lack of data for the earlier events) to 2016. The average acceptance rate for all series falls into the range 17% to 26% in that time window. The most exceptional acceptance rate was the acceptance rate of COLT in the second period (45%), which decreased to 36% in 2016. The average acceptance rate of CCS dramatically decreased. The number of submissions to this series increased over time; however, the acceptance rate remained approximately the same. Only the average acceptance rate of EuroCrypt significantly increased to 33% in 2007–2011 and then decreased again to 20% in 2012–2016.

Continuity. The continuity of conferences is calculated using the formula defined in Equation 4.1. For example, the continuity of CCS (ACM Conference on Computer and Communications Security) is 92%, where it was held every year from 1993 except for two years in 1995 and 2003. Moreover, the continuity of TPDL (The International Conference on Theory and Practice of Digital Libraries) is 100% since it has been held every year since the first year of establishment.

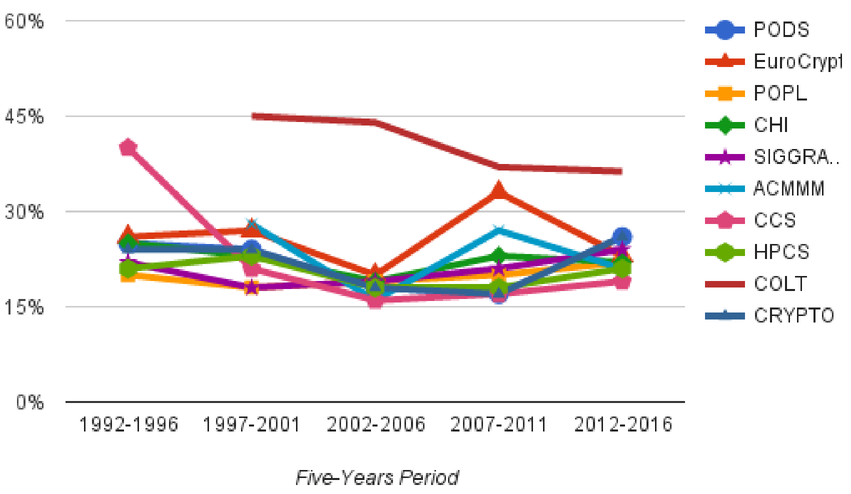


Figure 4.5: **Average Acceptance Rate.** The average acceptance rate for top-ten Computer Science Conference Series.

For illustration, the continuity of five conferences is shown in Table 4.4; for the others, the continuity is 100%. Overall, we observed a very high continuity among the renowned conferences.

```
{{#sparql:
    SELECT ?event ?country ?person
    WHERE {
      ?e  a                                category:Semantic_Web .
      ?p  property:Has_location_country    ?country .
      ?p  property:Has_affiliation         ?organization .
      [...]
      MINUS { ?e  property:Has_general_chair  :person . }
      FILTER (?startDate >= "2007-01-01"^^xsd:date && ?endDate < "2017-01-01"^^xsd:date)
    } LIMIT 10
      | format=maps
}}
```

Listing 4.3: A SPARQL query for finding the geographical distribution and affiliation changes of persons in the role of general chairs of events related to a certain field in the last ten years.

Geographical Distribution. The EUROCRYPT conference series has been held in a different country every year since 1987 but always in Europe. This is mostly related to the organization committee in this series since it is a European committee. For the same reason, the SIGGRAPH series has been held every year since 1974 in different North American countries

Table 4.4: **Continuity.** The continuity of a sample of five conference series.

Conference	Age	Editions	Regularity	Continuity (C)
ACMMM	23	22	1	96%
CCS	24	22	1	92%
CHI	35	34	1	97%
FOGA	27	13	2	96%
TPDL	21	21	1	100%

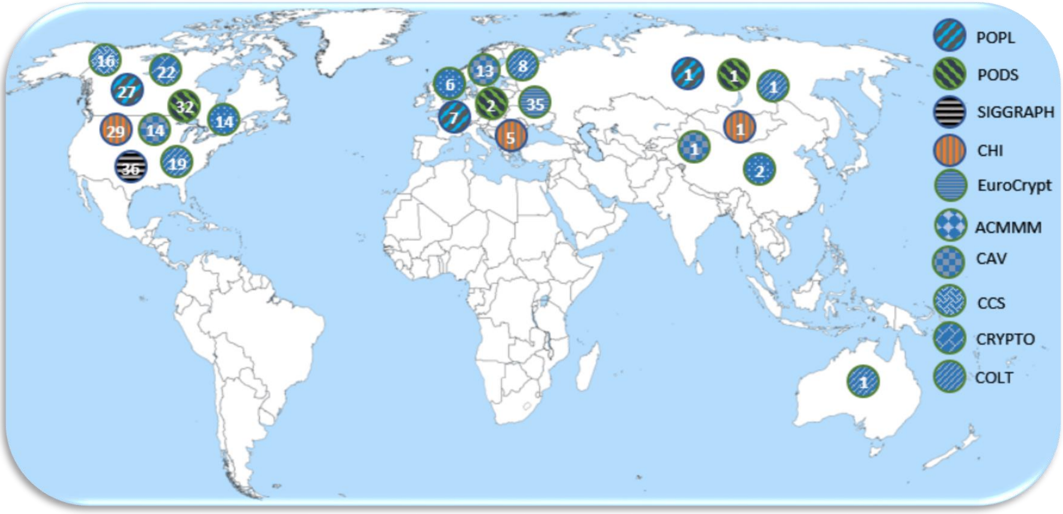


Figure 4.6: **Geographical distribution.** The geographical distribution of a sample of ten randomly selected conference series.

(mostly in the US). The FOCS series has been held 26 times in the US, every year since 1989 in North America, and in Europe only for one edition in 2004. On the contrary, ISSAC has been moving between different countries of different continents such as Japan, Canada, Germany, etc., since its first edition. Figure 4.6 shows the Geographical Distribution of a sample of ten conference series randomly selected. The most geographically diverse conference series is EUROCRYPT (diversity by country in Europe). The most static event series is FOCS, which has been held 26 times only in the US for the past 25 years.

Time Distribution. Most editions of top conference series are held around the same month of each year; see Figure 4.7. Namely, the PERCOM conference (IEEE International Conference on Pervasive Computing and Communications) has been held every year since 2003 in March and POPL (ACM SIGACT Symposium on Principles of Programming Languages) has been held every year since 1994 in January. Furthermore, almost all conferences in the study have been established around the same month. For example, EuroCrypt is always held in April or May, and SIGGRAPH always held in July or August.

Community Popularity. There are five groups labeled: Computer Security (SEC), Computer Graphics (GRA), Database Systems (DB), Programming languages (PROG), and Software Engineering (SE), each of which contains the top-five events belonging to this field. We considered two-time intervals: three 10-years periods for accepted papers but three 5-years periods for submitted papers due to the difficulty of obtaining information about the number of submitted papers for many conferences. Table 4.5 compares five Computer Science communities in terms of the number of accepted and submitted papers. GRA communities made the largest number of submissions in the whole period, even though GRA submissions began to decrease from 2005 until they reached their minimum value in the last period. The average number of accepted papers (Figure 4.9) in GRA doubled in the first time frame and increased by almost 150% in the past ten years, similarly in SE. The average number of accepted papers in DB slightly increased in the first period and then again increased in the last ten years by 50%. Over the three periods, the GRA community has attracted most, and PROG has attracted the least submissions. Overall, as shown in Figure 4.10, there is an increasing number of submissions for

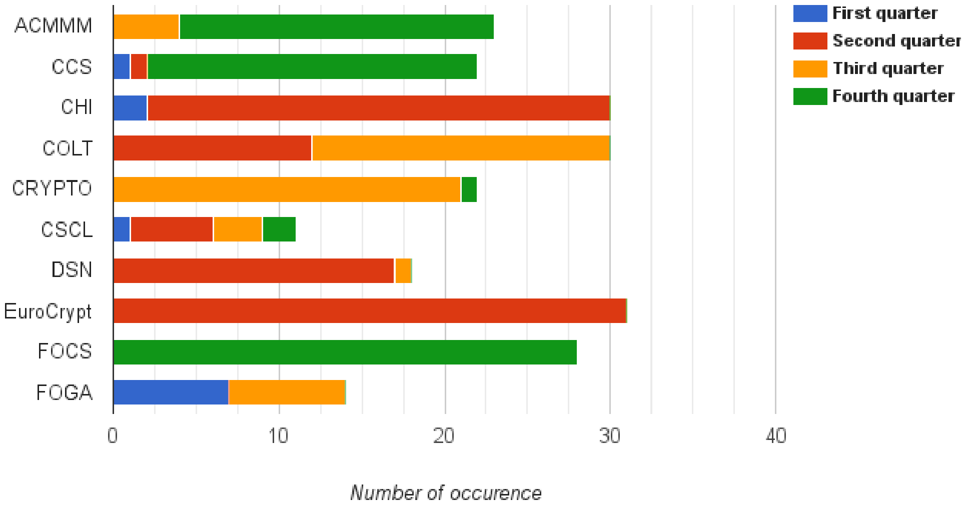


Figure 4.7: **Time distribution.** The time distribution for a sample of ten conference series.

all Computer Science communities we considered.

Field Productivity. We calculated the Field Productivity for the five communities in the study. The results are shown in Figure 4.8. We found that PROG and DB remained at the same FP with some ups and downs from 1987 to 2010 and then saw a slight increase. At the end of the 1980s and the early 1990s, GRA had the lowest FP with less than 1% until it began to increase to around 3% by 1993 and continued growing to approximately 10% before decreasing to only 4% by the end of the period. Moreover, all fields had an FP of approximately 3% from 1987 till 2006. For instance, FP of SE varied between 1.13% and 2.97%. Besides, GRA reached the

Table 4.5: **Accepted and Submitted Papers.** The number of accepted and submitted papers for five Computer Science communities over three 10-year and three 5-year intervals respectively.

		Accepted Papers			**Submitted Papers**		
		1987-1996	1997-2006	2007-2016	2000-2004	2005-2009	2010-2014
GRA	*Avg.*	78.9	198.4	302.2	927	1,188	1304
	Min	33	113	219	535	1,090	1,017
	Max	172	216	593	1,182	1,454	1,786
SEC	*Avg.*	106.2	130.1	235	633	855	1,12
	Min	68	86	144	513	607	936
	Max	135	157	337	849	988	1,264
PROG	*Avg.*	103.8	110.6	170.3	522	594	754
	Min	90	102	128	481	568	676
	Max	119	125	199	576	635	827
SE	*Avg.*	65.1	116.1	189.5	709	904	992
	Min	36	105	144	585	803	837
	Max	101	130	256	879	1,038	1,170
DB	*Avg.*	135.7	161.9	240.7	905	1,250	973
	Min	117	121	151	718	1,166	548
	Max	166	206	347	1,207	1,348	1,109

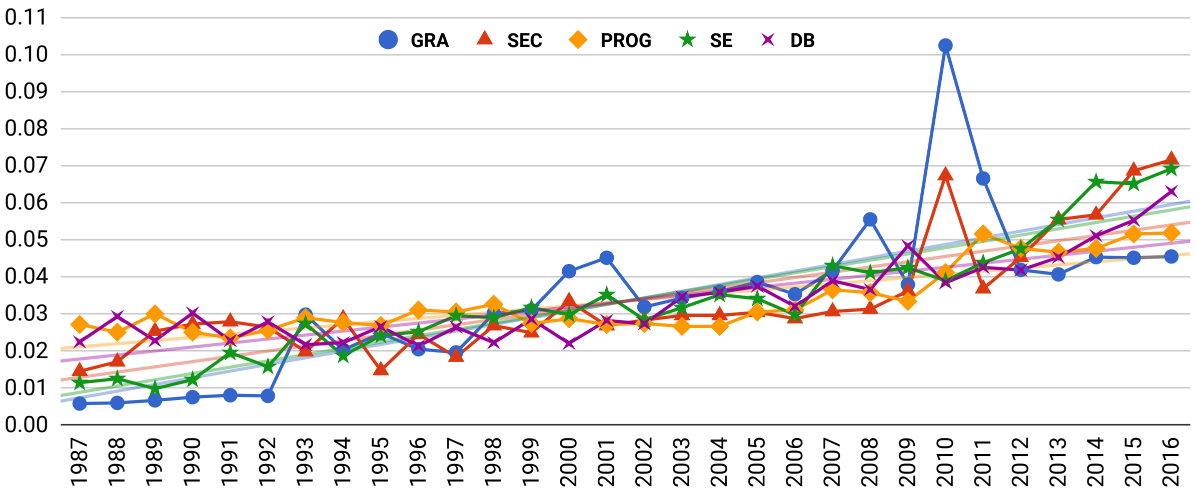


Figure 4.8: **Field Productivity.** The field productivity of five Computer Science communities in the period 1987–2016.

maximum FP in 2010 with 10%, and DB reaches the maximum FP with 6.45% in 2016. Overall, GRA has the highest FP with 5,795 publications over the other fields; the PROG community has the lowest FP with 3,707 publications. The DB community ranks second with 5,383 publications, followed by SEC with 4,715 publications and then PROG with 3,847 publications.

4.3.3 Summary

In this section, we presented a study for analyzing scholarly communication metadata of scientific events belonging to Computer Science. We combined descriptive and exploratory analysis with regard to a broad set of metrics, supported by spreadsheets, charts, and queries in the OpenResearch.org semantic wiki. Up to our knowledge for the first time, we were able to empirically validate the often-raised concern of a proliferation of submissions to major conferences.

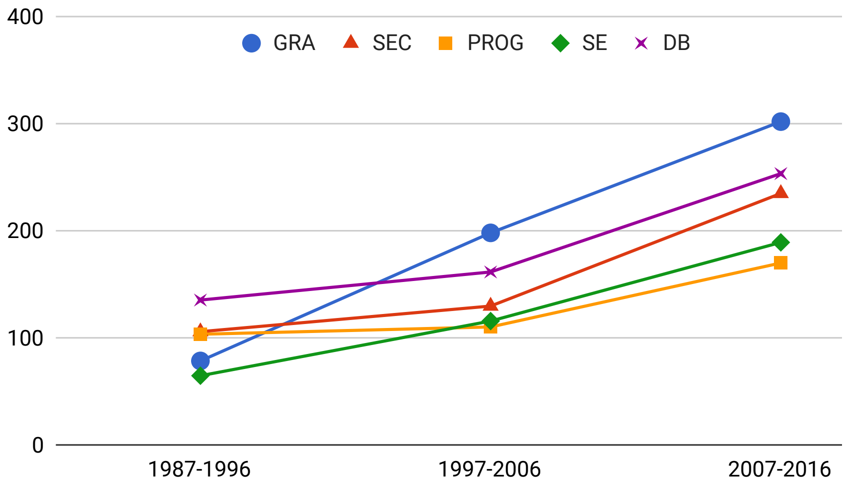


Figure 4.9: **Accepted Papers.** The average number of accepted papers of five Computer Science communities in three 10-year periods between 1987 and 2016.

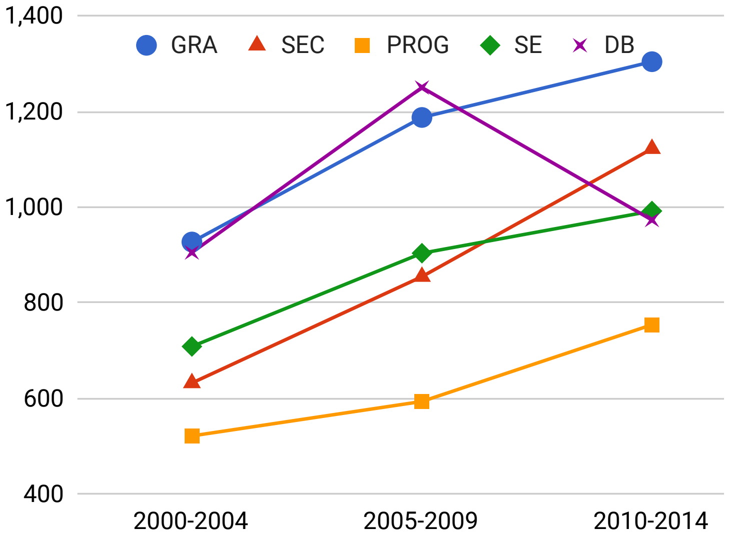


Figure 4.10: **Submitted Papers.** The average number of submitted papers of five Computer Science communities in three 5-year periods between 2000 and 2014.

Also, we were able to calculate and demonstrate with our method several other indicators, such as a new way to calculate conference continuity, the popularity of different communities, a new way to calculate field productivity or the geographic distribution of conferences. In addition to efficiency gains, the digitization of scholarly communication also has negative impacts, most significantly the proliferation of submissions, which substantially increases the reviewing workload with an already noticeable knock-on effect on reviewing quality, which is one of the core features of peer-review [172]. In summary, we made the following observations:

- With the number of submissions to the top conferences having tripled on average in the last three decades, acceptance rates are going down slightly.
- Most of those conferences that are A- or A*-rated today have a long continuity.
- Geographical distribution is not generally relevant; some good conferences take place in the same location; others cycle between continents.
- Good conferences always take place around the same time of the year.
- Some topics have attracted increasing interest recently, e.g., database topics thanks to the "big data" trend. This might be confirmed by further investigations into more recent, *emerging* events in such fields.

In the next section, we expand the analysis to other fields of science, including Mathematics, Engineering, and Physics, as well as considering smaller events, such as meetings and workshops. Furthermore, we study to assess the impact of digitization with regard to further scholarly communication stakeholders, such as sponsors and proceedings publishers.

4.4 Metadata Analysis of Scholarly Events in Four Fields of Science

The integration and harmonization between knowledge exchange channels, i.e., scholarly events, are based on the grown culture of any particular community and community-defined criteria for analyzing the quality of these channels. For example, some fields, such as medical science, use publishing in journals as the leading and most valuable channel; however, some other fields, such as computer science, publishes mostly in events. Furthermore, community-defined criteria distinguish highly ranked instances of any particular class of channels as well as popular events and journals. However, a systematic and objective analysis of metadata supports researchers in better dissemination of results to the right communities. In this section, we study the various characteristics of scholarly events in different fields of science using the proposed metrics-suite (cf. section 4.1) of standard criteria.

In order to obtain a better understanding of scholarly event characteristics in four fields of science, we analyzed the metadata of scholarly events of four major fields of science, namely Computer Science, Physics, Engineering, and Mathematics using the proposed metric suite, which comprises ten metrics. In particular, we analyzed renowned scholarly events of five communities within computer science, namely the World Wide Web (WEB), Computer Vision (CV), Software Engineering (SE), Data Management (DM) as well as Security and Privacy (SEC). This analysis is based on a systematic approach using descriptive statistics as well as exploratory data analysis. Metadata analysis refers to the statistical methods used in research synthesis for drawing conclusions and providing recommendations from the obtained results.

Aggregation of metadata from several data repositories, digital libraries, and scholarly metadata management services enables comprehensive analysis and services to the users of such services. Our comprehensive dataset of scholarly events of the aforementioned fields is openly available in a semantic format and maintained collaboratively at OpenResearch.org. The availability of specific metadata, e.g., citation count, restricts the objective impact measurements to metrics related to citations only. Furthermore, the diversity of the meaning of impact brings challenges for the development of robust, widely accepted impact measures. This diversity limits the scope and quality of possible evaluations. In this analysis, we address the following questions concerning impact:

- *What are the characteristics of scholarly events in Computer Science, Physics, Engineering, and Mathematics*?
- *What are the top citation impact events in computer science*?
- *How are top citation impact events assessed by ranking services*?

In section 4.3, the particular focus was on analyzing computer science events in terms of continuity, geographical and time distribution, field popularity, and productivity. In this section, we extended this work by studying the characteristics of scholarly events in four research fields (using SEMA 2.0), involving Computer Science, Physics, Engineering, and Mathematics, by:

- widening the research scope by adding more research questions,
- adapting the analysis methodology, since we found that several tasks, such as data gathering and events identification, need to be overlapped,
- expanding the analysis of the metadata of both Computer Science and non-Computer Science events, and
- providing a set of recommendations for event organizers, potential authors, proceedings publishers, and sponsors.

This study aims at answering the following research questions: RQ1.1) *How important are events for scholarly communication in the respective communities*?, RQ1.2) *What makes an event a high-ranked target in a community*? As well as RQ1.3) *How can scholarly events be assessed using a mixture of metrics*? The aim is to have a far-reaching influence on the contributions and information needs of the different stakeholders of scholarly communication:

Table 4.6: **Four Fields of Science.** The research fields considered in this study and the corresponding sub-fields.

Fields	Sub-fields
Computer Science	World Wide Web (WEB), Computer Vision (CV), Software Engineering (SE), Data Management (DM), Security and Privacy (SEC), Knowledge Representation and Reasoning (KR), Computer Architecture (ARCH), Machine Learning (LRN)
Physics	Astronomy, High Energy Physics, Particle Accelerators, Applied Physics and Mathematics, Nuclear Science, Nanomaterials, Neutrino Detectors, Geophysics
Engineering	Civil Engineering, Mechanical Engineering, Chemical Engineering, Electrical Engineering
Mathematics	Algebra, Mathematical Logic, Applied Mathematics, General Mathematics, Discrete Mathematics

- *event organizers*: to assess and elevate the development and impact of their events,
- *authors*: to identify renowned events to submit their research results,
- *proceedings publishers*: to evaluate the impact of the events whose proceedings are being published by them,
- *participants*: to identify renowned events to attend, and
- *event sponsors*: to tighten the collaboration between industry and academia.

4.4.1 Data Gathering

The relevant metadata of 3,704 various scholarly events was collected, involving conferences, workshops, symposiums, and meetings in the CS, PHY, ENG, and MATH fields. The corresponding sub-fields of each field involved in this study are listed in Table 4.6. Computer Science sub-fields were obtained from the ACM Computing Classification System (CCS)[8], while sub-fields of non-CS fields were obtained from the Conference Management Software COMS[9], a well-known conference management system. We focused on the WEB, CV, SE, DM, and SEC sub-fields of Computer Science because they were the top-5 sub-fields in our datasets in terms of data availability and had the highest number of submissions in the last decade. This data, including title, series, sub-field, start date, end date, homepage, country, and h5-index, has been collected from different sources. Only for Computer Science, metadata of scholarly events is available as Linked Data through DBLP (cf. Table 3.3).

Data harvesting sources: Two major datasets are used in this study: 1) *OpenResearch dataset* (ORDS) (5,500+ events at the time of carrying out the study) and 2) *SCImago dataset* (SCIDS) (2,200+ events). SCIDS stores metadata of each event in terms of SJR, h5-index, number of references in each paper, and the number of citations for each event's proceedings volume. On the other hand, ORDS stores different attributes, such as start date, end date, number of submissions, and number of publications. Therefore, different statistical methods can be used. The reason for collecting data from two separate sources is that the Computer Science

[8] `https://dl.acm.org/ccs/ccs.cfm`

[9] `https://www.conference-service.com/`

community, compared to other research fields, archives more information about past events, such as acceptance rate, location, date, and the number of submitted and accepted papers. Furthermore, there are many online services for archiving past events metadata and ranking their proceedings, such as DBLP, conference city, CASCONet, and AMiner (cf. Table 3.3). To identify renowned events from a large number of scholarly events available nowadays, we used several online services for ranking scholarly events described in subsection 3.1.1. These services are commonly used to identify high-quality events in each field of science.

4.4.2 Data Analysis

Exploratory analysis techniques are used for analyzing the metadata of the selected events over the past 30 years. Generally, the data analysis process is divided into three categories: The analysis presented in this work is based only on DSA and EDA because our purpose is to describe and explore new insights. SEMA 2.0 is used in this analysis. Since OpenReserach.org has its own SPARQL endpoint, SPARQL is used to query ORDS. For example, finding all events related to Computer Security, which took place in Europe[10] along with their acceptance rate (less than 20%) between 2013 and 2018; this requires joins between field/topic and event entities. The SPARQL query in Listing 4.4 is designed to retrieve these events.

4.4.3 Results and Discussion

In this section, we report the results of our analysis of events metadata within the two datasets over the past 30 years, according to the SEQA suite defined in section 4.1. One notable observation is that there is no comprehensive information about the number of submissions and publications in other fields than Computer Science. Therefore, metrics such as acceptance rate, field productivity, and progress ratio cannot be practically applied to events belonging to these fields. For the same reason, we categorize our results into three categories: 1) scientific fields analysis, 2) Computer Science communities analysis, and 3) individual events analysis.

```
SELECT ?title ?endDate ?startDate ?city ?country ?wikipage ?acceptanceRate
?continent
WHERE {
  ?e              rdfs:label                     ?title .
  ?e              a                              category:Computer_Security.
  ?e              icaltzd:dtstart                ?startDate .
  ?e              icaltzd:dtend                  ?endDate .
  ?e              property:Acceptance_rate       ?acceptanceRate .
  ?e              swivt:page                     ?wikipage .
  ?e              property:Has_location_country  ?country .
  ?country        rdfs:subClassOf                ?partContinent .
  ?partContinent  rdfs:subClassOf                ?continent .
  ?continent      rdfs:isDefinedBy               site:Category:Europe .
  FILTER ( ?acceptanceRate <  20.0  &&
  ?startDate >= "2013-01-01"^^xsd:date && ?endDate < "2018-01-01"^^xsd:date)
}
ORDER BY DESC(?acceptanceRate)
```

Listing 4.4: **SPARQL query.** A SPARQL query for finding all events belonging to Computer Security which took place in Europe along with their acceptance rate (less than 0.20) between 2013 and 2018; this requires joins between field/topic and event entities.

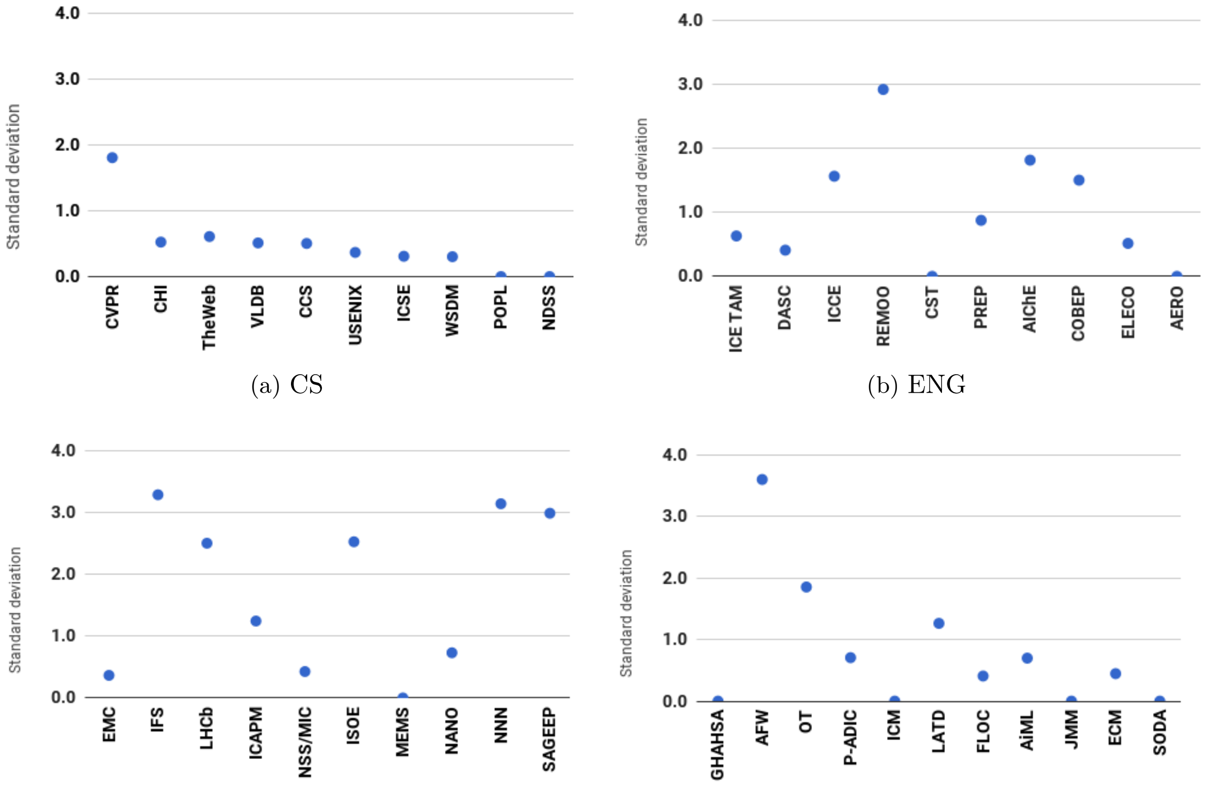
[10] The complexity of the relation between a country and its continent is owed to the way OpenResearch.org organizes such knowledge to provide convenient browsing by regions of continents such as “Western Europe”.

Scientific Fields Analysis

This section presents the results of analyzing metadata of events from all considered scientific fields, i.e., CS, PHY, ENG, and MATH, with respect to the metrics that can be applied, which are TD, GD, h5-index, continuity, SJR and citation count.

Time distribution. We analyzed the time distribution metric in terms of the standard deviation of the month of the year in which the event takes place for all events of CS, MATH, ENG, and PHY in ORDS in the last two decades. Namely, EMC (International Symposium on Electromagnetic Compatibility) has been held 17 times in August, and only three times in July, therefore, it has a low standard deviation of $\sigma = 0.4$, while NDSS (Annual Network and Distributed System Security Symposium) has been held in February every year since the beginning, therefore $\sigma = 0$. Notably, Computer Science events have the lowest σ among events of other fields. Overall, we observed that most editions of the high-quality events in all fields have always been held around the same month every year, i.e., their time distributions have low standard deviations (Figure 4.11).

Geographical distribution. We analyzed the geographical distribution metric for all CS, MATH, PHY, and ENG events in the last two decades. As shown in Figure 4.12, the USA hosted 50% or more of the scholarly events in all fields during the whole period. All other countries have significantly lower percentages. For instance, Canada hosted 7% of CS events and a significantly low percentage of events of the other fields, while France hosted 4% of both MATH and PHY events.



(a) CS

(b) ENG

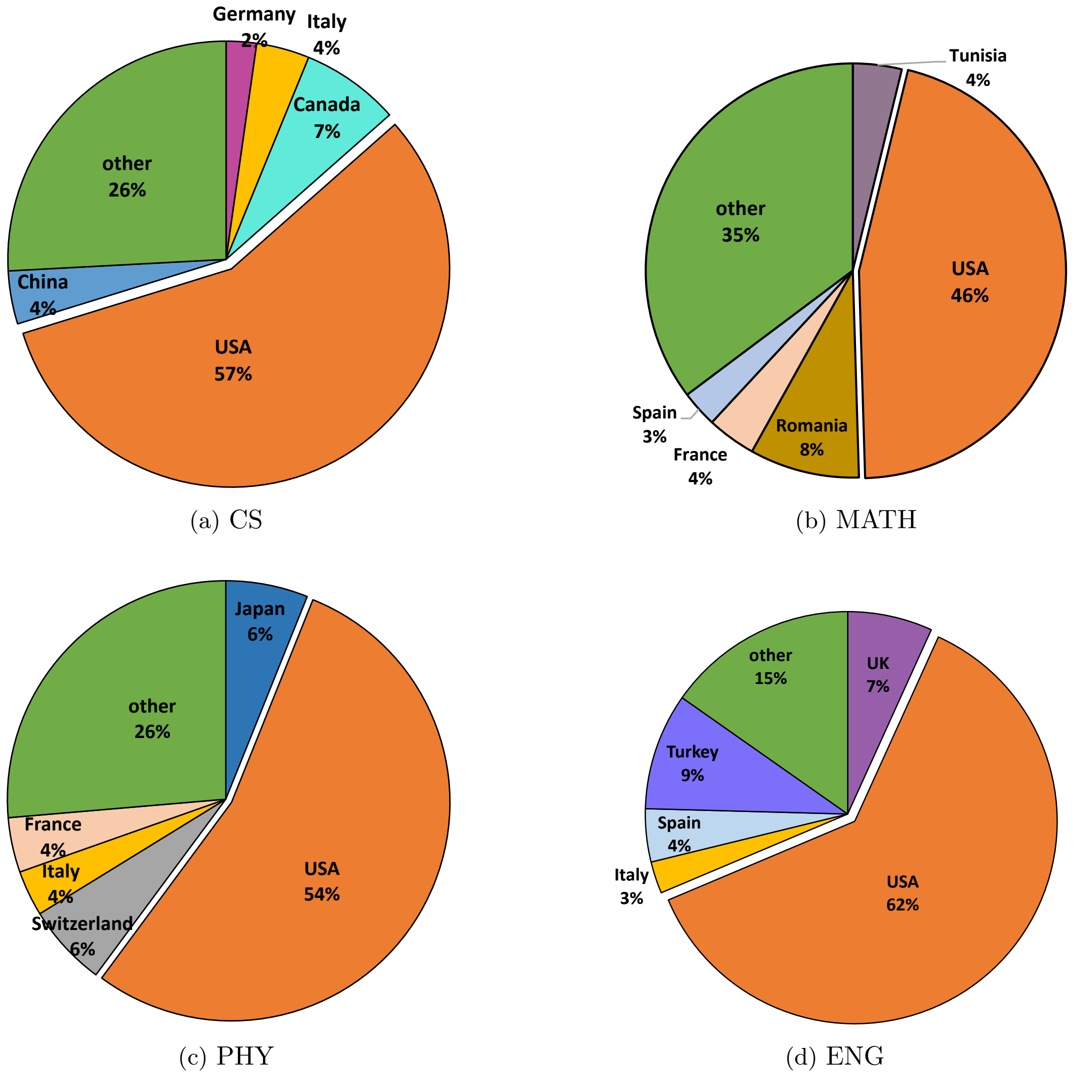
(c) PHY

(d) MATH

Figure 4.11: Time distribution of events in CS, MATH, PHY and ENG in the last two decades.

H5-index. To compare the impact of events of the four scientific fields, we analyzed the h5-index of the top-25 events in each field. Figure 4.13 shows the frequency distribution of events by categorizing the h5-index of the events into four ranges (0–10, 11–20, 21–30 and 30+). The slices of each pie chart compare the frequency distribution of events in each field concerning the h5-index. The CS community has the highest number of events (92%) with ($h > 30$), while the ENG community has the lowest one (16%). The number of MATH events with ($h > 30$) is as high as that of PHY, while each of them is almost twice as high as ENG. Also, the number of ENG events with ($21 \leq h < 30$) is as high as that of the PHY. Overall, we found that CS has the highest number of high-impact events, while ENG has the lowest. This can be, for example, attributed to the size of the field and its communities and their fragmentation degree, since a large community results in higher citation numbers. Also, it might be an indication of the importance of events for scholarly communication of this community (e.g., in comparison to journals).

Continuity. As shown in Figure 4.14, all events in all fields have a continuity higher than



(a) CS

(b) MATH

(c) PHY

(d) ENG

Figure 4.12: **Geographical distribution** of CS, MAT, PHY and ENG events in the last two decades.

90% except for NNN (International Workshop on Next generation Nucleon Decay and Neutrino Detectors) and ICE-TAM (Institution of Civil Engineers-Transport Asset Management), they have continuities of 88% and 86%, respectively. The reason is that NNN was not held in 2003 and 2004, and ICE-TAM was not held in 2013. For CS events, the continuity of USENIX (Usenix Security Symposium) is 93% because it was held every year from 1990 except for two years (1994 and 1997). This emphasizes that even the lowest continuity value of CS events is relatively high, in comparison to the other fields. Notably, we found that all MATH events involved in the dataset have continuity of 100% (green bar chart), even the oldest one (International Conference in Operator Theory), which has been holding every year since 1976. Overall, we observed a very high continuity among renowned events, which is an indication of stability and of the attractiveness of hosting and organizing such events.

SJR indicator. We calculated the average SJR indicator of all events, in SCIDS. As shown in Table 4.7, CS communities have an average SJR of 0.23, which is almost twice the value of PHY

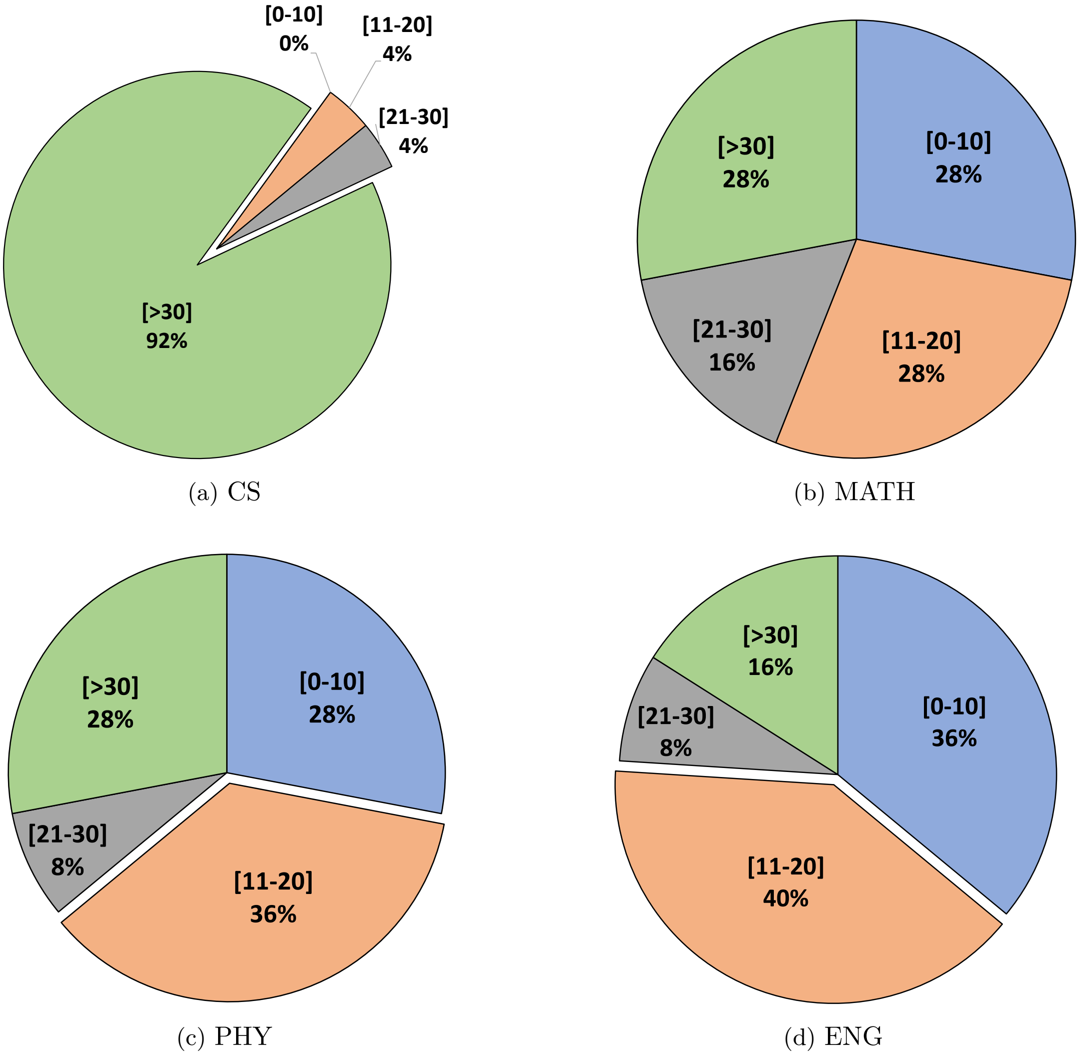
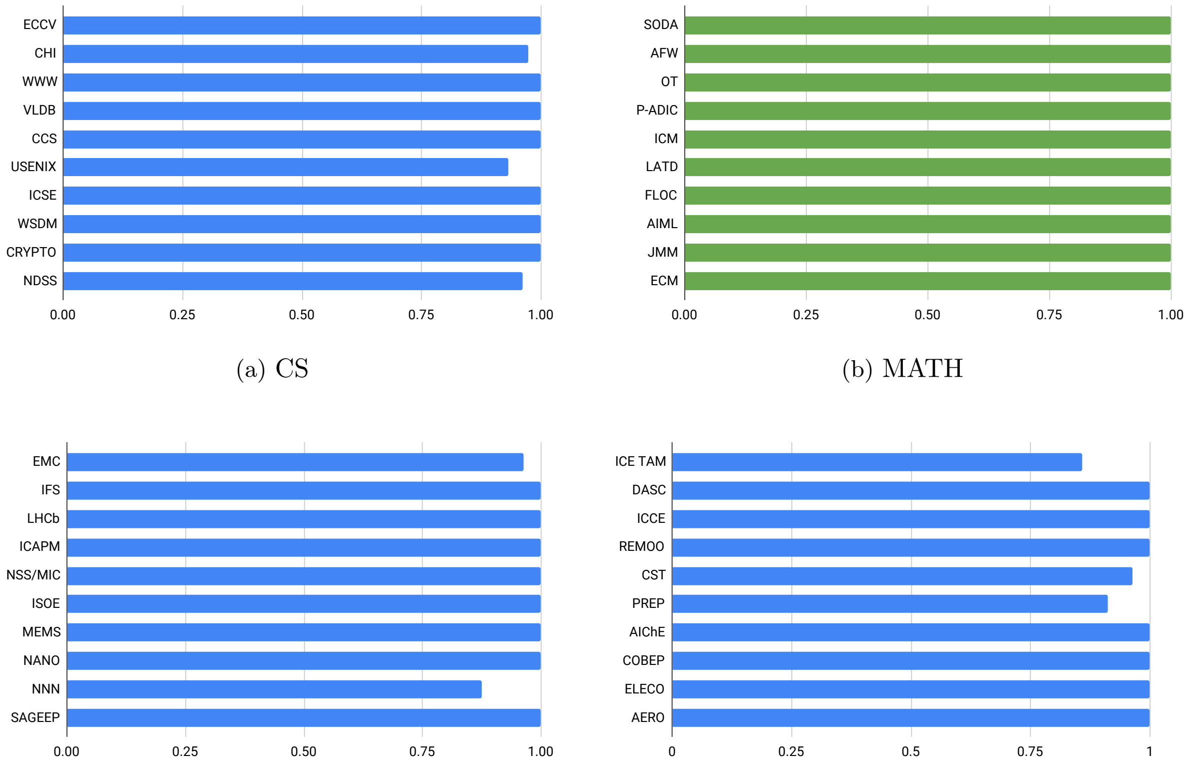


Figure 4.13:
Frequency of the top-25 events in CS, MATH, PHY and ENG in terms of their h5-index.

and ENG each; MATH comes next. As the SJR indicator is calculated based on the number of citations, we can infer that CS and MATH communities were more prolific or interconnected in terms of citations in 2016 compared to PHY and ENG. Since PHY had the highest number of articles published in 2016 (among other fields), it has not the highest SJR indicator. This can be somewhat attributed to the number of citations per article, which is lower in other fields. On average, a CS paper contains about 20 references (refs/paper), while a PHY paper contains only 15 references. In terms of the total number of references included in the papers published in 2016 (total refs.), CS has the highest number of references, while the ENG field has the lowest.

Citation count. We analyzed the number of citations of all proceedings papers of events that took place in Germany, for the CS, ENG, MATH, and PHY between 2007 and 2016. Figure 4.15 illustrates the development of the number of citations for each field over the period 2007–2016. This indicates that there is a relatively large number of researchers in Germany working in CS. While the number of citations has increased for all communities during this period, the most substantial increases were observed in CS and ENG. The leading role of CS has persistently increased throughout the whole period. The citations for PHY and MATH are relatively low and are almost similar. Overall, we can see a clear upward trend in the number of citations of CS publications, compared to a slight increase in the three other fields.



(a) CS

(b) MATH

(c) PHY

(d) ENG

Figure 4.14: **Continuity** of CS, MAT, PHY and ENG events in the last two decades.

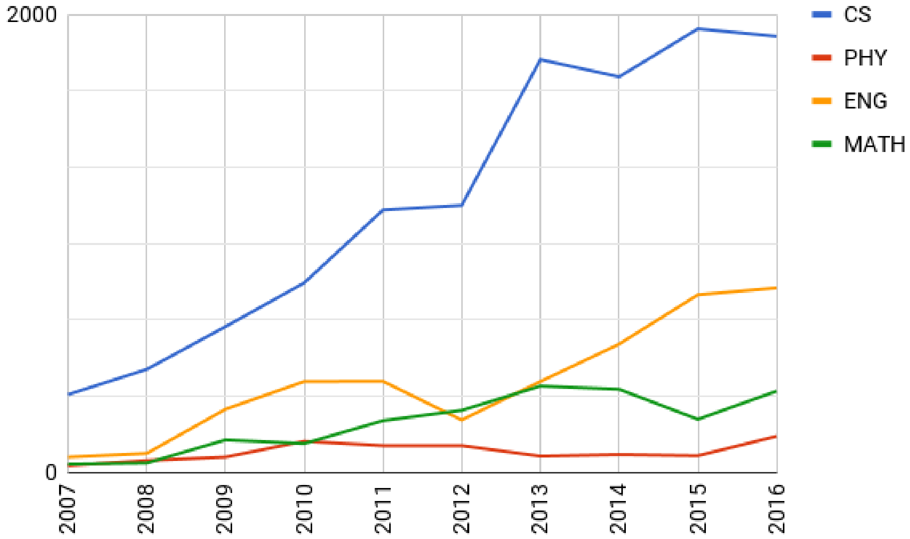


Figure 4.15: **Citation count** by different communities in Germany.

Computer Science communities Analysis

This section focuses on analyzing events of five Computer Science communities (WEB, CV, SE, DM, and SEC) based on the number of submissions and accepted papers, and all applicable metrics, such as AAR, FP, and PR.

Geographical distribution. We analyzed the geographical distribution of the top-5 events in each Computer Science community since 1973. As illustrated in Figure 4.16, the USA hosted most editions of events in all Computer Science communities. For instance, the USA hosted 40% (41 out of 96) of WEB events, 37.5% (54 out of 144) of CV events, 67.5% (104 out of 154) of SE events, 25.1% (34 out of 135) of DM events, and 66.4% (91 out of 137) of SEC events. The DM community has the broadest geographical distribution of events in 37 different countries hosting 137 events, while the WEB and SE communities have the narrowest geographical distribution with only 21 countries hosting 96 and 154 events, respectively. We observed that some events

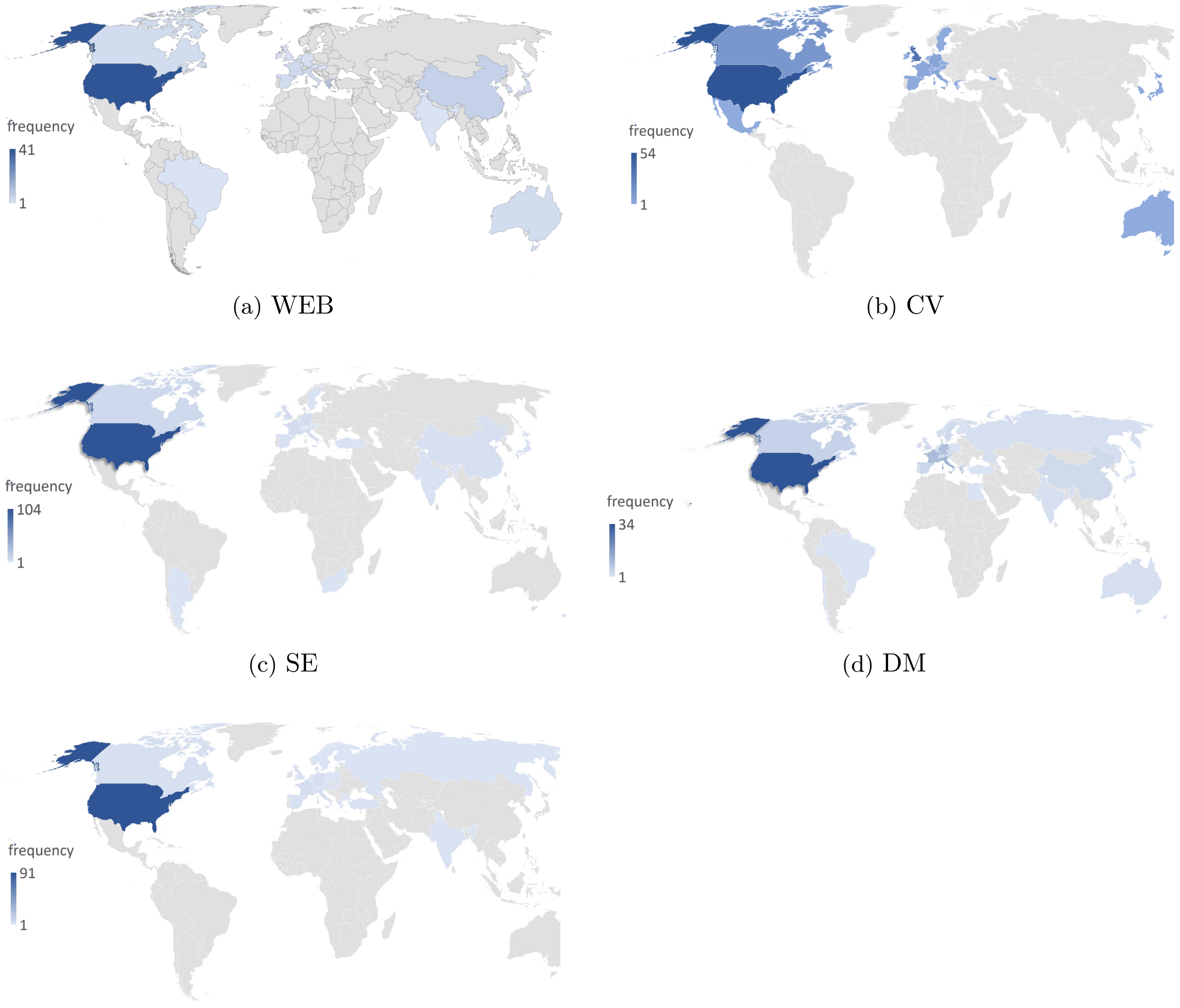
Table 4.7: **Scientometric profile** of the top CS, PHY, ENG and MATH events held in 2016. Data obtained from SCImago database.

Metrics	CS	PHY	ENG	MATH
max(h)	192	125	52	125
avg(h)	6.58	6.65	4.09	6.79
conf ($h > 10$)	151	28	21	25
avg. SJR	0.23	0.14	0.14	0.21
papers (2016)	13,234	16,795	1,675	16,585
papers (2013–2015)	163,556	90,245	46,790	68,814
total refs. (2016)	262,548	248,216	27,137	258,275
refs/paper	20	15	16	16

are restricted to one continent, such as EUROCRYPT, which has been held every year in Europe since 1982 and CRYPTO, which has been held every year in North America since 1995. Strikingly, we observed that most of the renowned events in SEC had been held in North America, particularly in the USA (83%), which indicates that the USA pays particular attention to this field. Notably, it is observed that the USA hosted most of the top-5 events in all communities.

Time distribution. We observed that most editions of top conference series are held around the same month each year (see Table 4.10). Namely, the WSDM Conference (ACM International Conference on Web Search and Data Mining) has been held every year since 2008 in February, and PLDI (conference on Programming Language Design and Implementation) has been held every year since 1987 in June.

Community popularity. We compared the popularity of the five Computer Science communities in terms of the number of submissions and accepted papers (Table 4.8). The CV community had the highest number of submissions and accepted papers during the three 5-year



(a) WEB

(b) CV

(c) SE

(d) DM

(e) SEC

Figure 4.16: **Geographical distribution** of the top-5 events in each Computer Science community since 1973.

time windows. The lead of CV in terms of submissions and accepted papers has continuously increased over the whole period, i.e., 2003–2017, until reaching nearly 4,000 submissions, on average, by the end of 2017 (highlighted in yellow). For example, the number of submitted papers in the period 2008–2012 (3,150 papers) is twice as large as of the period of 2003–2007 (1,148 papers) (highlighted in gray). Submissions, as well as accepted papers of the WEB community, have gradually increased throughout the whole period. The submissions of SEC have doubled in the last five years and, consequently, the accepted papers (highlighted in green). Differently, we observed that the average number of submitted papers of the DM community has slightly decreased in the last period, while the average number of submitted papers has slightly increased (highlighted in red). Overall, the CV community has had the most submissions among the Computer Science communities, while DM had the least.

Field productivity. The slices of the pie chart in Figure 4.17 compare the cumulative field productivity of eight Computer Science communities in the last ten years. We applied the FP metric to only the past ten years because not all data were available for all events in the earlier years. It is observed that CV is the most productive community over the other communities with an FP of 22%, then the DM community comes, while the computational learning community (LRN) is the lowest one of only 4%. As shown in Figure 4.18, DM, and WEB remained at the same FP with some ups and downs from 2008 to 2013; then, WEB had a slight decline in the next year, then began to rise again until it reached its maximum value in 2017. In 2015, the FP of SEC was the highest among all the others, i.e., about 17%, then dramatically decreased to 13.5% in the next year, then saw a slight increase to 14.9% in 2017. In summary, the FP of all communities has continued to increase gradually since 2008, ranging between 5.5% and 17% in the whole period, with the highest FP ever (17%) for SEC in 2015. In particular, FP of SE varied between 7.4% and 12.5%; for, CV it varied between 6.5% and 13.7%.

Table 4.8: **Accepted and submitted papers** measures for five Computer Science communities over three 5-year intervals.

		Accepted papers			**Submitted papers**		
		2003–2007	2008–2012	2013–2017	2003–2007	2008–2012	2013–2017
WEB	avg.	197	310	338	1,146	1,818	1,905
	min	143	264	251	921	1,739	1,491
	max	223	378	507	1,363	1,897	2,598
CV	avg.	342	866	965	1,148	3,150	3,914
	min	226	593	632	1,012	2,312	2,954
	max	473	1,177	1,255	1,909	4,047	4,901
SE	avg.	148	211	302	958	1,180	1,486
	min	116	190	261	751	1,094	1,405
	max	167	237	320	1,091	1,290	1,558
DM	avg.	211	327	383	1,279	1,543	1,481
	min	176	282	195	978	1,456	727
	max	265	364	503	1,727	1,611	2,248
SEC	avg.	145	195	397	912	1,103	1,915
	min	142	161	298	788	916	1,485
	max	152	258	508	980	1,326	2,353

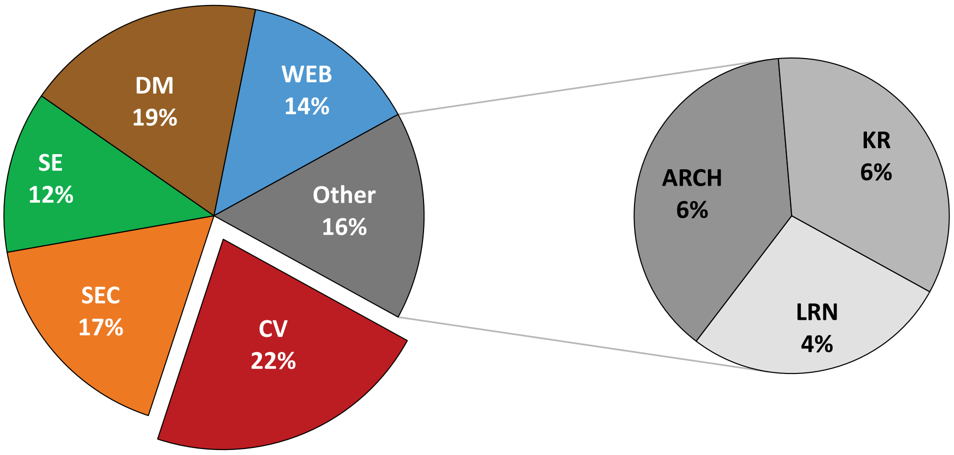


Figure 4.17: **Aggregated field productivity** of eight Computer Science communities over the last 10 years.

Individual Events Analysis

This section presents a study of the most renowned events within each Computer Science community.

Submitted and accepted papers. Figure 4.19(a)–(e) display the number of submissions as well as the number of accepted papers of the top-events, i.e., events with the highest h5-index, in each Computer Science community over the period 1995–2017. Among all events studied, in 2017, CHI had the highest number of submissions (2,400 submissions), while ICSE had the lowest one of 415 submissions. Accordingly, CHI had the highest (600 papers) and ICSE the lowest number of accepted papers (68 papers).

Average acceptance rate. Figure 4.22 shows the average acceptance rate (AAR) of each of the top-5 events in 2017 in each Computer Science community along with the country where most editions were hosted. The Web Conference, UIST (ACM Symposium on User Interface Software and Technology), VLDB (International Conference on Very Large Databases), ICSE (International Conference on Software Engineering), and USENIX (Usenix Security Symposium) have the lowest AAR among the top-5 events within WEB, CV, DM, SE, and SEC respectively.

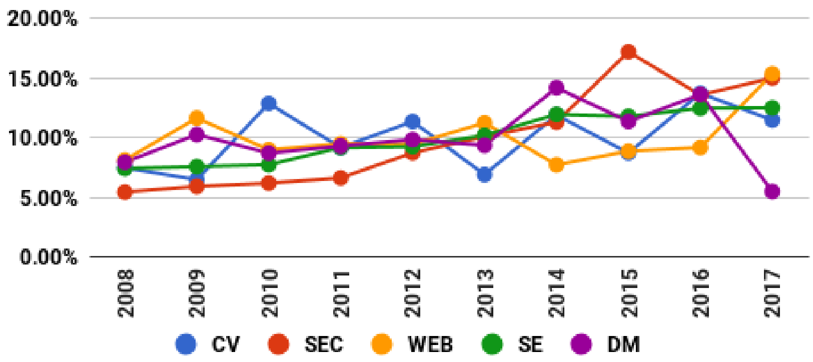
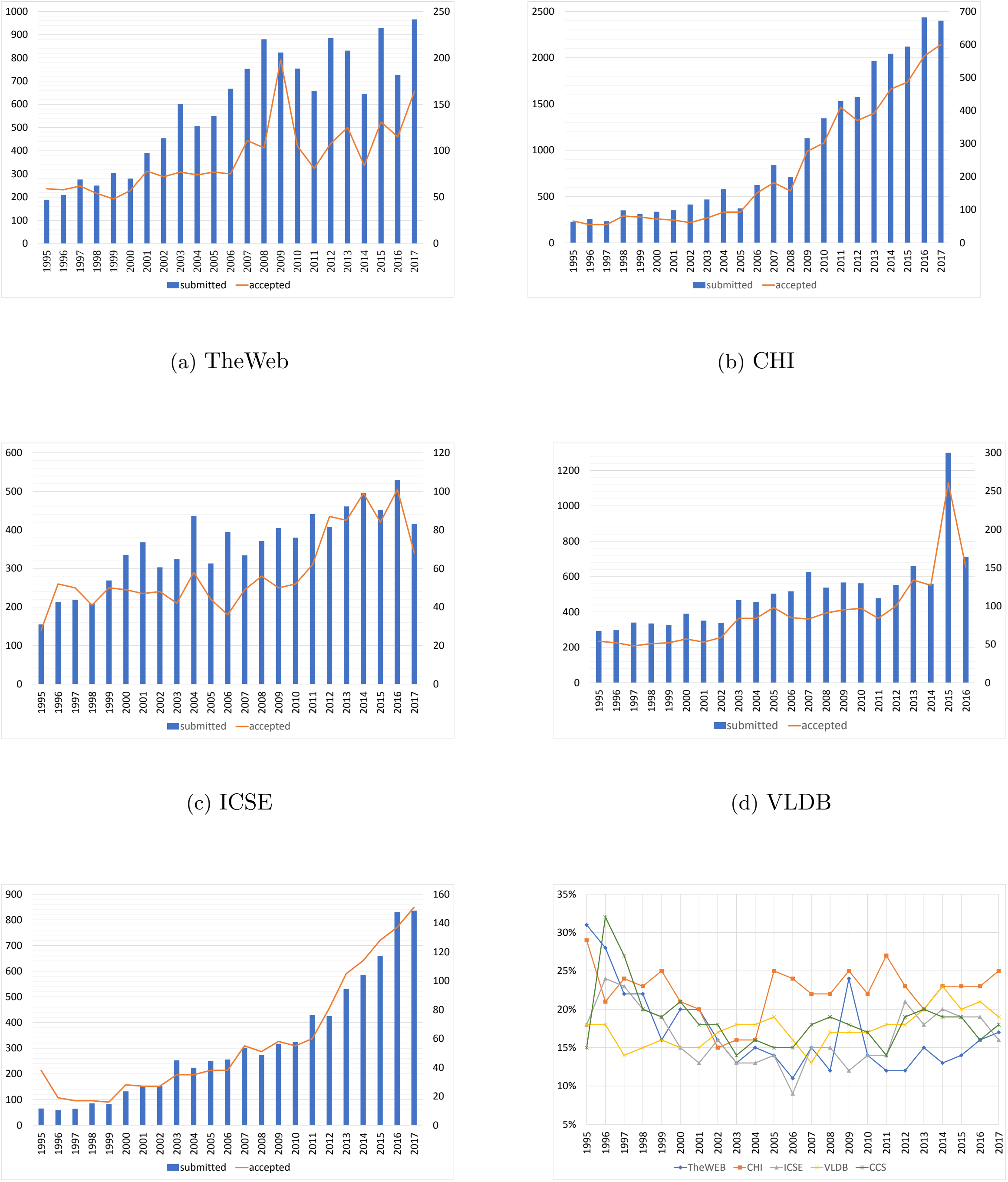


Figure 4.18: **Field productivity** of five Computer Science communities over the last 30 years.



(a) TheWeb

(b) CHI

(c) ICSE

(d) VLDB

(e) CCS

(f) average acceptance rate

Figure 4.19: **Number of submissions and accepted papers.** The number of submissions (main axis) and accepted papers (secondary axis) per year of the top event in each Computer Science community for the period 1995–2017.

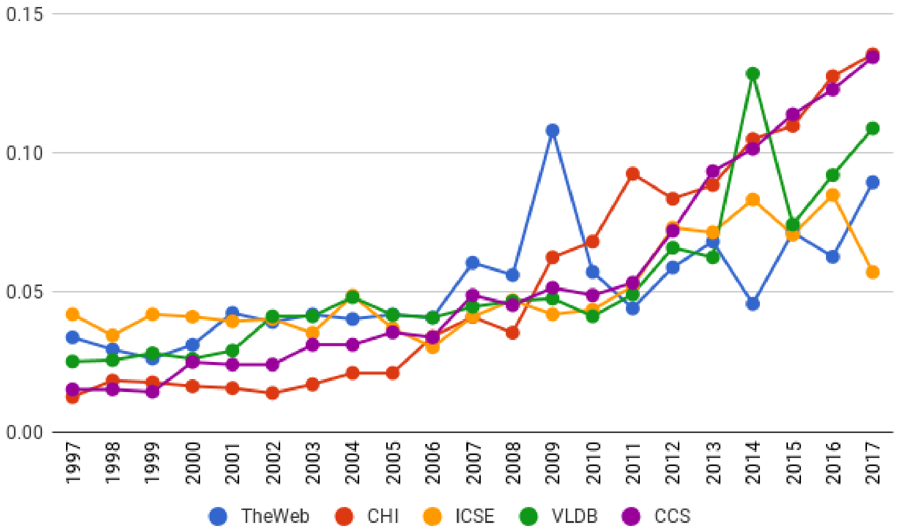


Figure 4.20: **Event Progress Ratio.** of the top events in each Computer Science community in the last two decades.

In general, the AAR for the top event in each Computer Science community is in the 13–25% range in the 20 year time window (Figure 4.19(f)). In addition, the acceptance rate of all events has remained relatively stable during the whole period. As can be seen from the charts Figure 4.19(a)–(e), the number of submissions has continuously increased over the whole period with slight ups and downs. However, the number of accepted papers increased steadily from the beginning until the end of the period, except for The Web Conference and VLDB in 2009 and 2015, respectively, where they showed peaks. The highest AAR ever, among these events, was

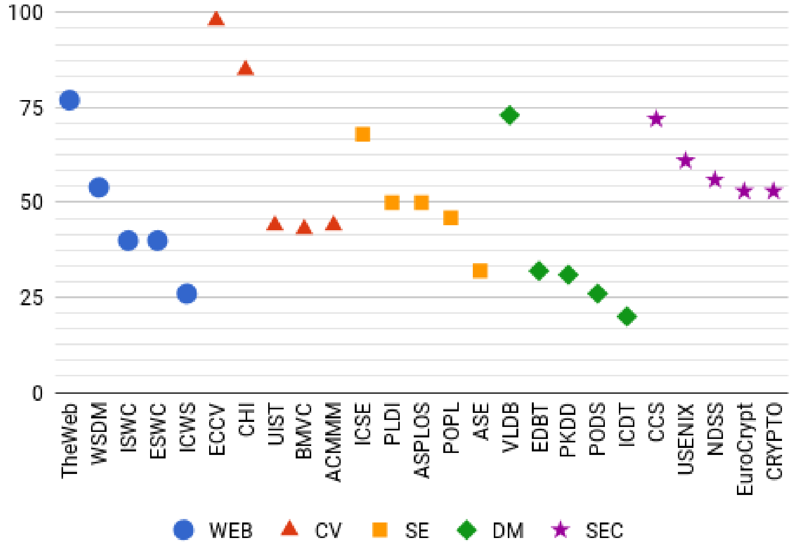


Figure 4.21: **H5-index.** Top-5 events in each Computer Science community according to their h5-index in 2017.

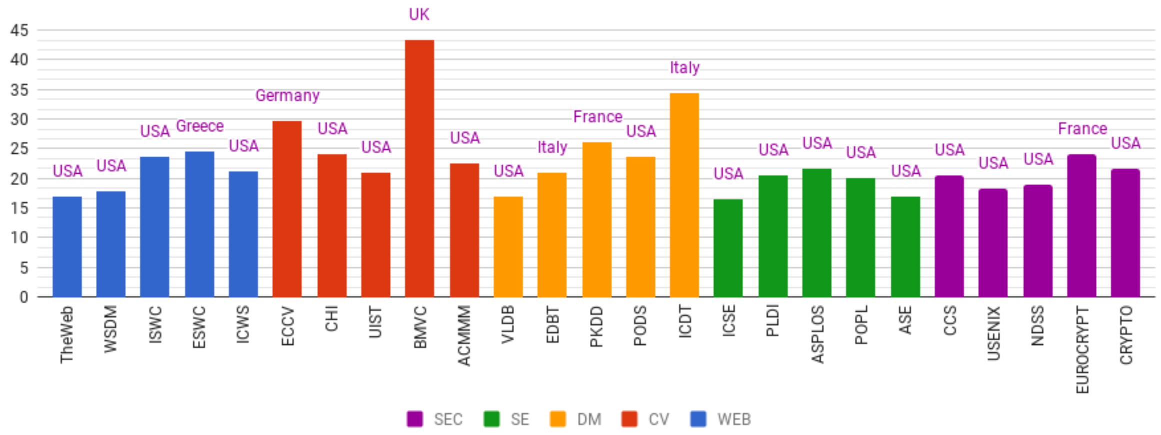


Figure 4.22: **The average acceptance rate** of top-5 events in 2017 in each Computer Science community indicating the most visited countries for each event series.

the one of CCS in 1996 (32%), which subsequently decreased to 18% in 2017. The AAR of The Web Conference was relatively high (31%) in 1996, then began to decrease until it reached 17% in 2017. The AAR of ICSE dramatically decreased from 24% in 1996 to only 9% in 2006, then increased to 15% in the next year and slightly increased to 16% in 2017. A reason for decreasing acceptance rates is the increasing number of submissions, with the number of presentation slots at an event being constant over time.

Continuity. The continuity of TheWeb and ISWC (International Semantic Web Conference) is 100%, whilst they were held every year since their inception. On the other hand, the continuity of CHI (Conference on Human Factors in Computing Systems) is 97% because it was held every year since 1982 except for 1984. We observed that some events, such as ASPLOS and EDBT, have changed the regularity from $R = 2$ to $R = 1$ due to the high demand of submissions. Therefore, we computed the average of the continuity within each of these periods. For instance, EDBT (International Conference on Extending Database Technology) had a regularity of 2 in the period 1988–2008, and then it continued to convene every year. Overall, we observed a very high continuity among the most renowned events (Table 4.9).

Event Progress ratio. We calculated the PR of the top-events in each Computer Science community in the period 1997–2016. As shown in Figure 4.20, the EPR of the top-5 events had a slight rise in the period 1997–2005; then, they all rose noticeably in the last decade. Overall, events of all Computer Science communities have shown a drastic increase in PR since the beginning, notably, CCS and CHI.

H5-index. Figure 4.21 shows the top-5 events in five Computer Science communities according to their h5-index, calculated in 2016. The conference with the highest h5-index among all the fields is ECCV (European Conference on Computer Vision) with 98 (in the CV field), and TheWeb comes next with 77 (in the Web technologies field). Overall, we observed that the renowned events in Computer Science usually have an h5-index greater than 20.

Geographical distribution. For country-level GD, VLDB and TheWeb have $\bar{x} = 1$, which means that they moved to a different country each year, while SP (IEEE Symposium on Security and Privacy) and NDSS have $\bar{x} = 0$, which means that they stayed in the same country every year. For continent-level GD, ESWC (European/Extended Semantic Web Conference) and NDSS were always held in Europe and North America, respectively, while ICDE (International

Table 4.9: **Continuity of top-5 events in five Computer Science communities**. The regularity column shows the most recent regularity of each event.

Series	Field	Start	Age	Editions	Regularity	Continuity
TheWeb	WEB	1995	24	24	1	100%
WSDM	WEB	2008	11	11	1	100%
ISWC	WEB	2002	17	17	1	100%
ESWC	WEB	2004	15	15	1	100%
ICWS	WEB	1995	24	24	1	100%
ECCV	CV	1992	27	14	2	100%
CHI	CV	1982	37	36	1	97%
UIST	CV	1988	31	31	1	100%
BMVC	CV	1987	32	32	1	100%
ACMMM	CV	1993	26	26	1	100%
ICSE	SE	1995	24	24	1	100%
PLDI	SE	1987	32	32	1	100%
ASPLOS	SE	1982	37	23	1	98%
POPL	SE	1973	46	45	1	98%
ASE	SE	1991	28	28	1	100%
VLDB	DM	1985	34	35	1	100%
EDBT	DM	1988	31	21	2	100%
PKDD	DM	1997	22	22	1	100%
PODS	DM	1982	37	37	1	100%
ICDT	DM	1986	33	21	2	100%
CCS	SEC	1993	26	26	1	100%
USENIX	SEC	1990	29	28	1	97%
NDSS	SEC	1993	26	25	1	96%
EUROCRYPT	SEC	1982	37	37	1	100%
CRYPTO	SEC	1995	24	24	1	100%

Conference on Data Engineering) alternatively moved across continents, i.e., North America, Europe, and Asia.

Time distribution. We computed the frequency of occurrence of the top-5 events (identified using the SER ranking) for each event each month of the year since its establishment. Table 4.10 shows the most frequent month in which events take place along with the percentage of occurrence in this month. We observed that most of the renowned events usually take place around the same month each year with a slight shift of maximum one month. For instance, 50% of the editions of TheWeb conference were held in May and 41% in April. The CVPR conference has been held 28 times (out of 31) in June, and the PLDI conference has been held 33 times (out of 36) in June. This helps potential authors to anticipate when the event will take place next year and thus helps them with the submission schedule organization.

Table 4.10: **Scientometric profile** of the top-5 events in each Computer Science community. N is the total number of editions. Prominent values are highlighted.

Comm.	Acronym	h5	CORE	Q	GII	TD	AAR	FP	PR	N	Birth	GD	Publisher	Sponsors
WEB	TheWeb	77	A*	A1	A++	May (50%)	17%	6.16%	8%	23	1995	USA (22%)	TheWeb	Google
	WSDM	54	A*	B1	A+	Feb (91%)	18%		14%	11	2004	USA (55%)	ACM	Google
	ISWC	40	A	A1	A+	Oct (63%)	24%		10%	21	1997	USA (57%)	Springer	Elseiver
	ESWC	40	A	A1	A	May (60%)	24%		11%	15	2004	Greece (60%)	Springer	Ontotext
	ICWS	26	A	A1	A	Jun (35%)	19%		17%	24	1995	USA (96%)	Springer	Springer
CV	ECCV	98	A	A1	A+	May(33%)	30%	5.25%	19%	14	1990	Germany(21%)	Springer	Google
	CHI	85	A*	A1	A++	Apr (64%)	22%		10%	35	1982	USA (57%)	ACM	Google
	UIST	44	A	A1	A+	Oct (58%)	21%		9%	31	1988	USA (68%)	IEEE	Autodesk
	BMVC	43	–	A2	A	Sep (89%)	41%		6%	32	1987	UK (97%)	Springer	Microsoft
	ACMMM	44	A*	A	A++	Oct (60%)	22%		11%	25	1993	USA (52%)	ACM	YouTube
SE	ICSE	68	A*	A1	A++	May (60%)	17%	8.38%	5%	23	1975	USA (43%)	ACM	Google
	PLDI	50	A*	A1	A++	Jun (92%)	21%		5%	31	1979	USA (68%)	ACM	Microsoft
	ASPLOS	50	A*	A1	A++	Mar (43%)	22%		8%	23	1982	USA (87%)	ACM	Google
	POPL	46	A*	A1	A++	Jan (89%)	17%		11%	44	1973	USA (80%)	ACM	Microsoft
	ASE	31	A	A1	A	Sep (52%)	21%		6%	27	1991	USA (59%)	IEEE	Intel
DM	VLDB	73	A*	A1	A++	Aug (59%)	17%	8.35%	4%	33	1985	USA (12%)	VLDB	Google
	EDBT	32	A	A2	A	Mar (100%)	21%		4%	21	1988	Italy (19%)	OP	IBM
	PKDD	31	A	A2	A	Sep (86%)	26%		9%	22	1997	France (14%)	ACM	IBM
	PODS	26	A*	A1	A+	Jun (45%)	24%		5%	36	1982	USA (81%)	ACM	Oracle
	ICDT	20	A	B1	A-	Mar (100%)	34%		5%	20	1986	Italy (20%)	Springer	Oracle
SEC	CCS	72	A*	A1	A++	Oct (46%)	20%	9.23%	11%	25	1993	USA (72%)	ACM	NSF
	USENIX	61	A*	A1	A-	Aug (61%)	19%		10%	28	1990	USA (82%)	USENIX	Google
	NDSS	56	A	A1	A+	Feb (96%)	19%		12%	25	1993	USA (100%)	NDSS	Cisco
	EuroCrypt	53	A*	A1	A++	May (62%)	23%		6%	36	1982	France (11%)	Springer	Intel
	CRYPTO	53	A*	A1	A++	Aug (100%)	21%		7%	23	1995	USA (100%)	Springer	Google

4.5 Summary

We analyzed metadata of scholarly events of four scientific fields (Computer Science, Physics, Engineering, and Mathematics) involving conferences, workshops, meetings, and symposiums. We report the results of our analysis of events metadata within the two datasets in the last 30 years, according to the proposed metrics suite. The results we obtained from this study reveal that the long continuity of events highlights the importance of such events for the CS, MATH, PHY, and ENG communities. Furthermore, the increasing number of submissions and the growing progress ratio of Computer Science events provide clear evidence of the weight of scholarly events in different Computer Science communities. Researchers consider scholarly events as a serious gate to disseminate their research results. They consider specific characteristics to select the target venue. As a result of domain conceptualization to provide the foundation for this study, a comprehensive list of event-related properties provides empiric evidence on what makes an event high-ranked in its community. The results also shed light on the publication policies of researchers in CS, PHY, ENG, and MATH in terms of publication venue and citation count. In the last decades, we observed an increasing trend in both submissions and accepted papers in all Computer Science events. For instance, the average number of submissions, i.e., submissions to renowned events per year, to SEC events has doubled. We summarize the contributions in this chapter as follows:

- The conceptualization of the scholarly communication domain and the development of an event quality framework,
- The creation of a dataset of scholarly events belonging to four scientific fields, which imported to the scholarly event knowledge graph of OpenResearch.org,
- A metric suite (SEQA) based on the domain conceptualization, which contains newly defined metrics for scholarly events' impact assessment, including continuity, community popularity, field productivity, and progress ratio,
- An empirical evaluation of the quality of scholarly event metadata of CS, PHY, ENG, and MATH research communities involving different event types such as conferences, workshops, meetings, and symposiums,
- A methodology (SEMA) for data curation and metadata analysis of scholarly events, and
- Support for communities by giving recommendations to different stakeholders of their events.

Generally, the acceptance rate is considered one of the most important characteristics of scholarly events; however, the findings of this study indicate that the success of events depends on several other characteristics as well, such as continuity, the popularity of events' topics, and citations of published papers (reflected by the h5-index of the event). The findings are on the one hand interesting to observe the general evolution and success factors of scholarly events; on the other hand, they allow (prospective) event organizers, publishers, and committee members to assess the progress of their event over time and compare it to other events in the same field. They also help researchers to make more informed decisions when selecting suitable venues for presenting their work. SEAQ can be used in further events' metadata analysis and for multi-criteria events ranking. After integrating and analyzing the results we obtained from this study, we found that the most noteworthy findings to record are:

- During data acquisition, we observed that there is not much information about events before 1990, in particular on the number of submissions and accepted papers. In addition, we observed that there is no much historical information available about publications in the PHY, ENG, and MATH fields,

- Among all considered Computer Science communities, SEC has the largest average h5-index, while CSO has the smallest one,
- The number of submissions has kept growing over the last five decades, while, roughly speaking, the acceptance rate has remained the same. The reason may be the digitization of scholarly communication.
- The average acceptance rate for all events, since the first edition, falls into the range 15% to 31%,
- ACM publishes most of the proceedings of the event, and IEEE comes next.
- Most editions of the top events in all communities have been held around the same time of the year with similar deadlines,
- Most of those events that are high-ranked and have a high h5-index also have a long continuity (greater than 90%),
- Among all countries hosting events, the USA has hosted about 50% of the scholarly events in all communities in the last two decades,
- The field productivity of all Computer Science communities continuously increased since 2008, and the top events kept the trend of acceptance rates mostly stable over time regardless of the number of submissions,
- Based on the SJR indicator, the CS and MATH communities are more prolific, and their publications have more citations among each other, compared to PHY and ENG,
- The CV community had the highest number of submissions and accepted papers during the three 5-year time windows,
- The Computer Science community has the largest number of events with h5-index exceeding 30 compared to other communities, which can be attributed to scholarly events having an even more important role in Computer Science, and
- Most of the research findings of non-CS communities were published as abstracts or posters, while research findings of Computer Science were published as full research articles in formal proceedings.
- The progress ratio of all events kept growing over the last two decades, most likely thanks to the digitization of scholarly communication,
- The USA has 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,
- 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 Computer Science events.

Based on these findings, a set of recommendations has been concluded to different stakeholders, involving event organizers, potential authors, and sponsors.

Organizers: The possibility of having a progress ratio overview of other events enables organizers to compare their event with competing events and to identify organizational problems, e.g., publicity issues, the reputation of the members, and location dynamics. Therefore, in order to provide a high-profile event to the community, following specific strategies to comply with the characteristics of high-ranked events is necessary, e.g., keeping event topic coverage up to date with new research trends, involving high-profile people and sponsors, maintain a high continuity of the event, increasing the geographic distribution of event venues, and minimizing the time distribution. In some cases, e.g., ISWC 2019, which has been held in New Zealand, visa restrictions have prevented many participants from attending the conference, which can apparently affect the choice of the hosting country.

Potential authors: Community productivity and popularity change the research direction of individual scientists. Submitting to events with a broad range of the newest topics keeps the research productivity and publication profile of researchers aligned with growing communities. While searching for a venue to submit research results, considering characteristics of renowned events may influence future visibility and the impact of the submission if accepted.

Sponsors and proceedings publishers: The progress ratio of renowned events and considered characteristics gives insights about events of small size or preliminary events. Sponsoring such small size, but reliable and valuable events may support their rapid growth and may influence the popularity and overall direction of the associated research topics. This study helps to shed light on the evolving and different publishing practices in various communities and helps to identify novel ways for scholarly communication, such as the blurring of journals and conferences or open-access overlay-journals as they already started to emerge. In addition, we anticipate that the findings will encourage researchers in MATH, ENG, and PHY to publish and archive more information about their events, which will help in the events' metadata analysis. Finally, we believe that this work can provide foundations for discovery, recommendation, and ranking services for scholarly events with well-defined, transparent measures.

CHAPTER 5

Publishing Scholarly Communication Metadata as Linked Open Data

Scientific events have become a vital factor in scholarly communication for many scientific domains. They are considered as the focal point for establishing scientific relations between scholarly objects such as people (e.g., chairs and participants), places (e.g., location), actions (e.g., roles of participants), and artifacts (e.g., proceedings) in the scholarly communication domain. Metadata of scientific events have been made available mostly in unstructured or semi-structured formats, which impedes the discovery of interconnected and complex relationships between them and prevents transparency. To facilitate the management of such metadata, the representation of event-related information in an interoperable form requires uniform conceptual modeling.

As mentioned in chapter 3, several data models have been developed for describing events, such as the Event Ontology (EO), Linking Open Descriptions of Events (LODE), the Simple Event Model (SEM), and the Semantic Web Dog Food (SWDF). However, there is yet no standard, well-formed ontology covering all those aspects related to scientific events that are covered by the proposed ontology (in section 5.1), such as types of scientific events, sponsors, publishers, and proceedings.

In section 5.1, we present the Scientific Events Ontology (OR-SEO) in order to tackle the problem of representing scientific events metadata semantically, i.e., integrating existing events vocabularies and making explicit the relationships and interconnections between event data, thus supporting the transformation of from a “Web of documents” into a “Web of data” in the scientific domain and making it easier to efficiently query and process the data. In section 5.2, we introduce a Linked Open Dataset, which offers a comprehensive semantic description of the renowned Computer Science event series that took place in the last five decades.

The following research questions are investigated in this chapter:

> **RQ2**: How can we represent and integrate heterogeneous scholarly event metadata in knowledge graphs to facilitate scholarly data management and retrieval?

> **RQ3**: How can ontologies represent semantics encoded in entities involved in scholarly events domain and relationships among them?

The work presented in this chapter is based on the following publications:

- **Said Fathalla**, Sahar Vahdati, Christoph Lange and Sören Auer, *SEO: A Scientific Events Data Model.* In International Semantic Web Conference (ISWC), pp. 79-95, Springer, 2019.
- **Said Fathalla**, Sahar Vahdati, Sören Auer, and Christoph Lange. *The Scientific Events Ontology of The Openresearch.org Curation Platform.* In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (SAC), pp. 2311-2313. ACM, 2019.
- **Said Fathalla**, Christoph Lange and Sören Auer, *EVENTS: A Dataset on The History of Top-Prestigious Events in Five Computer Science Communities.* In Proceedings of Semantics, Analytics, Visualization (SAVE-SD) at the World Wide Web conference. Springer, Cham, pp. 110-120, 2017.
- **Said Fathalla**, and Christoph Lange. *EVENTSKG: A Knowledge Graph Representation for Top-Prestigious Computer Science Events Metadata.* In International Conference on Computational Collective Intelligence (ICCCI), pp. 53-63. Springer, Cham, 2018.
- **Said Fathalla**, Christoph Lange, and Sören Auer. *EVENTSKG: A 5-Star Dataset of Top-Ranked Events in Eight Computer Science Communities.* In the European Semantic Web Conference (ESWC), pp. 427-442. Springer, Cham, 2019.

5.1 The Scientific Events Ontology

Scholarly information, emanating from scientific events, publishing houses and social networks (e.g., *ResearchGate*) is available online in an unstructured format (e.g., Call for Papers (CfP) emails) or semi-structured format (e.g., Event home page) which limits the visibility and hampers the discovery of interconnected relationships in humans as well as machines. This plethora of scientific literature and heterogeneity of the metadata makes it increasingly difficult to keep an overview of the current state of research. Therefore, establishing a knowledge-based representation of information in scholarly communication motivates the development of data models, ontologies, and knowledge graphs. Semantically enriched representation of such information makes it easier to efficiently query and process the data [14]. Consequently, collecting, integrating, and analyzing the metadata of scientific events, such as association with an event series, important dates, submitted and accepted articles, venue, event type, or the field of research, is of paramount importance for tracking the scientific progress. An important topic in semantic publishing is the development of semantic models related to various scholarly communication elements in order to describe the meaning and the relationships between data, thus enabling machines to interpret meaning. These models are crucial for facilitating scholarly data management and supporting information retrieval, which meets the needs of various stakeholders, including authors and publishers [15]. Given the heterogeneity of event metadata as input, semantic representation of such information involves modeling event metadata covering different types of entities involved, such as persons, organizations, location, roles of persons before/during/after the event, etc.

In this section, we present OR-SEO (with the namespace prefix `seo`), which enables a semantically enriched representation of scholarly event metadata, interlinked with other datasets and knowledge graphs. OR-SEO does not only represent what happened, i.e., time and place of a scholarly event, but also the roles that each agent played, and the time at which a particular agent held this role at a specific event. OR-SEO is now in use on thousands of *OpenResearch.org* events pages, which enables users to represent structured knowledge about events without having to deal with technical implementation challenges and ontology development themselves.

Standard methodologies and best practices have been considered when designing and publishing the ontology. OR-SEO has been developed using the Simplified Agile Methodology for Ontology

Development (SAMOD) [173], an iterative process that aims at building the final model through a series of small steps. OR-SEO has been designed with a minimum of semantic commitment to guarantee maximum applicability for analyzing event metadata from diverse sources, and maximum reusability by datasets using the ontology for modeling different aspects of scientific events. Following best practices, OR-SEO emphasizes the reuse of events-related vocabularies, the alignment with concepts between them, as well as the reuse of ontology design and visualization patterns. OR-SEO is available using persistent identifiers (`https://w3id.org/seo#`). Future versions can be collaboratively revised on a corresponding Git repository (cf. Table 5.1), and it is registered and indexed by Linked Open Vocabularies (LOV)[1].

A set of SWRL rules has been defined in order to support knowledge discovery by automated reasoning. The validation of the ontology is performed on syntactic and semantic levels using the W3C RDF validation service and description logic reasoners, respectively. This step is crucial for making OR-SEO reusable. We shed light on what OR-SEO contributes to the existing literature by reviewing the existing event-related models and pointing out their weaknesses. Furthermore, the ontology is aligned with existing event ontologies. A public SPARQL endpoint to query the ontology is available online (cf. Table 5.1).

5.1.1 Design Principles and Requirements

OR-SEO is developed to be used as a reference ontology for the conceptualization of scholarly event metadata and capturing the corresponding concepts. It follows the best state-of-the-art practices and design principles for relevant and reusable ontologies. We first point out general design principles, then introduce the terms that we defined for representing the metadata of scientific events.

The best practices within the Semantic Web community have been followed from the initial steps of the OR-SEO development [174]. The paramount intention behind our decision to develop an ontology for scholarly events is that, to the best of our knowledge, there is a need for a well-formed ontology in this domain to describe scholarly events. In particular, aspects related primarily to scholarly events are not covered by existing ontologies, such as roles of organizers, e.g., proceedings chair, sponsors, event proceedings, and quality metrics such as acceptance rate and the ranking of the event. Inspired by Linked Data principles [4], the following design decisions have been made while developing OR-SEO:

- *Addressing different stakeholders:* OR-SEO is developed to be used in the OpenResearch.org platform, supporting, e.g., authors to find high-impact events to submit their work to, and event chairs and proceedings publishers to derive useful facts to assess the impact of their events and the competing ones.
- *Broad coverage of the relevant concepts:* An event, according to OR-SEO, comprises everything that happens in scientific events, no matter whether there is a specific place or time, or agents involved, including organization of sub-events as well as associated social events.
- *Flexibility and ease of changes:* The use of any class and their corresponding properties are optional, i.e., there are no property or cardinality restrictions such as `owl:allValuesFrom`.
- *Reusability:* We only use `rdfs:domain` and `rdfs:range` to indicate where to use properties. This facilitates the reuse of OR-SE.

[1] `https://lov.linkeddata.es/dataset/lov/vocabs/seo`

Table 5.1: **Resources.** The URLs of the the resources related to OR-SEO.

Resource	URL
PURL	`https://w3id.org/seo#`
Turtle file	`http://kddste.sda.tech/SEOontology/SEO.ttl`
RDF/XML file	`http://kddste.sda.tech/SEOontology/SEO.rdf`
GitHub repository	`https://github.com/saidfathalla/SEOontology`
Issue Tracker	`https://github.com/saidfathalla/SEOontology/issues`
SPARQL endpoint	`http://kddste.sda.tech/SEOontology/sparql`
VoID	`http://kddste.sda.tech/SEOontology/VoID.nt`

- *Efficient reasoning:* In the development of OR-SEO, several logic rules have been taken into consideration in order to facilitate efficient reasoning.
- *Availability:* The ontology has been published under a persistent URL (cf. Table 5.1) under the open CC-BY 3.0 license. OR-SEO is published according to the best practices of the Linked Data community [174]; its source is available from a *GitHub* repository (cf. Table 5.1). The ontology has been made discoverable through LOV, a high-quality catalog of well-documented vocabularies for data on the Web.
- *Validation:* Two types of validation have been performed: *syntactic* and *semantic* validation. We syntactically validated OR-SEO to conform with the W3C RDF standards using the online RDF validation service[2]. The dereferenceability of the URIs of the OR-SEO terms over the HTTP protocol (cf. [175]) has been validated using Vapour[3]. We semantically validated OR-SEO using Protégé reasoners such as FaCT++[4], and the OOPS! Ontology Pitfall Scanner[5], for detecting inconsistencies.
- *Documentation:* The documentation for the ontology is available online through its PURL. Detailed information about entities and properties are also included in the ontology, i.e., as `rdfs:comments`.
- *Adoption and Sustainability:* OR-SEO is maintained and used by the editors of OR to represent metadata of scientific events so far mainly in Computer Science but also some other fields, including physics and chemistry. OR-SEO also has an issue tracker on its GitHub repository in order to make it easier to request new features, e.g., the reuse related ontologies that may appear in the future, and to report any problems.
- *Metadata completion*: We followed the best practices for completing the vocabulary metadata proposed in [176].

5.1.2 Challenges and Requirements

Towards the development of an ontology for scholarly events, challenges started with identifying the pitfalls in the state-of-the-art model. In addition, the scholarly events domain itself relates entities from diverse information sources, including bibliographical information, spatial, and

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

[3] `http://linkeddata.uriburner.com:8000/vapour?`

[4] `https://github.com/ethz-asl/libfactplusplus`

[5] `http://oops.linkeddata.es/`

temporal data. Therefore, data models necessitate an effective integration of concepts and their semantics. After studying the domain and the state-of-the-art model, the diversity of information representation, and a large amount of data pose high requirements to be addressed by OR-SEO. The ontology should be maintainable concerning the evolution of linked data vocabularies and adaptable to other domains of science. A part of these requirements is represented as a set of competency questions related to different use cases that the ontology should be able to answer. A detailed version of these competency questions and the corresponding SPARQL queries are available.[6] However, some of these questions are

- *Which events related to the target domain X, e.g., "Semantic Web", took place in country Y over a particular time span, with an acceptance rate less than a value Z*?
- *What are the top-X countries hosting most of the events belonging to "Security and Privacy" in the past decade?*", and
- *In which events did person X participate in the organization committee*?

5.1.3 Reuse of Existing Ontological Knowledge

Techniques for efficient and effective reuse of ontological knowledge are crucial factors in developing ontology-based systems [177]. A challenging task for ontology engineers is to decide in advance which of the available vocabularies are the most useful ones for reuse, mainly because the Web allows reuse across domains. By its nature, the scientific events domain involves entities from various other areas, including location, agents, time, and scholarly data, as shown in Figure 5.1. Therefore, the first step in building our ontology is reusing terms from related ontologies, since the more vocabularies a model reuses, the higher the value of its semantic data is [178]. We have selected the most closely related ontologies listed in the Linked Open Vocabularies directory (LOV). This reuse of vocabularies by explicitly linking to them brings OR-SEO its richness. We reuse several well-known ontologies to make OR-SEO interoperable in different contexts:

- The *Semantic Web Conference (SWC) ontology*, one of the vocabularies of choice for describing academic conferences [123], is used to represent, e.g., `Conference`s and `ConferenceSeries`.
- *Time-indexed Value in Context (TVC)*, a standard ontology design pattern to describe a time-indexed situation that expresses a particular role held by an agent at an event [15], is used to represent, e.g., `Duration` and `Interval`.
- *Dublin Core Metadata Initiative (DCMI)* is used to describe metadata of typical entities in scientific events,such as agents and proceedings,
- The *Friend-of-a-Friend (FOAF)* ontology describes involved persons and their social network profiles,
- *Semantically-Interlinked Online Communities (SIOC)* describes information from online communities such as `Role` and `Site` [179]
- SPAR ontologies [134] describe the research papers type (FABIO), publications identifiers (DATACITE) and document parts (DOCO).
- SemSur ontology describes research findings based on an explicit semantic representation of the knowledge contained in scientific publications (see subsection 7.2.3), and

[6] `https://www.openresearch.org/wiki/Sparql_endpoint/Examples`

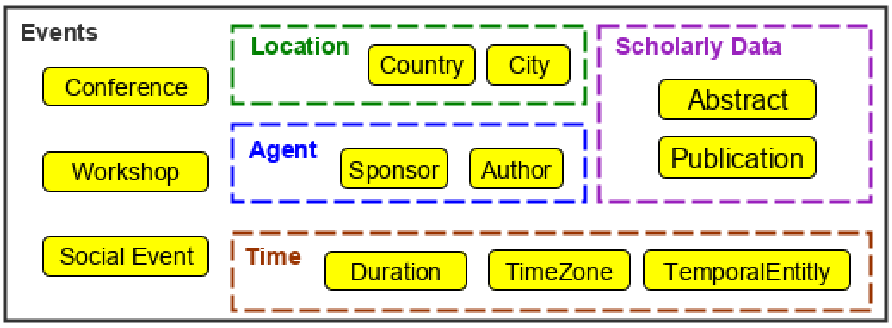


Figure 5.1: **Entities classification.** A domain-based classification of OR-SEO entities .

- DBpedia Ontology (dbo)[7] is used to represent geographical data, such as `dbo:Country` and `dbo:City`.

5.1.4 Ontology Description

The SAMOD [173] agile methodology is used for developing OR-SEO. SAMOD takes into consideration various issues when developing ontologies to achieve a “data-centric” model, such as avoiding inconsistencies, being self-explanatory, and giving examples of usage. This section describes the central entities in the ontology. We focus on core classes and properties, and reasoning support provided by the ontology. More details can be found in the online documentation of the ontology.

Core Classes

The OR-SEO ontology imports some of the main classes from the ontologies introduced in subsection 5.1.3. For the ones not explicitly matching with the concepts addressed by OR-SEO, new definitions have been developed. The core entities of the scholarly events in OR-SEO are: 1) *Event*, as the entity of main interest, including event types (e.g., conference or workshop) and metadata such as bibliographic and retrospective information (the numbers of submitted and accepted articles, information about the attendees, tracks), 2) *Agents*, including the *Organizations* hosting or sponsoring the event and *Persons* involved in the organization of the events in different roles, 3) *Role during event* of such stakeholders and persons, 4) *Location*, the city and country in which the event was held, 5) *Proceedings*, the proceedings produced by the event, and 6) *Time*, to describe the duration of events. Concretely, these entities are represented in OR-SEO as follows (see Figure 5.2 and Figure 5.3): `OrganizedEvent` represents the event itself and all the sub-events of those which are about the topic or theme of the main event, such as academic or non-academic events. `Agent` represents a person, group, company, or organization, which can be a sponsor or a publisher of the proceedings of the event. `RoleDuringEvent` represents a time-indexed situation that expresses a role held by an agent in the context of the event. Agents play different roles in scholarly events, including *Publishing Role During Event*, *Organizational Role During Event* and *Chair Role During Event.* `Country/City` represents the physical location of the event. `Proceedings` represents proceedings produced by academic events. `TemporalDuration` is a time interval representing the duration of the event.

[7] `http://dbpedia.org/ontology/`

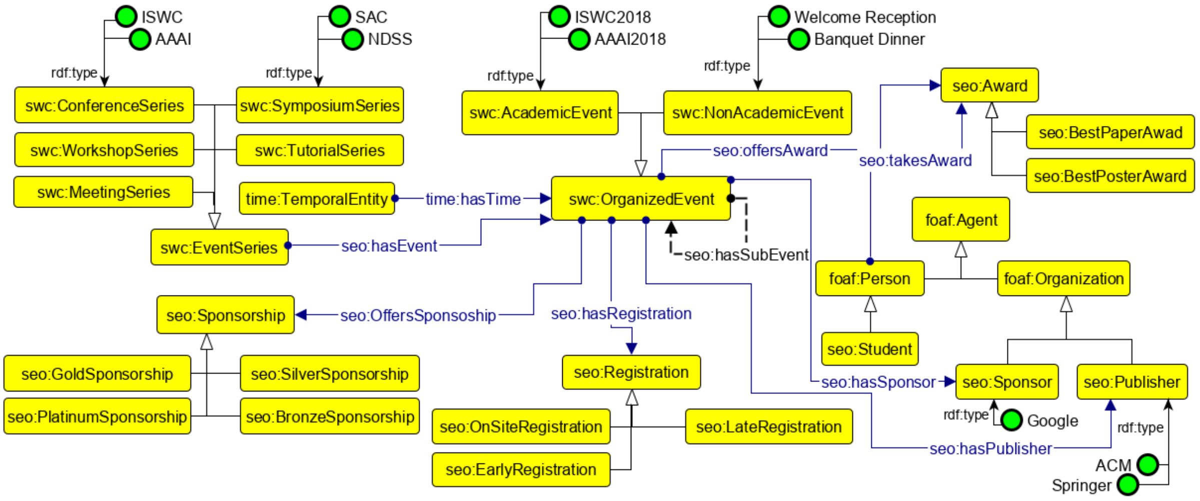


Figure 5.2: **Core concepts** in OR-SEO and their relationships. Arrows with open arrow heads denote `rdfs:subClassOf` properties between the classes.

Class Specialization: Because of the complexity and diversity of the concepts, some defined or reused classes need more specialization, so we created respective subclasses. For instance, `Symposium` has been added as a subclass of the `AcademicEvent` class. Another subclass (i.e., `SymposiumSeries`) to represent the series of such event type has been added to the superclass `EventSeries`. Moreover, a set of classes that is missing in the reused ontologies has been defined, for example, to describe agents and their roles more specifically, such as `Publisher` and `Sponsor`.

Class Disjointness: We assert pairwise disjointness, where applicable, between any of the classes in the ontology. For instance, the `IrregularRegistration` class is disjoint with `RegularRegistration`, and `LateRegistration` is disjoint with `EarlyBirdRegistration`.

Properties

OR-SEO's properties are divided into two categories: newly defined properties and directly reused properties. We indicate the classes to be used with several data and object properties by defining the domain and the range of these properties using `rdfs:domain` and `rdfs:range`. For instance, we capture the domain of newly-defined data properties for describing abstract and submission deadline, i.e., `seo:abstractDeadline` and `seo:submissionDeadline`, to be `swc:AcademicEvent` and the range to be `xsd:dateTime`. In addition, OR-SEO defines its own object properties, such as `seo:belongsToSeries`, `seo:hasTrack`, `seo:colocatedWith`, `seo:hasPublisher`. Some properties have complex ranges, e.g., the range of `seo:hasRegistrationType` is (`LateRegistration` ⊔ `EarlyBirdRegistration`) because these two classes are disjoint. Ontology design patterns are applied, e.g., the OWL patterns of Gangemi [180], such as n-ary relation pattern, to capture notions such as inverse relations and composition of relations. There are some inverse relations, e.g., `seo:isTrackOf` is the inverse of `seo:hasTrack` and `seo:isSponsorOf` is the inverse of `seo:hasSponsor`. Thus, if an event E `seo:hasTrack` T, then it can be inferred that T `seo:isTrackOf` E. Also, some symmetric relations are defined, such as `seo:colocatedWith`, e.g., if an event E_1 is co-located with another one E_2, then it could be inferred that E_2 is also co-located with E_1. Furthermore, it is a property whose domain is the same as its range, which provides the information that an organized event

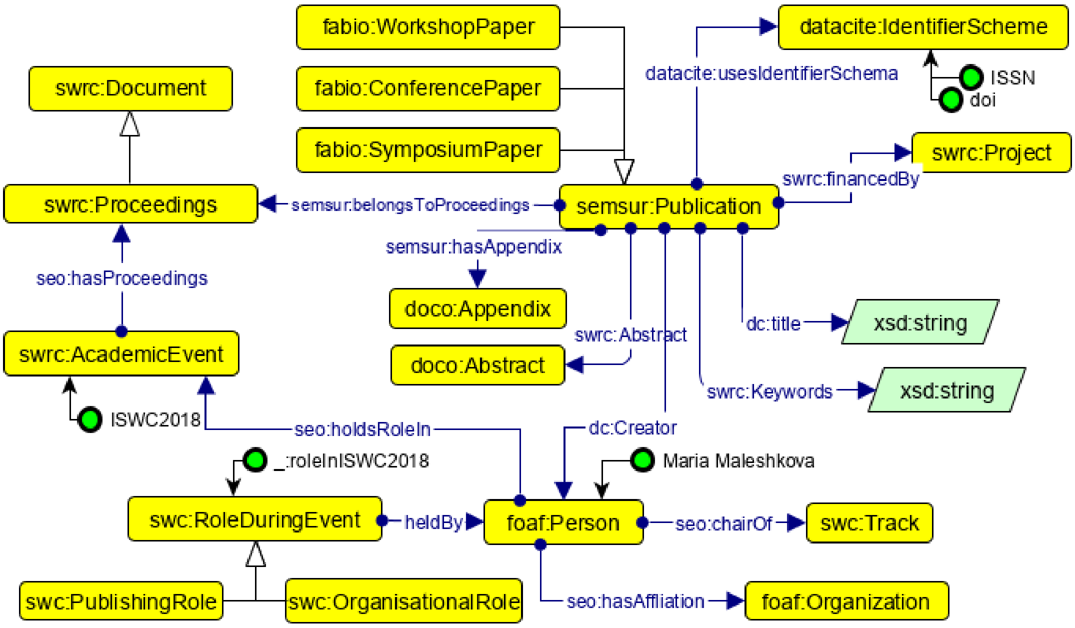


Figure 5.3: **Publications and roles** of agents during a scientific event.

can only be co-located with another organized event, and a reflexive relation, i.e., each event is co-located with itself. Such definitions allow to reveal implicit information and increase the coherence and thus the value of event metadata.

Representation pattern of n-ary relations. One common representation of n-ary relations is to represent a relationship as a class rather than property and using *n* properties to point to the related entities. Instances of such classes are instances of the n-ary relation, and additional properties can provide binary links to each argument of the relation, i.e., an instance of the relation linking *n* individuals. For more illustration, consider the case of *Maria Maleshkova*, the sponsorship chair in the ISWC conference in 2018. As shown in Figure 7.4, the individual *:roleInISWC2018* represents a single object encapsulating both the event, the person that had a role there, and the type of the role in that event.

Reasoning

Inference on the Semantic Web is additionally used to improve the quality of data integration in the ontology by combining rules and ontologies to discover new relationships, detect possible inconsistencies and infer logical consequences from a set of asserted facts or axioms. The Drools reasoner [181] is one of the reasoners that the Protégé ontology development environment uses for performing rule-based inference.

Our goal is to define a rule set for discovering new relationships and inferring new knowledge that did not explicitly exist in a knowledge graph. Therefore, a set of rules following the Semantic Web Rule Language (SWRL) [182] has been defined and written using the SWRLtab plugin for Protégé 5.2[8]. SWRL allows users to write Horn-like rules expressed in terms of OWL classes and properties to reason about OWL individuals. A set of rules to support the inference in

[8] `https://github.com/protegeproject/swrltab-plugin`

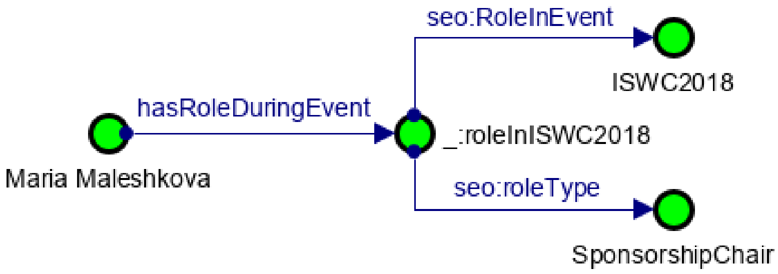


Figure 5.4: **N-ary relations.** Representation pattern of n-ary relations in OR-SEO.

OR-SEO has been defined. These rules have been semantically validated using Drools reasoner[9]. The rules can be used to infer new knowledge from existing OWL knowledge bases. According to O'Connor et al. [183], SWRL rules are written as antecedent (body)/consequent (head) pairs. Head and body consist of a conjunction/disjunction of one or more atoms. The rule set in OR-SEO contains the following SWRL rules (for readability, we omitted namespaces). Using Formula 5.1, participants in a specific event can be easily inferred while using Formula 5.2, the location of one event can be determined from a co-located event.

$$\mathit{Agent}\,(?a) \wedge \mathit{holdsRole}\,(?a, ?e) \rightarrow \mathit{participatesIn}\,(?a, ?e) \tag{5.1}$$

$$\mathit{colocatedWith}\,(?e1, ?e2)\ \wedge\ \mathit{hasLocation}\,(?e1, ?l) \rightarrow \mathit{hasLocation}\,(?e2, ?l) \tag{5.2}$$

5.1.5 Real-world Use cases

This section presents two real-world use cases for the OR-SEO ontology; OpenResearch.org and the EVENTSKG (cf. section 5.2) dataset.

Use case 1. As populating ontologies with instances is a time-consuming and error-prone task, OR-SEO is in use on 6,000+ event pages on *OpenResearch.org*, which facilitates the creation of instances of events and event series as wiki pages, without having to go into the details of the implementation of the ontology. It is an extended version of the original ontology of *OpenResearch*, which has been redesigned and systematically validated. Data acquisition in *OpenResearch* follows an approach that combines manual/crowd-sourced contribution and semi-automated methods. *OpenResearch* provides semantic descriptions of scientific events, publications, tools, and organizations using ontologies for each such entity type. OpenResearch employs one semantic form per core class of OR-SEO, combined with properties; they enable semantic annotations in the wiki markup. Semantic forms enable users to create and modify the knowledge graph via forms, without the need for actual programming. Listing 5.1 shows an example of an individual event (ISWC 2015) created on *OpenResearch*[10]. Furthermore, the semantically annotated text, found at the end of the wiki page of the ISWC series[11], represents the metadata of the event using corresponding terms of the ontology, such as chairs, country, or Twitter account. For instance, the *info box* on the right contains the metadata of the event series, including full title, bibliography, CORE ranking, and the average acceptance rate. Semantically

[9] `https://github.com/protegeproject/swrlapi-drools-engine/wiki/SWRLAPI-Drools-Engine`
[10] `http://openresearch.org/wiki/ISWC_2015`
[11] `https://www.openresearch.org/wiki/ISWC`

annotated metadata can be exported as RDF triples using the *"RDF feed"* feature. Interesting information can be exposed from OpenResearch, such as a list of upcoming events in a Calendar view[12], and top-ranked events along with their ranking and average acceptance rate[13].

Use case 2. The second use case of OR-SEO is the representation of a comprehensive dataset (EVENTSKG) of scholarly events sourced from several resources and curated semi-automatically (see section 5.2 for more details). Going beyond existing work (cf. chapter 3), EVENTSKG comprises metadata of 73 renowned events in eight Computer Science communities using OR-SEO as its schema. EVENTSKG is not only able to answer quantitative questions, but it also provides qualitative information, such as which countries hosted most events in a particular community (see section 5.2).

```
{{Event
    |Acronym=ISWC 2015      |Title=14th International Semantic Web Conference
    |Series=ISWC            |Type=Conference
    |Field=Semantic Web     |Start date=2015/10/11
    |End date=2015/10/15    |Homepage=iswc2015.semanticweb.org
    |City=Bethlehem         |State=PA
    |Country=USA            |Submission deadline=2015/04/30
    |Submitted papers=172   |Accepted papers=38
}}
```

Listing 5.1: **Use case 1.** Representation of metadata on OpenResearch.org in its markup language.

```
### https://w3id.org/seo#ISWC2015
eventskg:ISWC2015  rdf:type, conference-ontology:Conference;
     seo:belongsToSeries            eventskg:ISWC ;
     seo:acceptanceRate             "0.22"^^xsd:decimal;
     seo:submittedPapers            "172"^^xsd:integer;
     seo:acceptedPapers             "38"^^xsd:integer;
     seo:city                       <http://dbpedia.org/page/Bethlehem,_Pennsylvania>;
     seo:country                    <http://dbpedia.org/page/United_States>;
     seo:field                      seo:InformationSystems ;
     conference-ontology:startDate  "2015-10-11"^^xsd:date;
     conference-ontology:endDate    "2015-10-15"^^xsd:date;
     seo:eventWebsite               "http://iswc2015.semanticweb.org/"^^xsd:anyURI.
```

Listing 5.2: **Use case 2.** Using OR-SEO in metadata representation for ISWC 2015 in EVENTSKG, in Turtle.

The aim is to transform event metadata, distributed across different sources to Linked Open Data, which can be interpreted by machines to create innovative event-related services. Listing 5.2 shows the metadata of ISWC 2015 in EVENTSKG. Three major prefixes are used in metadata representation, namely: `eventskg`, `seo` and `conference-ontology` according to `http://prefix.cc/`.

5.1.6 Evaluation

Evaluating ontologies is the process of measuring the quality of the ontology content, ensuring that its definitions satisfy the requirements or perform correctly in the real world [184]. In other words, the quality of ontologies can be assessed using metrics that evaluate the success of the ontology in modeling a real-world domain (as illustrated in subsection 5.1.5). Ontologies can

[12] https://www.openresearch.org/mediawiki/index.php?title=Events_Calendar&field=Science
[13] https://www.openresearch.org/mediawiki/index.php?title=Series&field=Science

Table 5.2: **OR-SEO Evaluation.** The evaluation of OR-SEO using OntoQA model.

Ontology	Classes	Sub-classes	Attributes	Relations	AR	RR	IR
SWC	390	351	118	189	0.30	0.40	0.90
SEDE	122	46	47	56	0.39	0.60	0.38
SWRC	248	221	51	57	0.21	0.21	0.89
OR-SEO	165	197	93	177	**0.57**	**0.61**	**1.19**

be evaluated against a gold standard, or using a criteria-based or task-based evaluation [185]. This is majorly a manual task because it is challenging to construct automated tests to compare ontologies using such criteria [186]. We assess OR-SEO using a criteria-based evaluation, as proposed by Tartir et al. [187]. They proposed an ontology evaluation model, called OntoQA, which evaluates the ontology using schema metrics and instance metrics. We evaluate the ontology design by comparing it to the related work (with the best coverage of the domain, i.e., SWC, SEDE, and SWRC).

- *Attribute richness (AR)* refers to the average number of attributes (i.e., OWL datatype properties) per class. Formally, $AR = A/C$, the number of attributes for all classes (A) divided by the number of classes (C). The more attributes are defined, the more knowledge the ontology provides.
- *Relationship richness (RR)* refers to the diversity of relations and the placement of them in the ontology. Formally, $RR = R/(S + R)$, the number of relationships (R) defined in the schema, divided by the sum of the number of sub-classes (S), i.e., classes defined as sub-classes using `rdfs:subClassOf`, and the number of relationships. The more relations, except hierarchical relations, the ontology has, the richer it is.
- *Inheritance richness (IR)* refers to the average number of sub-classes per class. Formally, $IR = S/C$, the number of sub-classes divided by the sum of the number of classes. A high IR means that ontology represents a wide range of general knowledge, i.e., is of a horizontal nature.

As shown in Table 5.2, OR-SEO has a moderate size, but an overall beneficial knowledge structure. Among similar domain ontologies, it has the largest AR, which enables the provision of more knowledge per instance. Regarding RR, OR-SEO has a moderate diversity of relations and has much richer relations in comparison with SWC and SWRC, and slightly richer than SEDE. Regarding IR, OR-SEO has the highest value of all ontologies (1.19), which means that it represents a broader range of knowledge than state of the art. In terms of usability evaluation, most of the users of OpenResearch found it easy to populate the ontology via a user-friendly interface, i.e., SMW semantic forms. For instance, event organizers, or even any researcher interested in an event, can add event series or an individual event metadata using "*Add event series*"[14] and "*Add event*"[15] semantic forms, respectively. As mentioned before in subsection 5.1.5, the produced data are wiki pages presenting events metadata in a user-friendly way.

[14] `https://www.openresearch.org/wiki/Special:FormEdit/EventSeries`

[15] `https://www.openresearch.org/wiki/Special:FormEdit/Event`

5.1.7 Summary

This section presented OR-SEO, a reference ontology for capturing metadata of scientific events. Its real-world instantiation in the OpenResearch platform is discussed. Inference rules to discover new relationships, detect possible inconsistencies, and infer logical consequences from a set of asserted facts have been defined. We shed light on what OR-SEO contributes to the existing literature by reviewing the current event-related models, pointing out their weaknesses. OR-SEO covers issues closely related to scholarly events, which are not covered by other scholarly communication domain ontologies, such as types of scholarly events, sponsors, publishers, and proceedings. Furthermore, OR-SEO models scholarly events characteristics, such as acceptance rate, submission deadline, and notification date, and social media presence, e.g., Twitter accounts. The ontology is publicly available online, following ontology resource publication best practices. The ontology will continue to be maintained and extended in the context of the *OpenResearch* effort, aiming at large scale event data acquisition and analysis through applying semi-automated and crowd-sourcing methods. We hope that OR-SEO will thus contribute to facilitating the representation and analysis of the currently not yet well-structured space of scholarly event information, thus supporting all stakeholders of events, particularly organizers and potential authors.

5.2 Scholarly Events Datasets

Information emanating from scientific events, journals, organizations, institutions as well as scholars has become increasingly available online. Therefore, there is a high demand to assess, analyze, and organize such a massive amount of data produced every day. It is of primary interest to event organizers, as it helps them to assess the progress of their event over time and compare it to competing events. Furthermore, it helps potential authors looking for venues to publish their work.

In this section, we present the three releases of the scholarly events dataset (EVENTSKG), which contains historical data about the publications, submissions, start date, end date, location, and homepage for the most renowned event series belonging to eight Computer Science communities. Each of these releases is explained in more detail later in this chapter. Table 5.3 summarizes the differences between the three releases.

- *EVENTS* (the former name): contains historical data about the publications, submissions, start date, end date, location and homepage for 25 renowned event series (718 editions in total) in five Computer Science communities,
- *EVENTSKG* 1.0: offers a comprehensive semantic description of scientific events of six Computer Science communities for 40 renowned event series over the last five decades, and
- *EVENTSKG* 2.0: the new release of the EVENTSKG dataset, a 5-star dataset containing metadata of 73 top-ranked event series (almost 2,000 events) established over the last five decades.

Overview

The main goal of the development of EVENTSKG is to facilitate the analysis of events metadata, by enabling them to be queried using semantic query languages such as SPARQL. A key research

Table 5.3: **Statistics** about the three releases of EVENTSKG. due to subsequent updates, some measure are different from the ones in the published articles of these releases. N is the number of individual events.

Dataset	Series	N	Attr.	Fields	Triples	Format	Structure	API	LOD
EVENTS	25	718	14	5	9,810	RDF, CSV, XML	Individual RDF dumps	✗	✓
EVENTSKG 1.0	40	1048	15	6	13,952	RDF/XML, CSV, JSON-LD, Turtle	RDF graph	✗	✓
EVENTSKG 2.0	73	2000	15	8	33,236	RDF/XML, CSV, JSON-LD, Turtle	RDF graph	✓	✓

question that motivates our work is: *What is the effect of digitization on scholarly communication in Computer Science events*? In particular, we address the following questions:

- What is the orientation of submissions and corresponding acceptance rates of renowned events in Computer Science?
- How did the number of publications of a particular Computer Science community fluctuate?
- Did the date of renowned events changes from year to year?
- Which countries host most events in different Computer Science communities?

Concerning the events' impact, we address the following questions:

- What are the high-impact events of Computer Science?
- How are the high-impact events currently ranked in the available ranking services?
- Which country has hosted most high-impact Computer Science events?

By analyzing the dataset content, we gain some insights to answer these questions. Exploratory data analysis is performed, aiming at exploring some facts and figures about Computer Science events over the last five decades. Top-40 renowned event series have been identified based on several criteria (see subsection 3.1.1 for more details). These event series fall into six Computer Science communities[16]: information systems (IS), security and privacy (SEC), artificial intelligence (AI), computer systems organization (CSO), software and its engineering (SE) and Web (WWW). Events are linked by research fields, the hosting country, and proceedings publishers. For instance, EVENTSKG can answer competency questions such as:

- What are the events related to "*Software Engineering*" with an acceptance rate of less than 20% and proceedings published by "*Springer*"?
- Which countries have hosted most of the events related to "*Semantic Web*" over the last 20 years?
- Which of the six Computer Science communities has attracted growing interest (in terms of the number of submissions) in the last ten years?
- Which of the six Computer Science communities has a growing production (in terms of the number of accepted papers) in the last ten years?

[16] Using ACM Computing Classification System: `https://dl.acm.org/ccs/ccs.cfm`

We believe that EVENTSKG closes an critical gap in analyzing the progress of a particular event series and Computer Science community, using renowned event series in the community, in terms of submissions and publications over a long-term period. Furthermore, it will have a far-reaching influence on the research community, in particular:

- *Event chairs* – to assess the progress/impact of the event,
- *Potential authors* – to find out events with high impact to submit their work,
- *Proceedings publishers* – to trace the impact of their events.

5.2.1 EVENTS Dataset

EVENTS dataset contains metadata, including the number of submitted papers and accepted papers, start date, end date, location and homepage, about 25 renowned event series (718 editions in total) belonging to five Computer Science communities (Information systems (IS), Security and privacy (SEC), Artificial intelligence (AI), Computer systems organization (CSO) and Software and its engineering (SE). The dataset is publicly available online in several formats (i.e., CSV 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 section 4.3, we shed light on this by conducting a scientometrics study of the events metadata in EVENTS over the past 50 years.

Characteristics of the EVENTS Dataset

EVENTS dataset covers historical information about 25 renowned 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 the peculiarities of the temporal and geographical evolution of the event series. EVENTS contains two types of events, i.e., conferences and symposia[17]. Entries refer to all available attributes of all events.

Use Case. EVENTS dataset enables event organizers and chairs 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 that EVENTS will assist researchers who want to submit a paper to be able to decide which events they could submit their work, e.g., answering questions, such as "which events in a particular Computer Science field have a high impact?". Moreover, when a specific event is held each year, it helps them to prepare their research within the event's usual timeline.

Extensibility. EVENTS can be extended in three dimensions to meet future requirements by 1) adding events belonging to other communities, 2) creating a Knowledge graph of the renowned events based on scientific events ontologies found in the literature, and 3) adding more attributes, such as hosting university or organization, sponsors, and event steering committees or program committee chairs.

Availability and License. EVENTS is available online[18] and registered in the GitHub repository[19]. It is subject to the Creative Commons Attribution license. The RDF version has

[17] It would be correct to label a symposium as a small scale conference as the number of participants is small.
[18] `http://kddste.sda.tech/EVENTSKG-Dataset/EVENTS-Dataset/EVENTS.html`
[19] `https://github.com/saidfathalla/EVENTS-Dataset`

been validated using W3C Validation Service[20].

Content Overview. After analyzing the metadata of all events in the dataset, the following are the most interesting observations:

- AI community has the largest average h5-index of 89.9; SEC comes second with 62. Surprisingly, despite the Qualis ranking of RecSys is *B1*, the h5-index of RecSys is relatively high, and it is ranked as *A* by CORE and as *A*- by GII,
- ACM publishes most of the proceedings of Computer Science events, and IEEE comes next. However, we observed that some events, such as NDSS and USENIX, publish their proceedings on their 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.
- The USA leads by far, having hosted most editions of CVPR, ISCA, VLDB, ICSE, and CCS. Canada comes in second, hosting most editions of ISCA, VLDB, and ICSE.
- The highest acceptance rate is the one of PODC of 31%, while PERCOM has the lowest one of 15%.

5.2.2 EVENTSKG 1.0

EVENTSKG 1.0 is the subsequent release of EVENTS. It contains a comprehensive semantic description of 40 renowned scientific events (approximately 60% additional event series have been added) belonging to six Computer Science communities over the last five decades. A notable feature of the new release is the use of our Scientific Events Ontology (described in section 5.1) as a reference ontology for event metadata modeling and connect related data that was not previously, i.e., in EVENTS, linked. EVENTSKG is a knowledge graph containing events metadata as a *unified graph* rather than individual RDF dumps for each event series in the previous release, i.e., EVENTS.

At the end of this section, we shed light on these events by giving some outstanding comparative information about the events themselves and the communities they belong to. Generally, the benefits of publishing data as linked data are:

- *Data linking*: establish links between dataset elements so that machines can explore related information,
- *Semantic querying*: Linked Data can be queried using the SPARQL query language,
- *Data enrichment*: inference engines can be used to infer implicit knowledge which does not explicitly exist,
- *Data validation*: semantically validate data against inconsistencies.

Dataset Characteristics

EVENTSKG 1.0 covers three types of events since 1969: conferences, workshops, and symposia. It contains metadata of 1048 editions of 40 event series with 15 attributes each. It is available in four different formats: RDF/XML, Turtle, CSV, and JSON-LD. The number of submissions and

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

Table 5.4: **Scientometric profile** of all events in EVENTS dataset. N is the number of editions in 2018.

Acro.	**Comm.**	**CORE**	**GII**	**Q**	**h5**	**N**	**AAR**	**Loc.**	**Month (Freq.)**	**Start**	**Publisher**
CVPR	AI	A*	A+	A1	158	28	0.33	USA	Jun (26)	1985	IEEE
NIPS		A*	A++	A1	101	32	0.25	USA	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	USA	Aug (16)	1969	AAAI
AAAI		A*	A++	A1	56	32	0.26	USA	Jul (20)	1980	AAAI
ISCA	CSO	A*	A++	A1	54	45	0.18	USA	Jun(27)	1973	IEEE
HPCA		A*	A+	A1	46	24	0.20	USA	Feb (17)	1995	ACM
FOCS		A	A++	A1	45	30	0.28	USA	Oct (25)	1989	IEEE
PERCOM		A*	A+	A1	31	16	0.15	USA	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	USA	Aug (20)	1985	VLDB
RecSys		A	A-	B1	34	12	0.26	USA	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	USA	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	USA	May (25)	1975	ACM
PLDI		A*	A++	A1	50	33	0.21	USA	Jun (33)	1979	ACM
ASPLOS		A*	A++	A1	50	23	0.22	USA	Mar (10)	1982	ACM
ICDE		A*	A+	A1	51	34	0.20	USA	Feb (14)	1984	IEEE
UIST		A*	A+	A1	44	31	0.21	USA	Oct (18)	1988	ACM
CCS	SEC	A*	A++	A1	72	25	0.22	USA	Oct (12)	1993	ACM
SP		A*	A++	A1	68	39	0.28	USA	May (31)	1980	IEEE
USENIX		A*	A-	A1	61	27	0.19	USA	Aug (17)	1990	USENIX
NDSS		A*	A+	A1	56	25	0.20	USA	Feb (24)	1993	NDSS
EuroCrypt		A*	A++	A1	53	37	0.24	France	May (23)	1982	Springer

publications of each event involves all tracks' submissions and publications. There are several challenges to pursuing the maintenance of EVENTSKG for the future and keeping it sustainable; here is how we address them:

- *Availability:* EVENTSKG 1.0 is publicly available online under a persistent URL (PURL): `http://purl.org/events_ds`. It is subjected to the Creative Commons Attribution license.
- *Validation:* we perform two types of validation: *syntactic* and *semantic* validation. We syntactically validate EVENTSKG to conform with the W3C RDF standards using the online RDF validation service[21] and semantically validate it using Protégé reasoners.
- *Documentation:* the documentation of the dataset has been checked using the W3C Markup

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

Validation Service[22] and is available online on the dataset Web page[23].

A concrete use case for querying EVENTSKG is supporting the research community in taking decisions on what event to submit their work to, or whether to accept invitations for being a chair or program committee member. For example, what are the events belonging to *Artificial Intelligence*, which took place in the USA along with their acceptance rate requires joins between field/topic and event entities? The SPARQL query in Listing 5.3 is used to answer such a question.

Data Curation

While collecting data from different sources, several problems have been encountered, such as data duplication, incomplete data, incorrect data, and the change of event title over time. Therefore, a data curation process has been carried out, comprising data acquisition, preprocessing, augmentation, Linked Data generation, data enrichment, and publication. The curation of EVENTSKG dataset is an incremental process involving: *Data acquisition and completion*, *Data Preprocessing*, *Data Augmentation*, *Linked Data Generation*, *Linked Data Enrichment*, and *Data Publication.* SEMA methodology (see section 4.2) has been followed in the curation process of EVENTSKG. For the creation of the linked dataset, after the preprocessing step the following steps have been carried out:

```
SELECT ?event ?title ?acc ?topic
WHERE {
  ?event    rdfs:label          ?title .
  ?event    seo:country         <http://dbpedia.org/resource/United_States> .
  ?event    seo:field           ?topic .
  ?event    seo:acceptanceRate  ?acc .
  ?topic    rdfs:label          "Artificial Intelligence" .
}
```

Listing 5.3: **SPARQL query** for finding all events belonging to Artificial Intelligence which took place in the USA along with their acceptance rate.

Data Augmentation. The objective of the data augmentation process is to add new events to the dataset and fill in missing data. To achieve this objective, we periodically explore online digital libraries for the missing information. The output of this process is structured data in CSV format.

Linked Data Generation. The objective of the Linked Data Generation process here is to generate linked data from unlinked data in the CSV format. The structured data, in CSV format, has been converted to RDF triples using an ad-hoc tool[24]. Consequently, the next step is to validate (i.e., Syntactic Validation) the produced data using a standard validation tool (e.g., W3C RDF online validation service). The representation of AAAI conference in 2018, in Turtle syntax, is presented in Listing 5.4.

Linked Data Enrichment. The Linked Data enrichment (LDE) process is vital in order to discover the interlinking relationships between RDF triples by using inference engines, i.e., reasoners. The input of LDE is the RDF triples produced by the Linked Data generation, and the output is a set of *consistent* RDF triples, including the newly discovered relationships, where available. Semantic inference can be used to improve the quality of data integration in a

[22] https://validator.w3.org/

[23] http://kddste.sda.tech/EVENTSKG-Dataset/EVENTS-Dataset/EVENTSKGV1.html

[24] http://levelup.networkedplanet.com/

dataset by discovering new relationships, detecting possible inconsistencies, and inferring logical consequences from a set of asserted facts or axioms in an ontology. To enrich and semantically validate RDF data, generated from the previous process, we use two reasoners integrated into Protégé, FaCT++[25] and HermiT[26], which support three types of reasoning: (1) detecting inconsistencies, (2) identifying subsumption relationships, and (3) instance classification [188]. Detecting inconsistencies is a crucial step in LDE because inconsistency results in false semantic understanding and knowledge representation. We resolve detected inconsistencies and rerun the reasoner to ensure that no other inconsistencies arise.

Data Publication. The goal of Linked Data publishing is to enable humans and machines to share structured data on the Web. EVENTSKG is published according to the Linked Data community best practices [174, 189] and registered in a GitHub repository (`https://github.com/saidfathalla/EVENTS-Dataset`). The final step is to index the dataset in a public data portal (e.g., DataHub), which is the fastest way for individuals and teams to find, share, and publish high-quality data online. EVENTSKG dataset is published at DataHub[27]. DataHub is a tool for publishing, distribution, and sharing data on the Web[190].

```
eventskg:AAAI2018 rdf:type, conference-ontology:Conference ;
    seo:belongsToSeries eventskg:AAAI ;
    seo:city "New Orleans" ;
    seo:country "USA" ;
    seo:field "Artificial Intelligence" ;
    seo:state "Louisiana" ;
    conference-ontology:acronym "AAAI" ;
    conference-ontology:endDate "2018-02-07T00:00:00.0000000+00:00"^^xsd:dateTime ;
    conference-ontology:startDate "2018-02-02T00:00:00.0000000+00:00"^^xsd:dateTime ;
    eventskg:EventWebpage "https://aaai.org/Conferences/AAAI-18/" .
```

Listing 5.4: **Resource in Turtle.** A resource representing metadata of the AAAI 2018 conference in Turtle format.

Dataset Content Analysis

Dataset contents have been analyzed to answer the research questions presented at the beginning of this section. Table 5.5 provides the results obtained from the preliminary analysis of the dataset. The remarkable observation to emerge from analyzing the dataset content are:

- There is a clear upward trend in the number of submitted and accepted papers during the whole period, while, roughly speaking, the acceptance rate remains the same.
- Prestigious events usually take place around the same month each year (i.e., usual month in Table 5.5). This helps potential authors to expect when the event will take place next year, which allows in the submission schedule organization. *Usual month* refers to the number of times an event has occurred in a specific month. Namely, CVPR conference has been held 26 times (out of 28) in June, while POPL has been held 41 times (out of 45).
- The average acceptance rate for all event series falls between 15% to 30%, except for FOGA, COLT, and IJCAR. Roughly speaking, FOGA has the highest acceptance rate of 59%, while PERCOM has the smallest one of 15%.

[25] `https://github.com/ethz-asl/libfactplusplus`

[26] `https://github.com/phillord/hermit-reasoner`

[27] `https://datahub.ckan.io/dataset/eventskg`

- CVPR has the largest h5-index of 158, while FOGA has the smallest one of 9. Among all considered Computer Science communities, SEC has the largest average h5-index of 58.16, while CSO has the smallest one of 40.2.
- The USA hosted most editions of events in all communities and most editions of all SE events. France comes second, having hosted most editions of PKDD and EuroCrypt.
- Several event series organizers publish the proceedings of their events on their digital library, such as AAAI, VLDB, and TheWeb. On the other hand, ACM publishes the proceedings of most events, and IEEE comes next.
- IJCAI is the oldest series since it has been established in 1969 (i.e., 50 editions), while RecSys is the most recent one since it has been established in 2007 (i.e., 12 editions).

Summary

The new release of the EVENTS dataset, called EVENTSKG 1.0, is a unified RDF graph of top-40 renowned events based on the SEO ontology. To the best of our knowledge, this is the first time a dataset is published as a knowledge graph of metadata of renowned events in IS, SEC, AI, CSO, SE, and WWW. EVENTSKG closes an important gap in analyzing the progress of a Computer Science community in terms of submissions and publications, and it is of primary interest to steering committees, proceedings publishers, and prospective authors (see chapter 4 for more details). The most striking findings to emerge from analyzing EVENTSKG content can be found in section 4.5. These findings highlight the usefulness of EVENTSKG for the event's organizers as well as other researchers in the community.

5.2.3 EVENTSKG 2.0

This section introduces the second release of the EVENTSKG dataset (i.e., EVENTSKG 2.0). The new release is a Linked Open Dataset adhering to an updated version of the Scientific Events Ontology, leading to richer and cleaner data. Currently, EVENTSKG contains 75% series in addition to the first release from eight Computer Science communities[28]: 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.

Overview

EVENTSKG is a 5-star dataset [62], i.e., following a set of design principles for sharing machine-readable interlinked data on the Web. Generally, a 5-star dataset enables 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 learn about the data directly, 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. Interlinking is required to achieve the 5^{th} star of the 5-star deployment scheme proposed by Berners-Lee [62]. Research fields and both countries and cities have been mapped to OR-SEO and DBpedia, respectively. As in the previous release, events are also linked by research fields, hosting country, and proceedings

[28] These communities have been identified using the ACM Computing Classification System: `https://dl.acm.org/ccs/ccs.cfm`

Table 5.5: **Scientometric profile** of events in EVENTSKG. N is the number of editions in 2018.

Acro.	**Comm.**	**CORE**	**GII**	**Q**	**h5**	**N**	**AAR**	**Loc.**	**Month (Freq.)**	**Start**	**Publisher**
IJCAR	AI	A*	A	B1	45	10	0.41	UK	Jul (4)	2001	ACM
COLT		A*	A+	A2	33	31	0.49	USA	Jun (11)	1988	PMLR
KR		A*	A+	A2	26	16	0.28	USA	Apr (4)	1989	AAAI
ISMAR		A*	A	A2	26	21	0.24	USA	Oct (10)	1999	IEEE
VR		A	A-	A2	17	25	0.26	USA	Mar (24)	1993	IEEE
FOGA		A*	A-	B3	9	14	0.59	USA	Jan (7)	1990	ACM
CVPR		A*	A+	A1	158	28	0.33	USA	Jun (26)	1985	IEEE
NIPS		A*	A++	A1	101	32	0.25	USA	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	USA	Aug (16)	1969	AAAI
AAAI		A*	A++	A1	56	32	0.26	USA	Jul (20)	1980	AAAI
ISCA	CSO	A*	A++	A1	54	45	0.18	USA	Jun (27)	1973	IEEE
HPCA		A*	A+	A1	46	24	0.20	USA	Feb (17)	1995	ACM
FOCS		A	A++	A1	45	30	0.28	USA	Oct (25)	1989	IEEE
PERCOM		A*	A+	A1	31	16	0.15	USA	Mar (16)	2003	IEEE
PODC		A*	A+	A1	25	37	0.30	Canada	Aug (19)	1982	ACM
PODS	IS	A*	A+	A1	26	37	0.24	USA	Jun (17)	1982	ACM
VLDB		A*	A++	A1	73	33	0.18	USA	Aug (20)	1985	VLDB
RecSys		A	A-	B1	34	12	0.26	USA	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	USA	Jun (18)	2000	IEEE
PKDD		A	A	A2	31	22	0.25	France	Sep (19)	1997	ACM
POPL	SE	A*	A++	A1	46	45	0.20	USA	Jan (41)	1973	ACM
ICSE		A*	A++	A1	68	24	0.17	USA	May (25)	1975	ACM
PLDI		A*	A++	A1	50	33	0.21	USA	Jun (33)	1979	ACM
ASPLOS		A*	A++	A1	50	23	0.22	USA	Mar (10)	1982	ACM
ICDE		A*	A+	A1	51	34	0.20	USA	Feb (14)	1984	IEEE
UIST		A*	A+	A1	44	31	0.21	USA	Oct (18)	1988	ACM
OOPSLA		A*	A++	A1	37	33	0.22	USA	Oct (26)	1986	ACM
OSDI	SEC	A*	A+	A1	39	13	0.16	USA	Oct (7)	1994	USENIX
CCS		A*	A++	A1	72	25	0.22	USA	Oct (12)	1993	ACM
SP		A*	A++	A1	68	39	0.28	USA	May (31)	1980	IEEE
USENIX		A*	A-	A1	61	27	0.19	USA	Aug (17)	1990	USENIX
NDSS		A*	A+	A1	56	25	0.20	USA	Feb (24)	1993	NDSS
EuroCrypt		A*	A++	A1	53	37	0.24	France	May (23)	1982	Springer
TheWeb	WWW	A*	A++	A1	75	23	0.17	USA	May (12)	1989	TheWeb
WSDM		A*	A+	B1	54	11	0.18	USA	Feb (10)	2008	ACM
ISWC		A	A+	A1	40	21	0.24	USA	Oct (12)	1997	Springer
ESWC		A	A	A1	40	15	0.25	Greece	May (9)	2004	Springer
ICWS		A	A	A1	26	25	0.21	USA	Jun (6)	1995	IEEE

publishers as well as sponsors. A key benefit of EVENTSKG is the availability of the dataset as LOD, as well as a collection of open-source tools for maintaining and updating it. The objective is to ensure the sustainability and usability of it, which significantly supports the analysis of scholarly events metadata. The documentation page (cf. Table 5.6) describes the dataset structure and its releases. It also contains a description of each release and a chart comparing the statistics of each release. The URI of each resource, i.e., of an individual event 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 ESWC conference in 2018. EVENTSKG stores data relevant to these events in RDF, and each event's metadata is described appropriately through employing the data and object properties in the Scientific Events Ontology. 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. Further new features of the new release include the use of the latest version of the Scientific Events Ontology, 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.

To illustrate the potential use of EVENTSKG for tracking the evolution of scholarly communication practices, we analyzed the key characteristics of scholarly events (using exploratory data analysis techniques) 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.

Competency Queries

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 decade?*" In addition, the productivity and the popularity of a Computer Science community over time can be analyzed by studying the number of accepted and submitted papers, respectively. For instance, "*Q_2: Compare the popularity of the Computer Science communities in the past decade*"(Listing 5.5). Regarding country-level analysis, the popularity of a Computer Science community in a particular country can be determined by such a query: *"Q_3: What are the top-5 countries hosting most of the events belonging to Security and Privacy in the past decade?"* Listing 5.6 shows the corresponding SPARQL query. In fact, EVENTSKG is not only able to answer quantitative questions, but it also provides qualitative information, such as countries that hosted most events related to a particular community.

Dataset Characteristics

Currently, EVENTSKG covers three types of Computer Science events since 1969[29]: conferences, workshops, and symposia. EVENTSKG contains metadata of 73 event 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.

```
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 5.5: **SPARQL query** for comparing the popularity of the Computer Science communities.

```
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
```

Listing 5.6: **SPARQL query** for finding top-5 countries that host most of the events belonging to *Security and Privacy* in the past decade.

EVENTSKG is a 5-star dataset [62] in which each resource is identified by a URI and links to other datasets on the Web, such as DBpedia (to represent countries) and OR-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 5.6.

- *Availability and Best Practices*: 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 [4]. 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. Besides, we established a SPARQL endpoint (using Virtuoso) to enable users to query the dataset. EVENTSKG 2.0 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 Computer Science communities, and c) add event properties, such as deadlines, registration fees, and chairs.
- *Documentation*: The documentation of the dataset is available online[30] and has been checked using the W3C Markup Validation Service[31].

[29] the date of the oldest events in the dataset

[30] `http://kddste.sda.tech/EVENTSKG-Dataset/`

[31] `https://validator.w3.org/`

Table 5.6: **Related Resources.** The URLs of the EVENTSKG 2.0 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`
OR-SEO Ontology	`http://purl.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/EVENTSKG_R2.html`

- *Sustainability*: To ensure the sustainability of EVENTSKG, an API for updating and maintaining the dataset has been developed. The dataset is replicated on its GitHub repository and our servers.
- *Announcement*: EVENTSKG has been announced on several mailing lists, such as the W3C LOD list[32], the discussion list of the open science community[33], 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 [30].

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.

During the curation process, several problems have been encountered, such as 1) identification of top-ranked events in each Computer Science 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. Therefore, the curation process of EVENTSKG 1.0 has been adapted to an eight-step incremental

[32] `public-lod@w3.org`

[33] `open-science@lists.okfn.org`

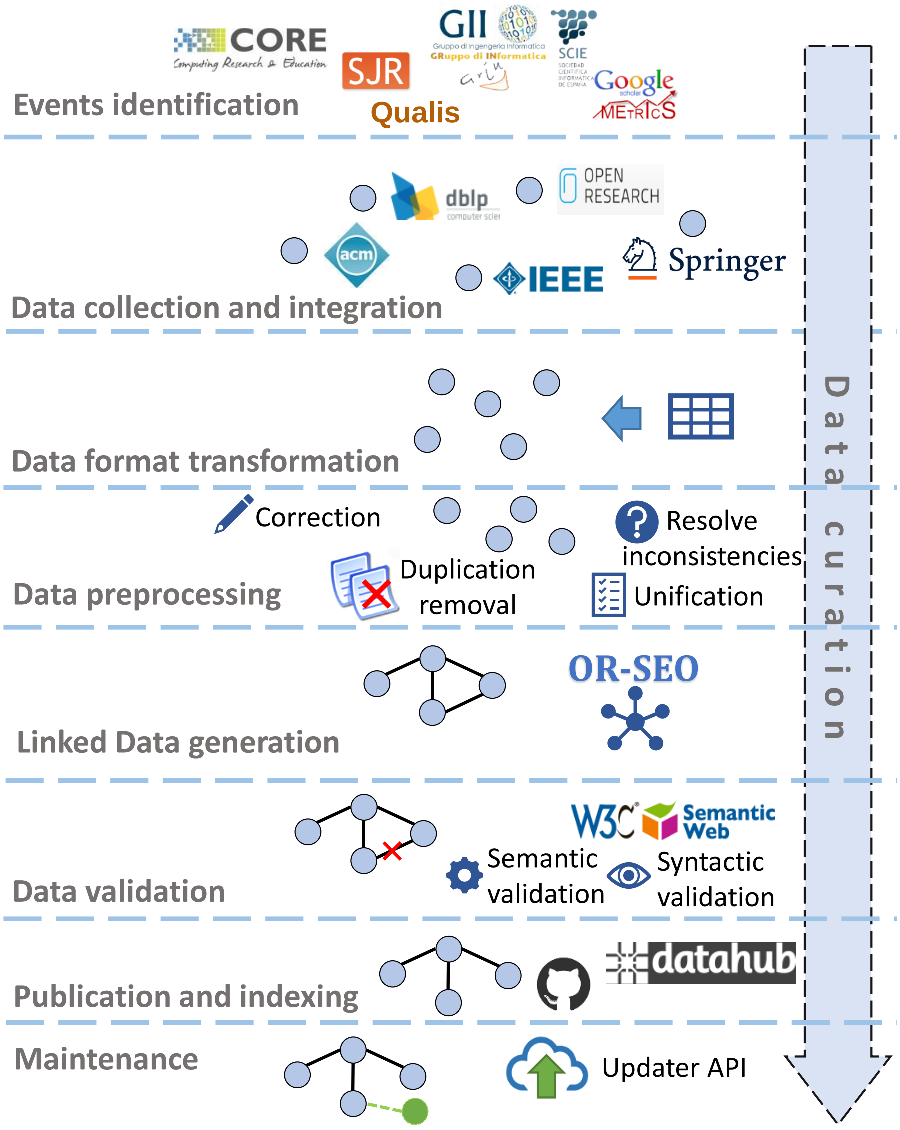


Figure 5.5: **EVENTSKG Curation.** The data curation process of EVENTSKG.

process, starting from the identification of top-ranked events until the maintenance phase, as shown in Figure 5.5. The curation of the EVENTSKG dataset is an incremental process starting from the identification of top-ranked events in each Computer Science community until the maintenance phase, which is performed continuously. In the following, 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 Computer Science community. To identify a subset of these events to be added to EVENTSKG, we used the metrics presented in subsection 3.1.1, which are used widely by Computer Science communities to identify top-ranked events in various Computer Science communities. While identifying top-ranked events in each community, we observed heterogeneity of the ranking of them in the aforementioned services, e.g., FOGA is ranked `A*` in CORE, i.e., ranked 1^{st}, while ranked `B3` in QUALIS, i.e., ranked 5^{th}, i.e., the 5^{th} highest category. 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 $SER\colon R \rightarrow S$, where $R = R_1, \ldots R_2$ is the set of existing rankings, and S is the set of SER classes. The range of $SER(x)$ is defined in Equation 5.3, where x is the sum of weights of each class in CORE, QUALIS, ERA, and GGS for each event series. We only chose the top-5 events, according to SER.

$$SER(x) = \begin{cases} A+ & \text{if } 100 < x \leq 75 \\ A & \text{if } 75 < x \leq 50 \\ B+ & \text{if } 50 < x \leq 25 \\ B & \text{if } 25 < x \leq 0 \end{cases} \tag{5.3}$$

Data Collection and Integration. Still, metadata collection is considered a time-consuming task because of the diversity of data sources available on the Web. 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.*

Linked Data Generation and Interlinking. The adoption of 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 [4]. The objective of this phase is to generate linked data from the less reusable, intermediate CSV representation. Using an ad-hoc transformation tool[34], we transformed the CSV data to an RDF graph, after mapping several events attributes given in the CSV file to the corresponding OR-SEO properties. Using a comprehensive ontology as the schema of the dataset gives the ability to obtain insights from the data by applying inference engines.

Data Validation. The next step is to semantically and syntactically validate the RDF graph to ensure the quality of the data produced. The syntactic validation has been carried out using the W3C RDF online validation service[35] to ensure conformance with the W3C RDF standards. The Hermit Reasoner is used to detect inconsistencies, i.e., Semantic Validation. Detecting inconsistencies is crucial because inconsistencies can result in a false semantic understanding of

[34] `http://levelup.networkedplanet.com/`

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

the knowledge. We periodically run the reasoner to ensure that no other inconsistencies arise after the interlinking process.

Data Publication. EVENTSKG is published in several RDF serializations, and it is registered in a GitHub repository (cf. Table 5.6). The commonly used way to make a dataset easier to find, share, and download is to index it in a public data portal, e.g., DataHub (cf. Table 5.6). A complete resource of the AAAI 2017 conference in the RDF/XML serialization can be found on the documentation page.

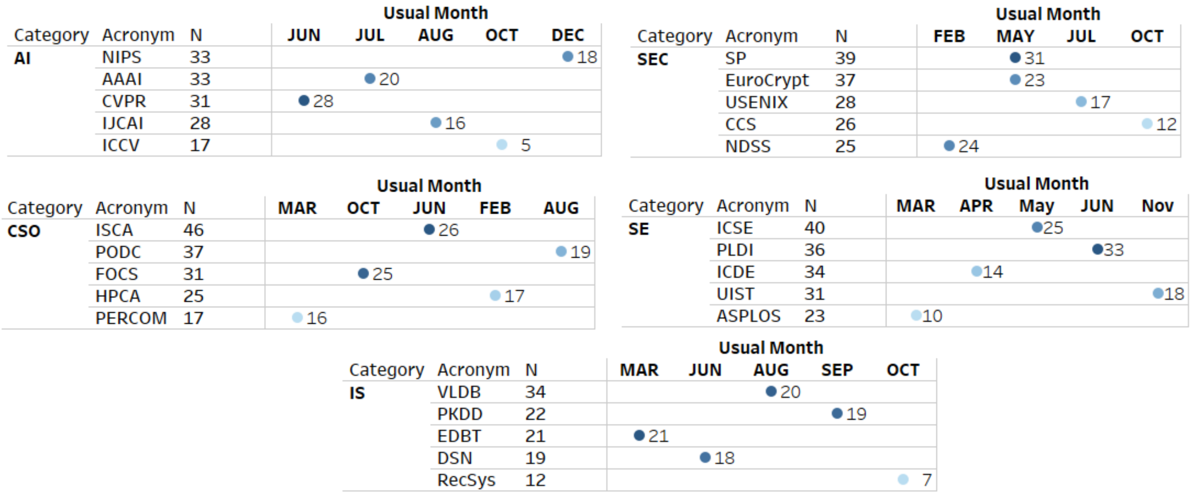
Maintenance. To maintain EVENTSKG and to keep it sustainable, there are several challenges to be considered; here is how we address them: 1) A *Java API* for updating and maintaining the dataset has been developed; source code is available on GitHub (cf. Table 5.6). A GUI has been developed on top of the API in order to facilitate 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. Furthermore, it also facilitates the addition of new events to the dataset. For instance, metadata for each individual event, e.g., TheWeb conference, can be easily updated or added using a friendly user interface, 2) *GitHub Issue tracker*: EVENTSKG has an issue tracker on GitHub, enabling the community to report bugs or to request features.

Dataset Content Analysis

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 Computer Science 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 aforementioned research questions.

Time distribution. 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 5.6 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.

Geographical distribution. 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 has 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. Regarding *Continent-level GD*, Europe hosted IS events the most, followed by SEC events (shown in Table 5.7). 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 among all communities. Regarding *Country-level GD*, it is observed that ICCV and ISMAR have $\bar{x} = 1$, which means that they



Category	Acronym	N	JUN	JUL	AUG	OCT	DEC
AI	NIPS	33					● 18
	AAAI	33		● 20			
	CVPR	31	● 28				
	IJCAI	28			● 16		
	ICCV	17				● 5	

Category	Acronym	N	FEB	MAY	JUL	OCT
SEC	SP	39		● 31		
	EuroCrypt	37		● 23		
	USENIX	28			● 17	
	CCS	26				● 12
	NDSS	25	● 24			

Category	Acronym	N	MAR	OCT	JUN	FEB	AUG
CSO	ISCA	46			● 26		
	PODC	37					● 19
	FOCS	31		● 25			
	HPCA	25				● 17	
	PERCOM	17	● 16				

Category	Acronym	N	MAR	APR	May	JUN	Nov
SE	ICSE	40			● 25		
	PLDI	36				● 33	
	ICDE	34		● 14			
	UIST	31					● 18
	ASPLOS	23	● 10				

Category	Acronym	N	MAR	JUN	AUG	SEP	OCT
IS	VLDB	34			● 20		
	PKDD	22				● 19	
	EDBT	21	● 21				
	DSN	19		● 18			
	RecSys	12					● 7

Figure 5.6: **Time distribution** of all events in terms of the most months when the event was held. N is the number of editions.

Table 5.7: **Normalized frequency of occurrence** of events by continent.

Comm.	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

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.

Publishers. It is observed that several event 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%.

Summary

This section 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 over the last 50 years. The OR-SEO ontology is used as the reference model for creating the dataset. We proposed the workflow of the curation process of EVENTSKG that can be used as a reference model for creating and publishing scholarly events datasets. In addition, we present a new scholarly event ranking service (SER), which combines the rankings of Computer Science 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 Computer Science communities has been published as a linked open dataset. EVENTSKG is coupled with an API for updating and maintaining it without going into the details of how this data is represented. The most striking findings from the analysis of EVENTSKG's data are highlighted in section 4.5. These findings highlight the usefulness of EVENTSKG for the event's organizers, researchers interested in data publishing, as well as librarians. Besides, sharing and reusing scholarly datasets became a new form of scholarly communication. Finally, we believe that EVENTSKG can close an important gap in analyzing the productivity and popularity of Computer Science communities, i.e., publications and submissions, and it is of primary interest to steering committees, proceedings publishers, and prospective authors.

CHAPTER 6

Facilitating Scholarly Data Retrieval

Recently, semantic data have become more distributed. Available datasets increasingly serve non-technical as well as technical audiences. This is also the case with our EVENTSKG dataset (presented in chapter 5), a comprehensive knowledge graph about scientific events, which serves the entire scientific and library community. A common way to query such data is via SPARQL queries. Non-technical users, however, have difficulties with writing SPARQL queries, because it is a time-consuming and error-prone task, and it requires some expert knowledge. This opens the way to natural language interfaces to tackle this problem by making semantic data more accessible to a broader audience, i.e., not restricted to Semantic Web experts.

In this chapter, we present SPARQL-AG, a front-end that automatically generates and executes SPARQL queries for querying EVENTSKG. SPARQL-AG helps potential semantic data consumers, including non-experts and experts, by generating SPARQL queries, ranging from simple to complex ones, using an interactive web interface. The ultimate goal behind this work is to widen the access to semantic data available on the Web by making it easier to generate and execute SPARQL queries with prior knowledge of neither the schema of the data being queried nor the SPARQL syntax. The prominent feature of SPARQL-AG is that users neither need to know the schema of the knowledge graph being queried nor to learn the SPARQL syntax, as SPARQL-AG offers them a familiar and intuitive interface for query generation and execution. Most SPARQL features are covered, such as optional, filters, aggregations, restricting aggregations, ordering, and limiting the number of results. It maintains separate clients to query three public SPARQL endpoints when asking for particular entities. The service is publicly available online[1] and has been extensively tested.

The following research question is investigated in this chapter:

> **RQ2**: How can we represent and integrate heterogeneous scholarly event metadata in knowledge graphs to facilitate scholarly data management and retrieval?

The work presented in this chapter is based on the following publication:

- **Said Fathalla**, Christoph Lange, and Sören Auer. *A Human-friendly Query Generation Front-end for a Scientific Events Knowledge Graph.* In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 200-214. Springer, Cham, 2019.

[1] `http://kddste.sda.tech/SER-Service/SPARQL-AG/SPARQL-AG.php`

6.1 Motivation and Research Problem

Nowadays, large amounts of semantic data have become widely available on the Web. This plethora of semantic data and the wide range of domains this data belongs to make it difficult to query this data. In addition, querying such data is a ponderous process, not only because of the syntax barrier but mainly because of data heterogeneity and diversity. Semantic data is queried by means of the widely-adopted W3C-standardized SPARQL query language [191]. SPARQL queries are executed against SPARQL endpoints, i.e., standardized query interfaces for semantic data stores. The advantages of SPARQL come from its expressiveness and scalability; however, people spend a large part of their time to learn how to write a SPARQL query to fulfill their needs and, in many cases, they fail. In this chapter, we present SPARQL-AG, a semantic web front-end that assists users in generating SPARQL queries for querying the EVENTSKG knowledge graph [103], a comprehensive knowledge graph for scientific events in computer science. The rationale to develop SPARQL-AG is to help potential semantic data consumers, including both SPARQL experts and non-experts, by automatically generating SPARQL queries, ranging from simple to complex ones, using an interactive web interface. It helps SPARQL experts by reducing the time required to write queries by modifying the generated query (*modify-before-execution*), i.e., removing the need to write the query from scratch. The generated query is displayed in an understandable way to make it easier to understand when a modification is needed before execution.

The architecture of SPARQL-AG is composed of five components: user interface, components selection, query composer, SPARQL clients manager, and query executor. This architecture integrates aspects of four research paradigms: query building (QB), semantic search (SS), human-computer interaction (HCI), and SPARQL query federation. Most of the SPARQL 1.1 Specifications [191] is covered, such as optional graph patterns, filters, aggregations, restricting aggregations, ordering, and limiting the number of results. SPARQL-AG maintains three SPARQL clients to query three public SPARQL endpoints (DBpedia SPARQL endpoint[2], the Scientific Events Ontology (SEO) SPARQL endpoint[3], and the EVENTSKG SPARQL endpoint[4]), asking for particular entities. Hence, there is no need to precisely know externally-defined entities; for instance, it is not required to know the DBpedia identifier for a country, some of which cannot be guessed trivially (e.g., `http://dbpedia.org/page/Georgia_(country)`). Querying external SPARQL endpoints is transparent to the user. A list of all currently existing countries is retrieved and cached by running a query against the DBpedia SPARQL endpoint. This list is periodically updated to obtain new updates if there are any. It is worth to mention that no special configurations for SPARQL endpoints are needed. Currently, SPARQL-AG is tailored to generate and execute queries over the EVENTSKG knowledge graph. This follows the motivation that the "*the more a system is tailored to a domain, the better its retrieval performance is*" [152]. However, the approach is easily transferable to other datasets and domain representations.

Our research aims at answering the following questions:

- How can users query semantic data without knowing the schema of this data?
- How can users query semantic data without learning RDF, OWL, or SPARQL?
- How can we combine data from several SPARQL endpoints to formulate a SPARQL query?

SPARQL-AG is a web-based user interface, which allows end-users to create and execute both

[2] `https://dbpedia.org/sparql`

[3] `http://kddste.sda.tech/SEOontology/sparql`

[4] `http://kddste.sda.tech/sparql`

simple and complex SPARQL queries over scholarly knowledge bases. Generally, we believe that SPARQL-AG bridges the gap between researchers outside the semantic web community, or even within the community but not being SPARQL experts, and the semantic data available on the Web. Several SPARQL experts have tested it by creating numerous successful queries. The source code is available on GitHub (cf. Table 6.1).

6.2 Design Principles

Several design principles considered when designing SPARQL-AG. Below, we outline them.

Research Paradigms. When designing SPARQL-AG, we integrate different research paradigms in the system architecture.

- *Query Building (QB)*: to build error-free SPARQL queries based on user selections from a visual interface. Query builders have the advantage of allowing for high expressivity while assisting users by listing eligible query elements without prior knowledge about the syntax of the language, which helps to avoid syntax errors,
- *Semantic Search (SS)*: for entities in various knowledge graphs that match the query pattern,
- *Human Computer Interaction (HCI)*: to make the user's interaction as simple and efficient as possible, in terms of accomplishing user goals, i.e., facilitating the task of querying semantic data without writing any piece of code, and
- *Federated Query*: SPARQL 1.1 Federated Query is a technique that is used for executing queries distributed over different SPARQL endpoints.

Portability. To promote portability, SPARQL-AG is a fully web-based service following web standards. A public SPARQL endpoint is used for querying knowledge graphs using HTTP requests, PHP and JavaScript for the application code, and HTML5/CSS3 (Bootstrap) for designing and styling the user interface (more details in section 6.5).

Availability. SPARQL-AG has been available online at `http://kddste.sda.tech/SER-Service/SPARQL-AG/SPARQL-AG.php` since December 2018. Users only need the URL of the service to be able to use it, i.e., no configuration or prerequisites are required.

Maintainability and Sustainability. SPARQL-AG is developed and maintained by the author of this thesis and hosted on the server mentioned above. To ensure the sustainability of SPARQL-AG, we use the issue tracker on its GitHub repository (cf. Table 6.1) in order to make it easier for users to request new features, e.g., features not covered in the initial release, and to report any problems/bugs.

6.3 Architecture

The architecture of SPARQL-AG is composed of five components (see Figure 6.1): user interface, components selection, query composer, SPARQL clients manager, and query executor. Later in this section, we denote the set of SPARQL variables with V, the set of predicates used in the

Table 6.1: **SPARQL-AG Resources.** The URLs of the resources related to SPARQL-AG.

Resource	URL
EVENTSKG	`http://kddste.sda.tech/EVENTSKG-Dataset/`
SEO Ontology	`https://w3id.org/seo#`
SPARQL-AG URL	`http://kddste.sda.tech/SER-Service/SPARQL-AG/SPARQL-AG.php`
GitHub repository	`https://github.com/saidfathalla/SPARQL-AG`
Issue Tracker	`https://github.com/saidfathalla/SPARQL-AG/issues`

query pattern with P, the set of RDF resources with R, the set of query restrictions with QR, and the set of RDF literals with L.

User Interface. The role of the user interface component is to 1) provide end-users a Graphical User Interface (GUI) through which they can select different query components using a web form, 2) display the generated query (GQ) to the user, who can then modify it before it is submitted to the SPARQL endpoint for execution, and 3) send the queries to the EVENTSKG SPARQL endpoint and display the results in a human-readable format (HTML table). SPARQL-AG features are: 1) generating and executing simple SPARQL queries, 2) generating SPARQL queries with aggregation, and 3) executing predefined query templates. Users can 1) select columns they want to appear in the result using checkboxes, 2) restrict the results by selecting the checkbox corresponding to each predicate and by entering/selecting possible object values for these predicates by – depending on the datatype – direct input of numeric values, selecting from a list, or picking a date from a calendar. This avoids the problem of resolving the ambiguity that might arise when processing natural language queries.

Components Selections. User selections of the different SPARQL query components, listed below, are passed to the Query Composer (QC) in order to formulate the SPARQL query based on these selections.

- *Prefix declarations*: the set of namespace prefixes used in the query,
- *Dataset definition*: we omit the *dataset definition* part of the SPARQL query from user selections because it is implicitly given, i.e., the EVENTSKG URI,
- *Result clause*: identifying what information to return from the query,
- *Query patterns*: specify the query's graph pattern that matches the data, such as `UNION`, `MINUS`, `FILTER`, and `OPTIONAL`, and
- *Query modifiers*: a set of modifiers for the query results, such as order, projection, distinct, offset, and limit.

Each selection is mapped to a query element in the generated query. To address scalability issues, users can limit the number of results retrieved using the `LIMIT` modifier.

Query Composer and SPARQL Clients Manager. The query composer is the core component of SPARQL-AG, as it formulates the SPARQL query based on user selections and data received from the SPARQL Clients Manager, which forms the link between SPARQL endpoints and the query composer. SPARQL Clients Manager is responsible for managing SPARQL clients

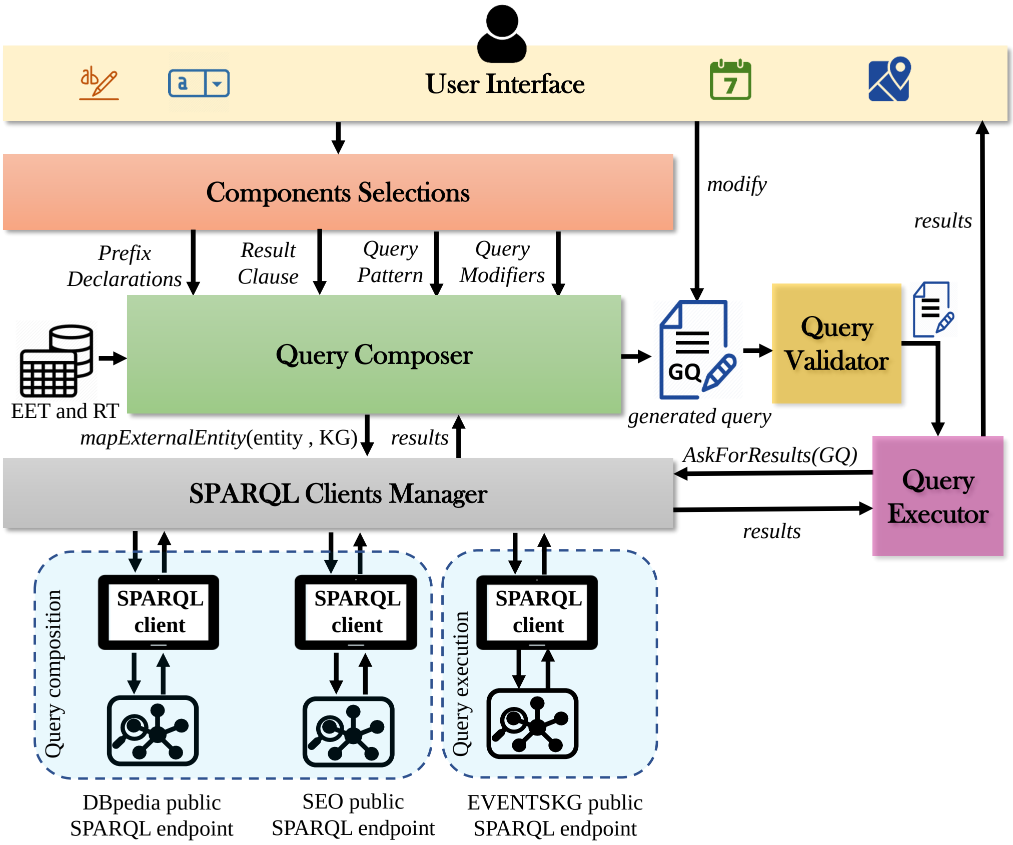
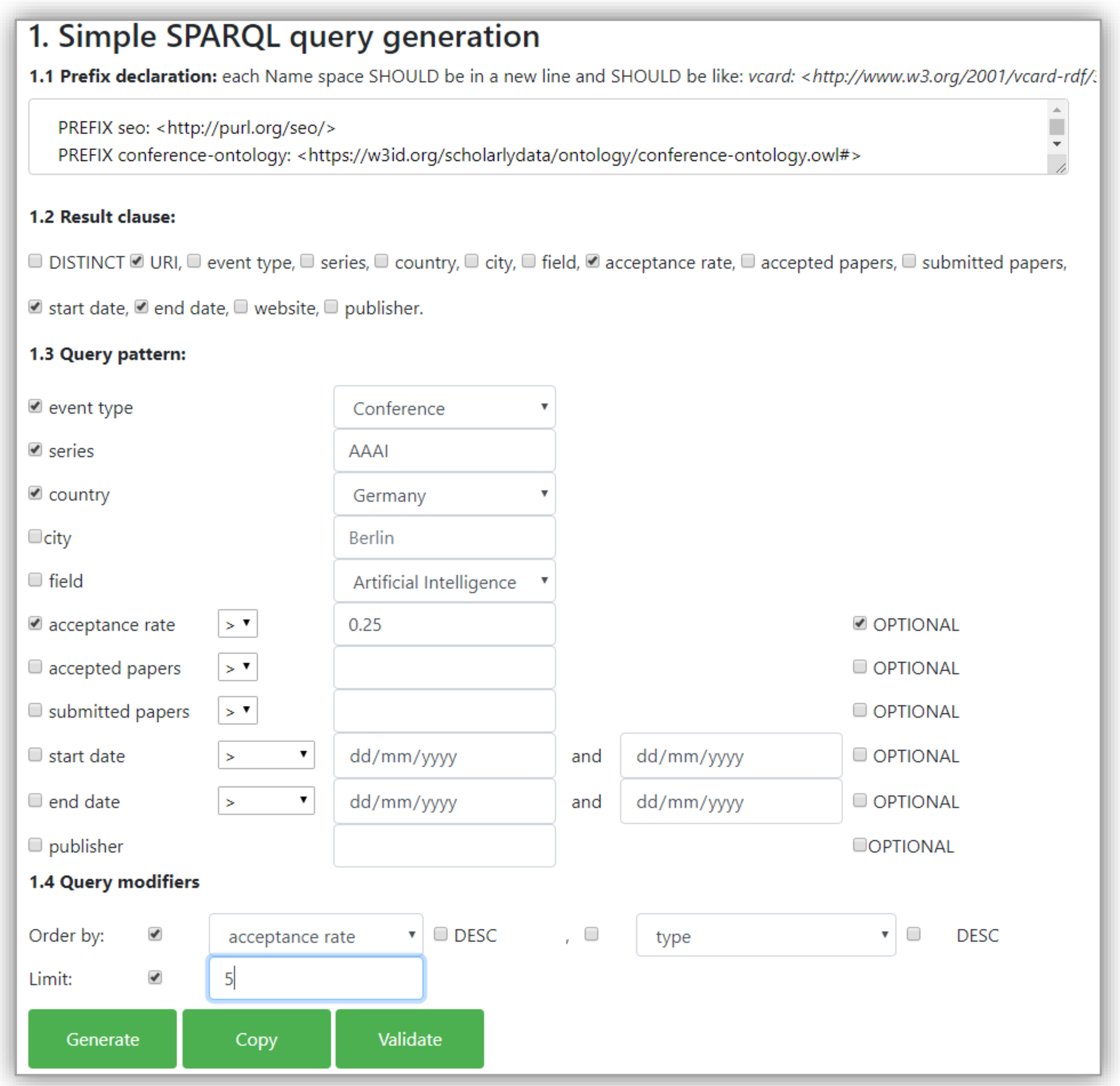


Figure 6.1: **System Architecture.** The complete system architecture of SPARQL-AG.

in order to be able to query external SPARQL endpoints, DBpedia, and OR-SEO in this case. SPARQL clients allow executing SPARQL queries against remote SPARQL endpoints using the SPARQL protocol [192]. After a successful connection to the endpoint, the SPARQL client sends the SPARQL query to the endpoint and waits for the result. When requests are received from the Query Composer, asking for the resource URI of an externally defined entity, e.g., Germany, the Clients Manager formulates a SPARQL query and sends it to the SPARQL client responsible for this type of request and waits for the requested resource URI, e.g., `http://dbpedia.org/resource/Germany`.

The main steps carried out by the Query Composer are summarized in Algorithmus 1, where "+" denotes string concatenation.

In the beginning, users should define the namespaces used in the query in the prefix set (PS). Currently, as SPARQL-AG is provided for the EVENTSKG dataset, the prefix declaration is automatically generated with the required predicates, i.e., `seo` and `conference-ontology`. For future purposes, if a namespace mapping is not available in the system, then the user should add this namespace to the prefix declaration. Each selection in the result clause (RC) (upper part in Figure 6.2) is mapped to a new variable in the SPARQL query (call `mapResultClause` in Algorithmus 1). For instance, when *country* is selected, it represented by the variable ?*country*. Formally, the mapping is defined as follows: Each RC selection is assigned a unique variable via the function $mapResultClause : RC \mapsto V$. Each query pattern is represented as a tuple of $(prop, op, val)$, where *prop* is the property being restricted, *op* is the operation,



1. Simple SPARQL query generation

1.1 Prefix declaration: each Name space SHOULD be in a new line and SHOULD be like: *vcard: <http://www.w3.org/2001/vcard-rdf/:*

```
PREFIX seo: <http://purl.org/seo/>
PREFIX conference-ontology: <https://w3id.org/scholarlydata/ontology/conference-ontology.owl#>
```

1.2 Result clause:

□ DISTINCT ☑ URI, □ event type, □ series, □ country, □ city, □ field, ☑ acceptance rate, □ accepted papers, □ submitted papers, ☑ start date, ☑ end date, □ website, □ publisher.

1.3 Query pattern:

☑ event type		Conference ▾			
☑ series		AAAI			
☑ country		Germany ▾			
□city		Berlin			
□ field		Artificial Intelligence ▾			
☑ acceptance rate	> ▾	0.25			☑ OPTIONAL
□ accepted papers	> ▾				□ OPTIONAL
□ submitted papers	> ▾				□ OPTIONAL
□ start date	> ▾	dd/mm/yyyy	and	dd/mm/yyyy	□ OPTIONAL
□ end date	> ▾	dd/mm/yyyy	and	dd/mm/yyyy	□ OPTIONAL
□ publisher					□OPTIONAL

1.4 Query modifiers

Order by: ☑ acceptance rate ▾ □ DESC , □ type ▾ □ DESC

Limit: ☑ 5

Figure 6.2: **SPARQL-AG User interface.** A part of the User interface of SPARQL-AG.

and *val* $\in L$ is the value. For example, $(\mathit{acceptanceRate}, \geq, 0.25)$ represents the results with an acceptance rate greater than or equal to 0.25. A query pattern might contain externally defined entities, such as countries and cities (defined in DBpedia), and the research field which the events belong to (defined in the SEO ontology). Here, the *External Entities Table* (EET) plays its role, in which all external entities are stored along with the URL of the public SPARQL endpoint of the knowledge graph in which these entities can be found. For instance, countries and cities can be found through DBpedia endpoint. Therefore, these entities should be identified (call `isExternallyDefined`). The function $\mathit{isExternallyDefined} : P \mapsto \mathbb{B}$ is defined as: $\mathit{isExternallyDefined}(p) := \mathit{true}$ if p is found in EET, and *false* otherwise. The URIs of these external entities should be retrieved as well by sending requests to the SPARQL Clients Manager via the function: $\mathit{mapExternalEntity} : P \mapsto R$. Each request is assigned a unique number via the function *reqID* : $RR \mapsto \mathbb{N}$, where RR is the set of requests made by the Query Composer component.

After successful retrieval of the requested data, all these requests are stored in the requests table (RT) for answering further requests, instead of sending them to the Clients Manager, i.e., caching requests. Therefore, after a period of time, when all external entities within a specific query have been requested before, then there is no need to query external endpoints for this query. This reduces the workload of querying external knowledge graphs for every request. In addition, it performs results aggregations when more than one client, i.e., SPARQL endpoints, returns a result.

In order to map query restriction (QR) to the corresponding predicate in the dataset, the function *mapToPredicate* : $QR \mapsto P$ is used. For instance, when the user wants to filter results by, e.g., country, then he/she should select the country checkbox (Figure 6.2), which should be mapped to the corresponding predicate in the dataset, i.e., `seo:heldInCountry`. Each variable in the result clause must be bound in the query pattern, therefore the function *mapToVariable* : $P \mapsto V$ is used to obtain these variables, which are defined in the result clause to be bound in the query pattern. For example, when users want to display events along with their start and end dates, these attributes are bound to two variables in the result clause, which are `?SD` and `?ED` respectively. Mapping query modifiers is straightforward, e.g., the *order by* modifier specifies columns (currently limited to one or two) to order the results by using the `ORDER BY` keyword.

Query Validator. When users modified the generated query before execution, they can validate the modified query using the "*Validate*" button. Validation is performed by the SPARQL parser associated with the SPARQL engine provided by RAP (RDF API for PHP), more details in section 6.5.

Query Executor. This component is responsible for sending the validated query to the Clients Manager (`AskForResults`) and displaying the results in a human-readable format. Since the query results are returned as an array of variable bindings, which are difficult to understand for end-users who are not familiar with SPARQL, we decided to display the results as an HTML table.

Generating SPARQL with aggregation functions. Aggregation functions are useful when users want to study data in an analytical fashion, e.g., finding the total number of publications of all events in each research field. Results are grouped using the `GROUP BY` clause, and these groups can be filtered using the `HAVING` clause. Here, it is worth to mention that SPARQL-AG enforces some SPARQL rules to be applied while users select different query elements. For instance, a group column is added to the `GROUP BY` clause when aggregation functions are used, and aggregation functions are automatically restricted using `HAVING`. For example, when the user uses a column in an aggregation function, this column must be added to the `GROUP BY` clause; this is a SPARQL restriction.

6.4 Possible Use Cases

In this section, we present use cases for SPARQL-AG for supporting scholarly communication stakeholders by providing figures about computer science events in the context of eight computer

```
Algorithmus 1 : QueryGeneration
  Input   : PS: prefix set (array),
            RC: result clause set (array),
            QP: query pattern (array).
  Output  : query: SPARQL query (string)
  query = null ;
  foreach namespace $n_i \in PS$ do
  |  query= query + "PREFIX "+ $n_i$ ;
  end
  query = query +"SELECT ";
  while result clause $rc \in RC \neq$ null do
  |  query+="?"+mapResultClause(rc) // add result clause elements
  end
  query= query + "WHERE { ";
  foreach pattern $p_i \in QP$ do
  |  if uri=isExternallyDefined($p_i$) then
  |  |  value=mapExternalEntity($p_i$)
  |  else
  |  |  value=UI.control.value // to get literal values, e.g. numeric values
  |  |     from the UI
  |  |  ;
  |  end
  |  query+= "?e " + mapToPredicate($p_i$.prop) + mapToVariable($p_i$.prop) + " FILTER
  |   ( mapToVariable($p_i$.prop) + $p_i$.op + $p_i$.val )";
  end
  query+="}"
```

science communities. Listing 6.1 shows the SPARQL query generated for the query *"(Q1) List the top-10 events with topics related to* Artificial Intelligence *with an acceptance rate lower or equal to 0.20, which have been held in Germany in the last decade. Order the results by ascending acceptance rate"*.

Listing 6.2 shows the SPARQL query generated for a query with an aggregation: *"(Q2) List the subfields of computer science whose events have a large number of submissions, i.e., greater than 10,000. Order the results by ascending field name"*.

6.5 Implementation and Evaluation

This section describes the implementation, the evaluation, and discusses the results of a usability study for testing SPARQL-AG.

Implementation. SPARQL-AG is implemented in PHP as a web-based service, using a client-server architecture. We have implemented all functions described in section 6.3 using PHP 7.2.10 and the RAP (RDF API for PHP) toolkit[5]. In addition, JavaScript is used to validate input data

[5] http://wifo5-03.informatik.uni-mannheim.de/bizer/rdfapi/index.html

and to enforce some rules, such as GROUP BY rules, as mentioned in section 6.3. SPARQL-AG only needs the URL of an endpoint to explore it, without any further required configuration. Thus the approach is easily transferable to other datasets by just changing the components selections and the dataset URL. Queries to SPARQL endpoints are sent directly from the client browser, using HTTP requests, which makes SPARQL-AG independent from a server.

```
PREFIX seo: <http://purl.org/seo/>
PREFIX conference-ontology:
<https://w3id.org/scholarlydata/ontology/conference-ontology.owl#>
SELECT DISTINCT ?event ?acceptance ?field
WHERE {
  ?event  rdf:type  ?type .
  FILTER (?type = conference-ontology:Conference ||
  ?type = conference-ontology:Workshop  ||
  ?type = seo:Symposium) .
  ?event  seo:heldInCountry  ?country .
  FILTER (?country = <http://dbpedia.org/resource/Germany>) .
  ?event  seo:field  ?field .
  FILTER (?field = <http://purl.org/seo#ArtificialIntelligence>) .
  ?event  seo:acceptanceRate  ?acceptance .
  FILTER (?acceptance < 0.20) .
}
ORDER BY ?acceptance
LIMIT 10
```

Listing 6.1: **SPARQL query** generated by SPARQL-AG for Q1.

```
SELECT ?field SUM(?submissions) AS ?SP_SUM
WHERE {
  ?event  seo:submittedPapers  ?submissions .
  ?event  seo:field            ?field .
}
GROUP BY ?field
HAVING (SUM(?submissions) > 10000)
ORDER BY ?field
LIMIT 10
```

Listing 6.2: **SPARQL query** generated by SPARQL-AG for Q2.

Experimental Setup. To ensure the usefulness of SPARQL-AG, we confirm that it can answer several competency queries listed in Table 7.9. These queries were thoroughly selected to assure that they cover all features provided by SPARQL-AG. Besides, we test the usability of SPARQL-AG by letting real users use the system, i.e., Usability testing [193]. Nielsen and Landauer [193] argue that the best usability evaluation results come from testing no more than five users and running as many small tests as possible. This type of evaluation was performed by several Semantic Web query interfaces [152, 154] since it gives direct insight into how real users use the system. The goal is to improve the usability of the system being tested. In this experiment, casual end-users should test and assess the usability of the service. To apply the usability test for SPARQL-AG with real-world end-users, we promoted the usability study on its web site, several mailing lists, and between colleagues. A total of 12 participants were recruited for this study. They are distributed over a wide range of backgrounds and professions. In addition, we ensure the anonymity of the participants in order to obtain unbiased results.

The participants were split into two groups (six users each) based on their SPARQL experience: 1) experienced SPARQL users (all are computer scientists), and 2) casual end-users from other fields and professions, such as dentistry and engineering. In the beginning, we informed all participants that the query interfaces were being tested and not the users themselves. This is an

Table 6.2: **Queries used in evaluating SPARQL-AG.** Each variable, such as X and Y, is a placeholder for an appropriate replacement.

No.	Query
Q1	List events related to field X that took place in country X.
Q2	List events related to field X with an acceptance rate less than Y along with their sponsors, publishers, website, and start date.
Q3	List events that took place in a particular month X, along with their publishers and the field of research.
Q4	List conference series that have been held in country X in a particular month Y.
Q5	List the number of submitted and accepted papers of a series X in a particular time period.
Q6	List the top-X countries that hosted the most events in CS overall.
Q7	List the subfields of CS for which a country X has hosted most events since a particular date Y.
Q8	Compare the popularity of different computer science research communities, in terms of the number of submissions to the respective events.
Q9	List the top-X research fields, in terms of the number of events they have.
Q10	Find the average acceptance rate for events in each computer science research community.

important issue that can severely influence the test results. Inexperienced users are confused when given too many interaction options. Therefore, at the beginning of each experimental run, we gave each participant all information and instructions concerning the experiment, either in a face-to-face meeting or in a call (for remote participants). Most of the users found that the experiment can be easily understood by casual end-users and does not require expert knowledge. After testing the service, experienced users were explicitly asked to fill in a satisfaction questionnaire (accessible through the link `https://goo.gl/55TbRU`) in which they were asked about their assessment of the interface, generated query, the presentation of the results, and the usefulness of SPARQL-AG for both casual and experienced users. In addition, casual users were explicitly asked to fill in the System Usability Scale (SUS) questionnaire [194] (available at `https://goo.gl/Mxj9Uu`). SUS is a standardized usability test, which contains ten questions with five possible responses ranging from 1 (strongly disagree) to 5 (strongly agree). The best way to interpret one's results is by normalizing the scores to produce a percentile ranking. We also asked experienced SPARQL users for further qualitative feedback, including positive and negative aspects, and suggestions for future improvements.

Results. Five (out of six) experts strongly agreed that SPARQL-AG is helpful for users with no prior knowledge of SPARQL, and they are satisfied with the design of the user interface. For the analysis of results, the SUS scoring method [195] has been used for the casual end-user questionnaire. The average SUS score falls into seven adjective ratings, ranging from *Best Imaginable* (above 90.9) to *Worst Imaginable* (below 12.5) [196]. Most strikingly, findings showed that SPARQL-AG scored a high SUS satisfaction score of 93, i.e., Best Imaginable, thus reaching excellent usability. Table 6.3 contains the statistics for each question of the SUS Questionnaires for casual end-users. Since all questions measure positive agreement, notably, the mode of almost every response is 5, which means that most participants responded with *strong agree* for most

Table 6.3: **SUS Questionnaires Results.** Statistics for questions of the SUS Questionnaires for casual end-users.

Metrics	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Mean	4.57	4.14	4.71	4.57	4.57	4.00	4.86	4.43	4.29	3.86
SD	0.53	1.46	0.49	0.53	0.53	1.00	0.38	1.13	0.76	0.90
Median	5.00	5.00	5.00	5.00	5.00	4.00	5.00	5.00	4.00	4.00
Mod	5.00	5.00	5.00	5.00	5.00	4.00	5.00	5.00	5.00	3.00

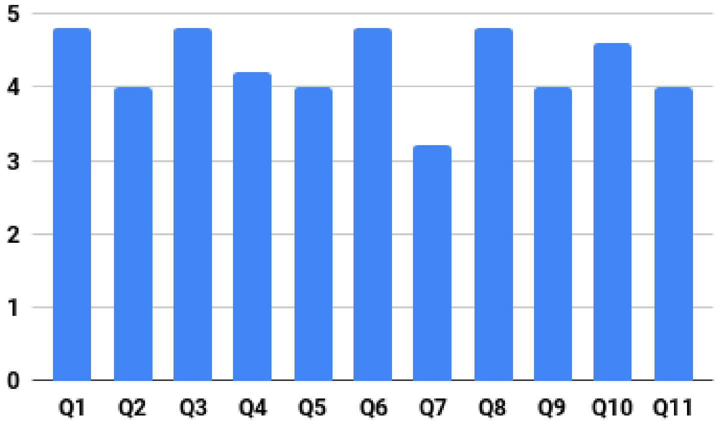


Figure 6.3: **Satisfaction Questionnaire Results.** The mean of experts response in the satisfaction questionnaire.

questions. As shown in Figure 6.3, the mean of expert responses to the expert questionnaire falls into the range of 3.2–4.8, which is rather high. This implies a reasonable satisfaction of all experts for all questions since all of them have been designed to measure positive satisfaction. Q7 (“Do you agree that the translation of SPARQL queries to natural language is useful?”) has the lowest score of 3.2, which means that there is no need to translate SPARQL queries to natural language to let casual users confirm their actual intention.

6.6 Summary

This chapter presents SPARQL-AG, which aims at improving data access to semantic knowledge graphs by generating SPARQL queries for users who may be challenged by a lack of schema or data knowledge regarding a knowledge graph. The architecture of SPARQL-AG can be used for generating SPARQL queries for any semantic data. It combines research techniques from different disciplines in an integrated fashion. We highlight the importance of using NLIs to make semantic data accessible to a broader community. One aim of this study is to show the potentials of NLIs that give a chance to casual users to benefit from the Semantic Web’s capabilities without having to study them. SPARQL-AG also significantly lowers the barrier of writing SPARQL queries from scratch by providing additional support for SPARQL experts. It generates SPARQL queries of three kinds: simple query generation, SPARQL query generation with aggregation, and parameterized execution of predefined queries. We believe that we are on the way towards increasing users' understanding of SPARQL by lowering the syntax barrier, since generating queries using user interactions will increase the understanding of the syntax itself, enabling users to improve their understanding of the query language incrementally. Our usability study with twelve participants using a list of 21 different questions showed that the usability of the system is excellent, with a SUS score of 93. The results of our evaluation show that both experienced and casual users agree that writing SPARQL queries in a blank sheet, where they must type commands, is cumbersome and time-consuming.

As anticipated, SPARQL-AG enables the successful generation of error-free and readable queries, which potentially saves much time and effort. To the best of our knowledge, this is the first web-based user interface that allows end-users to create and execute both simple and complex SPARQL queries over scholarly knowledge bases. Nevertheless, it can be applied to any domain with little adaptation effort. Our work has some limitations. Still, not all SPARQL 1.1 Specifications are covered. Also, currently, it is restricted to only one dataset. Nevertheless, our work provides a framework for developing query builders for querying scholarly data, making such data available for further analysis and improvement.

CHAPTER 7

Towards a Knowledge Graph for Science

Despite significant advances in technology, the way how research is performed has not changed much.

Over the past four years, we have been engaged in the development of the Science Knowledge Graph Ontologies (SKGO), which is a set of ontologies for modeling the scientific knowledge in various fields of science resulting in a knowledge graph of the scientific findings in modern sciences, such as natural, social, and formal sciences. We demonstrate the utility of the resulting knowledge graph by using it to answer queries about the different research contributions presented in scientific papers. We evaluate how well the query answers serve readers' information needs, in comparison to having them extract the same information from reading such papers.

In this chapter, a knowledge graph-based approach as well as a set of OWL ontologies are proposed for providing a semantic representation for 1) representing various branches in modern science with the aim of representing relationships among them; 2) modeling research findings in Computer Science; and 3) modeling research data in various natural sciences, involving Physics and Pharmaceutical Science. We affirm the applicability of developing a knowledge graph-based approach for exploring scholarly Knowledge Graphs using a semantic wiki platform. Section 7.1 provides an overview of the limitations of the traditional representation, i.e., unstructured format, of the scientific knowledge published on the Web as well as the long-term vision of the semantic representation of such knowledge. In section 7.2, we introduce this suite of ontologies and discuss the design considerations taken into account during the development process. Section 7.3 presents a semi-automatic knowledge graph-based approach (Aurora) that captures information about research contributions in the *OpenResearch.org* semantic wiki.

The following research question is investigated in this chapter:

> **RQ4**: How can published research in various fields of science be understood by machines making information retrieval, analysis, and scholarly data management more efficient?

The work presented in this chapter is based on the following publications:

- **Said Fathalla**, Sahar Vahdati, Sören Auer, and Christoph Lange. *Towards a Knowledge Graph Representing Research Findings by Semantifying Survey Articles.* In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 315-327. Springer, Cham, 2017.
- **Said Fathalla**, Sahar Vahdati, Sören Auer, and Christoph Lange. *SemSur: A Core Ontology for The Semantic Representation of Research Findings.* Procedia Computer

Science 137 (SEMANTiCS), pp. 151-162, 2018.

- Sahar Vahdati, **Said Fathalla**, Sören Auer, Christoph Lange, and Maria-Esther Vidal. *Semantic Representation of Scientific Publications.* In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 375-379. Springer, Cham, 2019.
- **Said Fathalla**, Sören Auer, and Christoph Lange. *Towards The Semantic Formalization of Science: The Science Knowledge Graph Ontologies Suite.* In Proceedings of the 35th ACM/SIGAPP Symposium on Applied Computing (SAC), pp. 2057-2059, 2020.
- Zeynep Say, **Said Fathalla**, Sahar Vahdati and Sören Auer. *Ontology Design for Pharmaceutical Research Outcomes.* In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 119-132. Springer, 2020.
- Aysegul Say, **Said Fathalla**, Sahar Vahdati, Jens Lehmann and Sören Auer. *Semantic Representation of Physics Research Data.* In 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, In press, SciTePress, 2020.

7.1 Semantic Representation of Scientific Knowledge

Digitization has significantly eased the preparation, submission, and publication of scientific work to various venues around the world, resulting in a plethora of scientific literature. This plethora of scientific literature makes it laborious to obtain an overview of the current state of research results in different disciplines of science. From one day to the next, researchers produce a considerable number of scholarly articles, mostly in PDF format, that needs to be explored, found, interpreted, and understood by the community. The vision of representing knowledge for effective interaction with scholarly artifacts dates back to the era of emerging electronic information processing [197]. However, the sheer amount of information being published in this way poses challenges for scholars. To stay up to date and keep an overview of the state-of-the-art in a particular research area, researchers tend to write a distinct type of scientific publications, called survey/review articles. However, exploring, analyzing, and comparing such articles require significant effort and time.

7.1.1 Limitations of Traditional Representation

The unstructured or semi-structured representation of research data published on the Web still has deficiencies. The content is not represented in a formal, i.e., machine-comprehensible format, which restricts knowledge discovery and the development of intelligent agents, as well as the browsing applications on top. In addition, current scholarly knowledge infrastructures do not take advantage of modern information systems and emerging technologies to their full potential [198]. Modern information systems can boost knowledge-discovery applications, such as scientific exploration, by introducing highly structured knowledge representation [199]. Unfortunately, most existing scholarly communication infrastructures adopt traditional, keyword-based information retrieval strategies. Subsequently, knowledge-based representation of scholarly data, which motivates the development of data models, ontologies, and knowledge graphs, can support a richer representation of this data [14]. This representation will make it easier to query and process this data. Thus, there is a pressing need to represent this knowledge in a richer format.

7.1.2 Long-term Vision

Establishing knowledge-based scholarly communication makes analyzing and exploring scientific data in digital libraries both easier and efficient than it is now. We have the vision that ultimately, researchers will work on a standard knowledge base comprising comprehensive descriptions of their research, thus making research contributions transparent and comparable. This knowledge base will help to increase the impact that researchers can make by enabling them to immediately contribute to a common knowledge base comprising comprehensive descriptions of their research. Currently, the way of structuring, systematizing, and comparing research results is via writing survey or review articles. Such articles usually select a number of articles describing comparable research and (a) develop a standard organization scheme with feature classifications, (b) provide a conceptualization of the research domain with mappings to the terminologies used in the individual articles, (c) compare and possibly benchmark the research approaches, implementations, and evaluations described in the articles, and (d) identify directions for future research. As a result, survey and review articles significantly contribute to structuring a research domain and make its progress more transparent and accessible. However, such articles still share the same deficiencies as their original research counterparts, i.e., the content is not represented according to formal knowledge representation and not machine comprehensible.

In 2017, we took an initial step towards establishing Knowledge-based scholarly communication by developing the comprehensive Semantic Survey Ontology (SemSur) for capturing the scientific data within survey articles in Computer Science (more details in subsection 7.2.3). SemSur defines how surveys for research fields can be represented in a semantic format, resulting in a knowledge graph that represents individual research problems, approaches, implementations, and evaluations in a structured and comparable way [6]. As a result of structuring and representing research advances according to such a semantic scheme, scientific knowledge will become more comparable and accessible. For example, research addressing a particular problem can be automatically retrieved, as well as approaches can be compared according to their features or evaluation results in a specific setting. The ultimate aim of semantically representing research data is to enable the provision of better and more intelligent services for the discovery of scientific work.

7.2 Science Knowledge Graph Ontologies Suite

In this section, we introduce the Science Knowledge Graph Ontologies (SKGO) suite. This suite captures the knowledge of scientific information typically presented in publications by interlinking domain-specific concepts in a highly structured format, thus enabling access to these data in a machine-readable, transparent, and comparable manner. Currently, SKGO comprises four OWL ontologies for scientific work in various fields of science, including Computer Science (SemSur), Physics (PhySci), and Pharmaceutical science (PharmSci) as well as an upper ontology on top of them called Modern Science Ontology (ModSci). This suite can support the digital transformation of scholarly communication from documents to knowledge-oriented representation in the form of structured and interlinked knowledge graphs, aiming at analyzing, exchanging, and exploiting scholarly knowledge in an efficient manner. SKGO ontologies have been made publicly available in standard formats, at permanent URLs, following ontology publication best practices [174]. Technically, to facilitate the ontology reuse, PharmSci ontology is available in multiple RDF serializations, including Turtle, JSON-LD, RDF/XML, and N-Triples, from the SKGO GitHub repository. The development methodology, design considerations, and an overview about the ontologies are provided in the next subsections. In the this chapter, Graffoo diagrams [200] is used for visualizing the main ontological entities.

7.2.1 Ontologies Development

SKGO ontologies are developed through cross-disciplinary interaction between ontology experts and researchers belonging to the respective fields of science. The Systematic Approach for Building Ontologies (SABiO) [201] has been followed in the development process of SKGO ontologies. It comprises the following five phases:

1. *Ontology Purpose Identification and Requirements Elicitation*: this phase involves identifying the ontology's purpose and its intended uses, then eliciting its requirements. Ontology requirements refer to the knowledge (content) to be represented by the ontology. They can be stated as competency questions, which help to determine what is relevant to the ontology and what is not, i.e., defining the scope.

2. *Ontology Capture and Formalization*: this phase involves capturing the domain conceptualization, i.e., conceptual data modeling, based on the competency questions created in the first phase in order to produce a reference ontology,

3. *Ontology Design*: this phase involves transforming the conceptual specification of the reference ontology into a design specification,

4. *Ontology Implementation*: this phase involves implementing the ontology in a particular language, e.g., OWL, and

5. *Ontology Testing*: this phase involves verifying and validating the behavior of the ontology on a finite set of test cases against the expected behavior regarding the competency questions, hence the competency questions-driven testing.

The following design considerations have been taken into account in the development process of SKGO ontologies.

- *Publishing Ontologies on the Web*: All SKGO ontologies are published with dereferenceable URIs. We configure the server to provide human-readable HTML content from the vocabulary URI using the recipes provided in [174]. These recipes have been designed to be consistent with the architectural principles of the World Wide Web [202], thus enabling, e.g., the server to perform content negotiation, to serve humans as well as machine clients. Besides, the SKGO Ontologies follow the FAIR principles for data publication [17].
- *Findability:* All SKGO ontologies have been published under a persistent URL (`https://w3id.org/skgo/[ontologyprefix]#`) under the open CC-BY 4.0 license and they are available from a *GitHub* repository[1] (cf. Table 7.1). All SKGO ontologies are available in diverse RDF serializations, including Turtle, JSON-LD, RDF/XML, and N-Triples, from the SKGO GitHub repository. The SKGO ontologies can be browsed online, through web-based repository front-end for browsing and visualizing published ontologies, at: BioPortal[2],

[1] https://github.com/saidfathalla/Science-knowledge-graph-ontologies

[2] http://bioportal.bioontology.org/ontologies/MODSCI

Linked Open Vocabularies[3], and AberOWL[4]. These activities are a step towards increasing the findability of the ontology.

- *Logical correctness:* The dereferenceability of the URIs of the ontologies over the HTTP protocol has been validated using Vapour[5]. We validated the ontologies against inconsistencies using the HermiT reasoner, and the OOPS! Ontology Pitfall Scanner[6]. Inconsistent ontologies can arise for several reasons, including instantiating disjoint classes with the same instance and instantiating an unsatisfiable class.
  Documentation: The Widoco wizard for documenting ontologies [203] is used to create HTML documentation for ModSci, thus enabling human understanding of the ontology and increasing its reusability. The documentation is available online through the persistent URI of the ontology. The `rdfs:comment` property is used to provide a human-readable description of each resource. Besides the HTML documentation, we created an interactive visualization (available through the documentation page via visualization using D3.JS) that allows users to navigate the ontology hierarchy seamlessly in order to paint the whole picture of the branches of modern science.
  Adoption and Sustainability: The first author of this paper maintains ModSci. Furthermore, we are currently integrating ModSci to the Open Research Knowledge Graph (currently funded by an ERC project running until 2024[7]), thus enlarging the circle of maintainers of the ontology to further ORKG contributors. Suggested improvements, e.g., reusing related ontologies that may appear in the future, or reports of problems by the community can be submitted through the GitHub issue tracker via two types of issues: *Problem report* and *Improvement request*.
- *Metadata completion:* A checklist for completing the vocabulary metadata proposed in [176] has been used to complete the ontology's metadata, e.g., authorship information in terms of Dublin Core. This metadata makes it easier for academia and industry to identify and reuse ontologies effectively and efficiently [204].
- *Reasoning capability*: several properties characteristics, including reflexivity, symmetry, disjointness, and transitivity, have been asserted [205].
- *Ontology design*: several ontology design patterns [206] have been used in the development of SKGO ontologies, such as the OWL patterns of Gangemi [180], which is used to capture inverse relations and composition of relations.
- *Announcement*: several mailing lists, such as the W3C LOD list[8], the discussion list of the open science community[9], and discussion forums, such as those of the Open Knowledge Foundation (OKFN)[10] have been used for announcing the latest release of the ontology. We got valuable feedback regarding reusing existing ontologies and improving the documentation from several parties, including researchers in our community.

[3] `https://lov.linkeddata.es/dataset/lov/vocabs/modsci`
[4] `http://aber-owl.net/ontology/ModSci/`
[5] `http://linkeddata.uriburner.com:8000/vapour?`
[6] `http://oops.linkeddata.es/`
[7] `http://orkg.org/`
[8] `public-lod@w3.org`
[9] `open-science@lists.okfn.org`
[10] `https://discuss.okfn.org/`

Table 7.1: **SKGO ontologies.** The URLS, prefix and domain of the SKGO ontologies.

Resource	URL	Prefix	Domain
ModSci	`https://w3id.org/skgo/modsci#`	modsci	Branches in modern science and related entities (e.g., phenomenon, scientist, instrument, etc.)
PhySci	`https://w3id.org/skgo/physci#`	physci	Concepts in Physics publications (e.g., Laser, Law, etc.)
PharmSci	`https://w3id.org/skgo/pharmsci#`	pharmsci	Concepts in pharmaceutical science publications (e.g., Drug, Disease, etc.)
SemSur	`https://w3id.org/skgo/semsur#`	semsur	Concepts in computer science publications (e.g., Algorithm, Dataset, etc.)

7.2.2 Modern Science Ontology

Recent developments in the context of semantic technologies have given rise to ontologies for modeling scholarly information in various fields of science. However, most studies focused only on creating ontologies for describing scholarly articles' components, such as the structure of document elements, metadata, and bibliographic information, rather than the scientific work itself. Over the past four years, we have been engaged in the development of the Science Knowledge Graph Ontologies (SKGO), a set of ontologies for modeling the research findings in various fields of modern science resulting in a knowledge graph. In this section, we pay specific attention to the Modern Science Ontology (ModSci) for modeling relationships between modern science branches and related entities, such as scientific discoveries, phenomena, renowned scientists, instruments, etc. ModSci is a poly-hierarchical ontology that provides a unifying framework for the various domain ontologies that make up the Science Knowledge Graph Ontology suite. FAIR principles have followed in the development of the ontology. We present several use cases, particularly, two real-world use cases; the OpenResearch.org collaboration platform, and the Open Research Knowledge Graph. We deem that within the next years, a science knowledge graph is likely to become a crucial component for organizing and exploring scientific work.

Modern science is commonly divided into three major branches: Natural sciences, Social sciences, and Formal sciences. Each of these branches comprises various specialized yet overlapping scientific disciplines that often possess their nomenclature and expertise [207]. For instance, *Ecology* is a new branch of biology that deals with the relations of organisms to one another and their physical surroundings, hence an overlap between biology and Earth sciences occurs here. Another example is that Astrometrical studies use statistical methods to compute data estimates and error ranges; hence an overlap between Astrometry and Statistics occurs here. Modern science follows a set of core procedures or rules in order to determine the nature and underlying natural laws of the universe, which requires, of course, collaborations between scientists from different fields of science. For example, biologists require mathematics to process, analyze, and report experimental research data and to represent relationships between some biological phenomena. Also, statistics is used in economics in the measurement of correlation, analyzing demand and supply, forecasting through regression, interpolation, and time series analysis. ModSci is designed to represent various branches in modern science with the aim of representing relationships among them. Applications and phenomena related to each of these branches of science, as well as scientists, are also covered.

Motivation

ModSci ontology is motivated by real-life requirements that we encounter during day-to-day research and supervision work: 1) Determining which field of science best matches the interests of researchers in the early stages and what are the applications of this field, 2) Getting an overview of the instruments used in, and applications of, a particular field of science, and 3) Getting a comprehensive overview of other fields of science that study a given phenomenon. Indeed, the classification of research topics supports a diversity of research areas, such as information exploration (e.g., in digital libraries), scholarly data analytics and integration, and modeling research dynamics [208]. It greatly helps researchers who submit their work to multidisciplinary journals to position their work in the right track, thus avoiding out-of-scope rejections. Furthermore, it helps editorial teams of multidisciplinary journals in classifying submissions according to a taxonomy of research topics. To the best of our knowledge, there is yet no semantic model that organizes major fields and related sub-fields of science and emerging areas of study. In the following, we present four motivating scenarios. The objective of presenting motivation scenarios is to express the motivation behind developing the ontology and, therefore, its potential uses. The following are typical examples of the potential future uses of ModSci.

- *Scenario 1:* Alice is an undergraduate student who is interested in studying Biology and Chemistry. She wants to find out which field of science best matches her interests and what are the applications of this field.
- *Scenario 2:* Bob is a member of the biomedical research community and wants to get a comprehensive overview on the instruments used in biomedical studies, such as biomedical reactions studies.
- *Scenario 3:* for her interdisciplinary research, Charlie (a Chemistry researcher) needs to study another field of science, e.g., Biology, but she does not know the sub-fields there and the differences between them.
- *Scenario 4:* Dan is an astrophysicist studying the concrete phenomenon of black holes, who wants to get a comprehensive overview on other fields of science studying the same phenomenon.

Objective

We aim at increasing the impact of research data by making research contributions transparent, easily findable, and directly comparable (realizing the FAIR principles [17]). This goal can be achieved by providing a comprehensive and granular semantic resource that is capable of supporting ontology-based applications. This will support the digital transformation of scholarly communication from a document- to a knowledge-oriented representation in the form of structured and interlinked knowledge graphs, aiming at analyzing, exchanging, and exploiting scholarly knowledge efficiently. Besides, classification allows research and experimental development activity to be categorized by field of research. ModSci has been made publicly available (under the SKGO suite) in standard formats, at a permanent URL, following ontology publication best practices [174]. The ModSci classification tree of modern science branches is freely available for research purposes in HTML format through its documentation page.

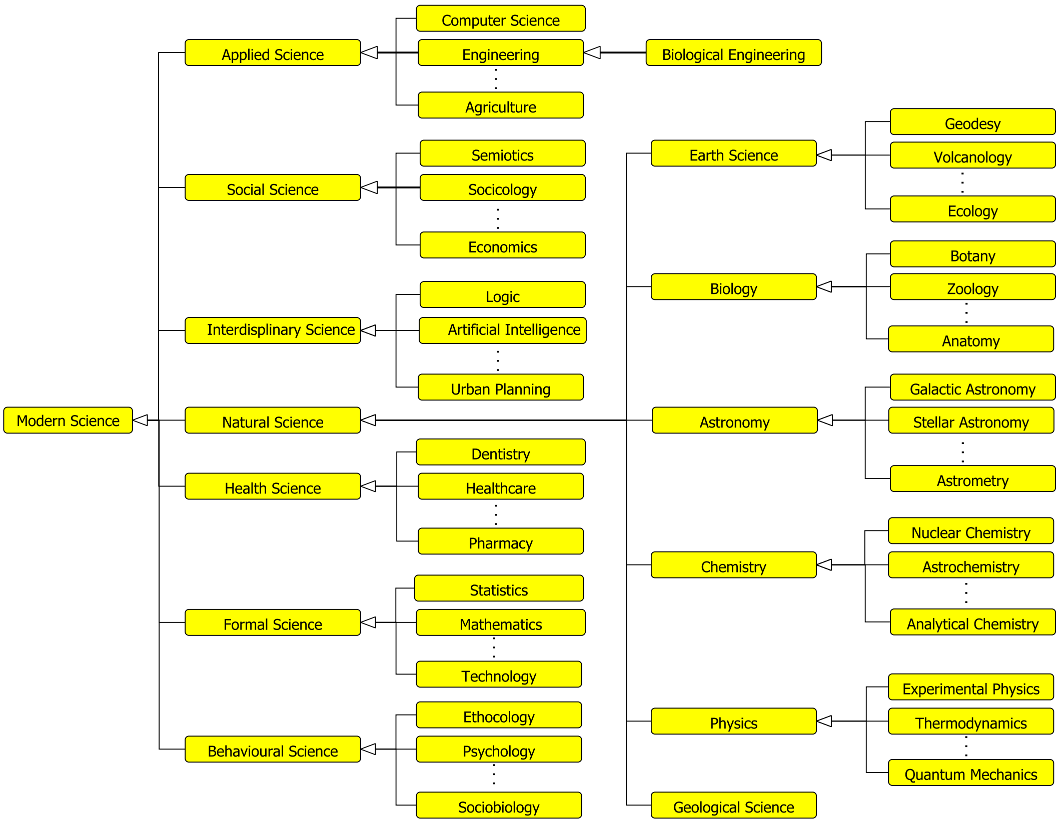


Figure 7.1: **ModSci Hierarchy.** A Graffoo diagram illustrating some of the ontological entities defined in ModSci. Arrows with open arrow heads denote the `rdfs:subClassOf` relation between the classes. For conciseness, many classes have been omitted as indicated by dots between siblings classes.

Ontologies Reuse

Several well-known taxonomies of research fields, such as Field of Research (FoR) [209], Dewey Decimal [210], and Library of Congress Classification [211] have been reused. FAIR principles, which are proposed to make data Findable, Accessible, Interoperable and Reusable [17], have followed in the development of the ontology for maximizing the added-value of the ontology. Ontologies exhibit the foundation for achieving data FAIRness.

In addition, several ontology design patterns [206] have been used in the development of ModSci, such as the OWL patterns of Gangemi [180], which are used to capture inverse relations and composition of relations. Graffoo diagrams [200] are used to illustrate the key ontological entities visually. The ontology is available in multiple RDF serializations, including Turtle, JSON-LD, RDF/XML, and N-Triples, from the SKGO GitHub repository. We reuse existing models developed for describing the scientific work in various fields of science, such as EXPO [143]. Several existing standard models have also been reused, such as the SWEET ontologies [212], SKOS, and FOAF. In addition, several taxonomies of research fields, such as Field of Research (FoR) by ANZSRC [209], Dewey Decimal [210], and Library of Congress [211] Classifications have been integrated for expanding various science branches, including mathematical, physical, and chemical sciences.

Further data sources included domain expert interviews, Wikipedia classification of science, and research area classifications by universities. The Ontology Lookup Service[11] has been used to explore biomedical ontologies that contain biomedical research classification.

Core Concepts

The pivotal concepts of the ModSci ontology are the branches of modern science and its sub-branches (a part of the hierarchy is shown in Figure 7.1). In addition, there are seven entities (we follow the definitions found in [213]) related to such concepts:

- *Modern Science*: is a systematic endeavor that builds and organizes knowledge in the form of testable explanations and predictions about the universe,
- *Scientific Discovery*: is the process or product of successful scientific inquiry, which can be things, events, and properties as well as theories and hypotheses [214],
- *Phenomenon*: is something that is observed to occur or to exist,
- *Applications of Science*: is any use of scientific knowledge for a specific purpose, whether to do more science; to design a product, process, or medical treatment; to develop new technology, or to predict the impacts of living organisms actions,
- *Scientific Organization*: is an organization that facilitates progress in a specific area through promotion of research,
- *Scientist*: is a person who conducts scientific research to advance knowledge in an area of interest, and
- *Scientific Instrument*: is a device or tool used in scientific experiments for particular purposes, such as Ammeters, for measuring the current in an electric circuit.

Concretely, these entities are represented in ModSci as `owl:Class` as follows: `ModernScience`, `ScientificDiscovery`, `Phenomenon` `ApplicationsOfScience`, `ScientificOrganization`, `Scientist`, and `ScientificInstrument`, respectively. We observed a great extent of collaboration between various fields of science, which in turn gave rise to new fields of science. For example, ecology, a branch of science that studies the distribution and interactions between living things and the physical environment, is a new field of science that combines methods and techniques from both Biology and Earth sciences. Thus, the `Ecology` class is defined as a subclass of both the `Biology` and the `EarthSciences` classes. Another example is `Biochemistry`, which is a subclass of both `Biology` and `Chemistry`.

Class Specialization: Specialization can be defined as the creation of a set of sub-classes according to one or several attributes [215]. In ModSci, an example of such specialization is the creation of `Ethology`, `Psychology`, `SocialPsychology`, and `Sociobiology` as sub-classes of `BehavioralSciences`.

Class Disjointness: expresses that, speaking in colloquial terms, two classes are incompatible, i.e., they do not share any common entities. Adding disjointness axioms to ontologies enables a wide range of noteworthy applications [216]. We explicitly asserted the pairwise disjointness of various classes in the ontology. For instance, the `AstronomicalPhenomena` class is disjoint with `BiologicalPhenomena`.

[11] `https://www.ebi.ac.uk/ols/index`

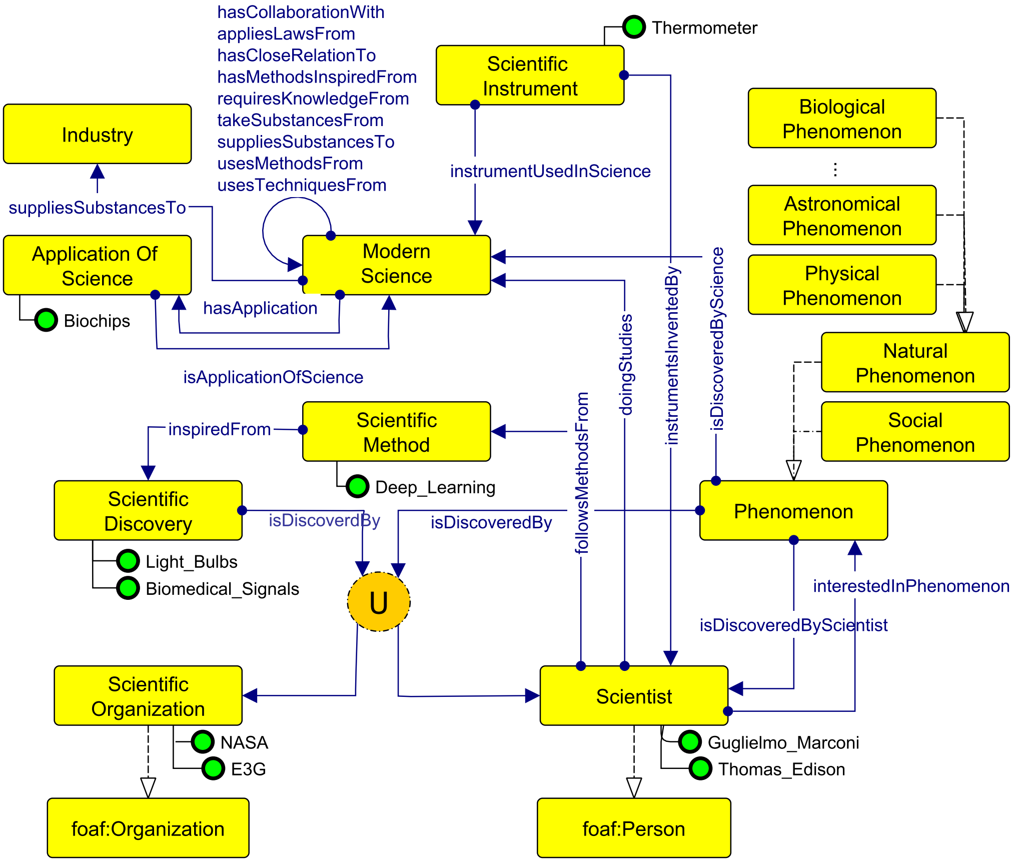
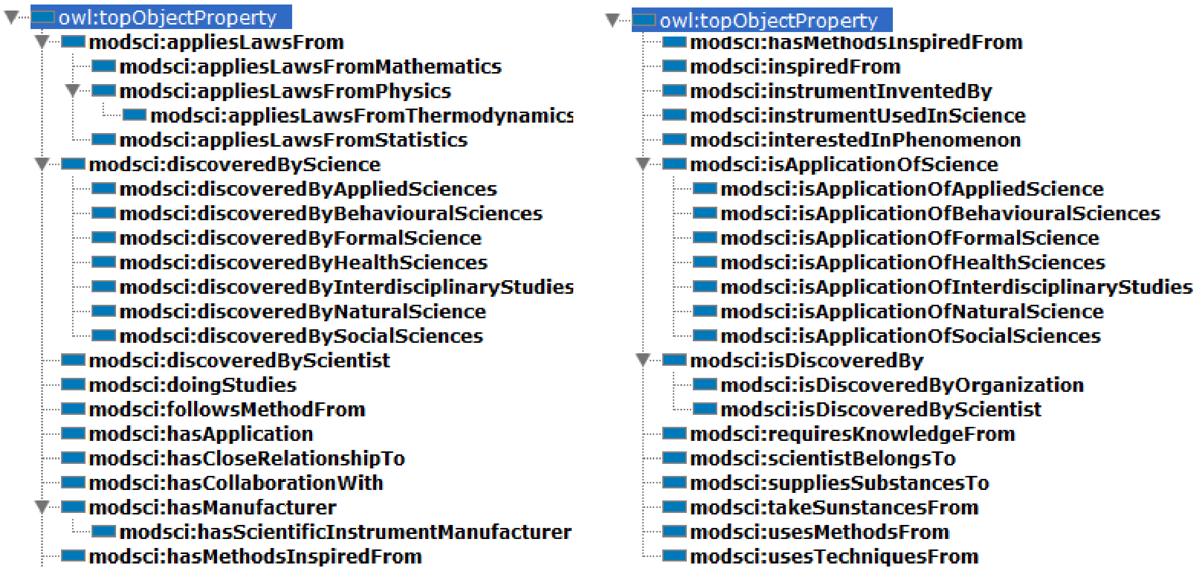


Figure 7.2: **Core Concepts of ModSci.** A Graffoo diagram illustrating the core concepts of ModSci and their interlinking relationships. Open arrow heads denote `subClassOf` properties between the classes. Several reflexive properties are represented as loops for better readability (upper-left corner). Circle with "U" inside refers to the union operator.

Semantic Relations

Object properties in ModSci have been restricted by defining their domain and range. For instance, the domain and range of `hasApplication` are `ModernScience` and `ApplicationOfScience` respectively. *Logical disjunction.* Some properties have complex ranges and domains, e.g., the domain of `discoveredByScience` is (`Phenomenon` ⊔ `ScientificDiscovery`), which means that a Phenomenon or a Scientific Discovery can be discovered by a particular Science. A full view of the properties defined in ModSci, including their domains and ranges, is shown in Figure 7.2.

Representation pattern of n-ary relations. One common representation of n-ary relations is to represent the relation as a class rather than property and use n properties to point to the related entities [217]. Instances of such classes are instances of the n-ary relation, and additional properties can provide binary links to each argument of the relation, i.e., an instance of the relation linking the n individuals [217]. For example, consider the case of representing that Biology facilitated Physics in the discovery of *Energy conservation* phenomenon. This case can only be represented as an n-ary relation. As shown in Figure 7.4, the individual *:helpInDiscoveryOfEnergyConservation* represents a single object encapsulating the information that both sciences helped in the discovery of the phenomenon "*Energy Conservation*". Figure 7.3 depicts the object properties



- owl:topObjectProperty
  - modsci:appliesLawsFrom
    - modsci:appliesLawsFromMathematics
    - modsci:appliesLawsFromPhysics
      - modsci:appliesLawsFromThermodynamics
    - modsci:appliesLawsFromStatistics
  - modsci:discoveredByScience
    - modsci:discoveredByAppliedSciences
    - modsci:discoveredByBehaviouralSciences
    - modsci:discoveredByFormalScience
    - modsci:discoveredByHealthSciences
    - modsci:discoveredByInterdisciplinaryStudies
    - modsci:discoveredByNaturalScience
    - modsci:discoveredBySocialSciences
  - modsci:discoveredByScientist
  - modsci:doingStudies
  - modsci:followsMethodFrom
  - modsci:hasApplication
  - modsci:hasCloseRelationshipTo
  - modsci:hasCollaborationWith
  - modsci:hasManufacturer
    - modsci:hasScientificInstrumentManufacturer
  - modsci:hasMethodsInspiredFrom

- owl:topObjectProperty
  - modsci:hasMethodsInspiredFrom
  - modsci:inspiredFrom
  - modsci:instrumentInventedBy
  - modsci:instrumentUsedInScience
  - modsci:interestedInPhenomenon
  - modsci:isApplicationOfScience
    - modsci:isApplicationOfAppliedScience
    - modsci:isApplicationOfBehaviouralSciences
    - modsci:isApplicationOfFormalScience
    - modsci:isApplicationOfHealthSciences
    - modsci:isApplicationOfInterdisciplinaryStudies
    - modsci:isApplicationOfNaturalScience
    - modsci:isApplicationOfSocialSciences
  - modsci:isDiscoveredBy
    - modsci:isDiscoveredByOrganization
    - modsci:isDiscoveredByScientist
  - modsci:requiresKnowledgeFrom
  - modsci:scientistBelongsTo
  - modsci:suppliesSubstancesTo
  - modsci:takeSunstancesFrom
  - modsci:usesMethodsFrom
  - modsci:usesTechniquesFrom

Figure 7.3: A part of the object properties hierarchy in ModSci (*max depth* = 3).

hierarchy in ModSci in which, e.g., `appliesLawsFromThermodynamics` is a sub-property of `appliesLawsFromPhysics`, which in turn is a sub-property of `appliesLawsFrom`. *Property restrictions.* A property restriction provides a type of logic-based constructors for complex classes by defining a particular type of class description, which is a class of all individuals that satisfy the restriction [218]. OWL defines two kinds of property restrictions: value constraints (restricting the range of the property) and cardinality constraints (restricting the number of values a property can take). One example of a property restriction in ModSci is the use of `owl:minCardinality` for restricting `discoveredByScientist` to assure that a phenomenon is discovered by at least one scientist (`owl:someValuesFrom`). Another kind of property restriction is the *owl:allValuesFrom* constraint, which restricts the individuals used as objects with a given property to be either a member of a particular class or data values within a specified set of values. For instance, the property `discoveredByScientist` has been restricted by *owl:allValuesFrom* to the class `Scientist`.

Supporting Inference

Near the top of the well-known Semantic Web technology stack [219] one finds “inference”, which comprises reasoning over linked data through rules. On the Semantic Web, the inference is used to detect relationships between ontology concepts (which not explicitly given), detect inconsistencies, and infer logical consequences of a set of asserted axioms [220]. Several ontology design patterns have been applied to support the inference process, e.g., the OWL patterns of Gangemi and Presutti [180]. The `isApplicationOfScience` property being an inverse of `hasApplication` is an example of an inverse relation. Thus, if an application of science (e.g., a biochip) A `isApplicationOfScience` S, then it can be inferred that S `hasApplication` A. Another way of supporting the inference process is to define symmetric relations. For instance, `hasCloseRelationshipTo` is a symmetric relation where, if Statistics is connected to Mathematics via this property, then the opposite also holds. Furthermore, some properties whose domain is the same as their range, e.g., `hasCollaborationWith`, having both domain

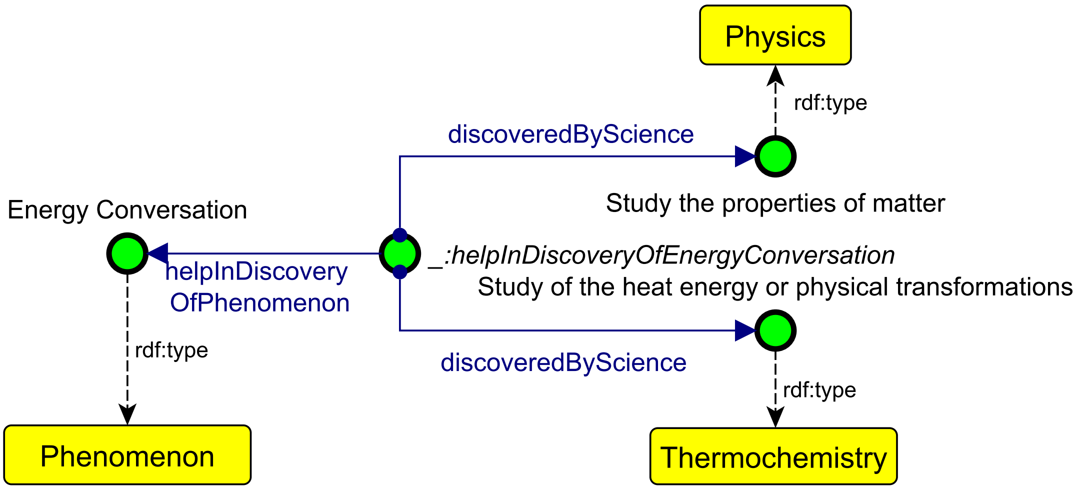

Physics
rdf:type
discoveredByScience
Study the properties of matter
Energy Conversation
helpInDiscovery OfPhenomenon
_:helpInDiscoveryOfEnergyConversation
Study of the heat energy or physical transformations
rdf:type
discoveredByScience
rdf:type
Phenomenon
Thermochemistry

Figure 7.4: **N-ary relation Example.** An example of the representation pattern for n-ary relations in ModSci.

and range `ModernScience` and thus providing the information that there exist collaborations between two modern sciences, are defined as reflexive relations, i.e., scientists in a particular field of science have collaborations with themselves. We have also defined symmetric properties, such as `hasCollaborationWith`. Finally, a set of SWRL [182] (Semantic Web Rule Language) rules (listed below) have been defined for discovering new relationships and inferring new knowledge that is not explicitly given in the ontology. These rules have been semantically validated using the Drools reasoner [181].

$$discoveredByScientist\,(x,y)\ \wedge\ discoveredByScience\,(x,z)\ \ \rightarrow\ \ scientistBelongsTo\,(y,z) \tag{7.1}$$

$$Scientist\,(x) \wedge isDiscoveredBy\,(a,x) \rightarrow isDiscoveredByScientist\,(a,x) \tag{7.2}$$

$$Scientist\,(x)\ \ \wedge\ \ doingStudies\,(x,s)\ \ \rightarrow scientistBelongsTo\,(x,s) \tag{7.3}$$

$$ScientificOrganization\,(x) \wedge isDiscoveredBy\,(a,x) \rightarrow isDiscoveredByOrganization\,(a,x) \tag{7.4}$$

Ontology population

To aid the development and testing of ModSci, we have created over 150 instances (e.g., `ScientificInstrumentManufacturer` (17), `ScientificInstrument` (35), `AtmosphericPhenomena` (5), `Scientist` (10), and `ScientificOrganization` (8)). Figure 7.5 depicts the relationships between a sample of instances in ModSci. These instances help to assist in characterizing core concepts within the ontology and to provide links (where available) between ModSci and the reused ontologies. Even though some of these instances are not required for evaluating the ontology, they are essential for understanding the domain; hence they help in the development process. Instances are defined with individual axioms, also called "facts"; green

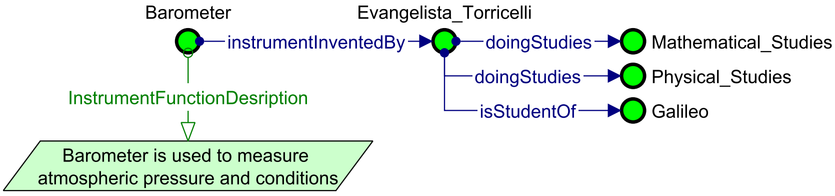


Figure 7.5: Relationships between some instances in ModSci.

circles in Figure 7.2 and Figure 7.5 present some of these individuals. We manually created two types of facts: 1) facts about class membership and property values of individuals: for example, deep learning algorithms (an individual of `algorithms`), or Non-Negative Matrix Factorization (NMF), are based on biological data called "biomedical signals" (also called *Biosignals*), and 2) facts about identical individuals. The OWL `owl:sameAs` construct is used to establish the identity of individuals, i.e., states that two URIs refer to the same individual. For example the scientist "Thomas Alva Edison" and "Edison" refer to the same thing as follows:

```
<rdf:Description rdf:about="#Thomas_Alva_Edison">
<owl:sameAs rdf:resource="#Edison"/>
</rdf:Description>
```

Real use cases

Several concrete real-world use cases illustrate the added value of ModSci in several application areas, including cross-disciplinary indexing, enriched bibliographic data, and network analysis within cross-disciplinary scientific communities. Particularly, we present two use cases of ModSci;

Use case 1 (*Open Research Knowledge Graph*[12]): ModSci is being integrated into the Open Research Knowledge Graph (ORKG) [221] to support the classification of research papers. ModSci is used in the classification system for research papers in the Open Research Knowledge Graph (ORKG). The ORKG is a step towards the next generation of digital libraries for semantic scientific knowledge communicated in scholarly literature [221]. ModSci is being integrated into the step of selecting the research field of the research papers added to the knowledge graph[13], which provides more than 200 research fields in various fields of modern science. Besides, it can be used in browsing the research papers by fields through "Browse by research field" option[14].

Use case 2 (*OpenResearch.org*): OpenResearch.org contains scholarly information in several fields of science, i.e., not restricted to particular fields. This semantic wiki aims at making scholarly information more accessible and shareable. We are developing a wiki-based approach to generate overviews of research domains and their relevant artifacts in OpenResearch.org [222]. ModSci is used to categorize information about scientific events, research projects, scientific papers, publishers, and journals.

Data-driven Evaluation

We performed an ontology Verification & Validation (V&V) following the guidelines proposed in [201]. The evaluation has been conducted in two strategies: Verification by human evaluation

[12] https://projects.tib.eu/orkg/
[13] https://www.orkg.org/orkg/add-paper
[14] https://www.orkg.org/orkg/

Table 7.2: **Competency Questions.** The main competency questions defined for ModSci. A variable like X is a placeholder for any suitable value.

Id	**Question text**
CQ1	What are the main branches of modern science and their sub-branches?
CQ2	Are there any collaborations of scientists from various fields of science to produce a product X? (*derived from F5*)
CQ3	What are the instruments used in a particular study X belonging to the scientific field Y?
CQ4	What scientists discovered the phenomenon X?
CQ5	What are the fields of science that combine both science X and science Y?
CQ6	What are the differences between different subfields of science X?
CQ7	What are typical applications of science X?
CQ8	What discoveries were made by scientists belonging to a particular science X?
CQ9	In which field of science was the phenomenon X discovered?
CQ10	Which fields of science have a close relationship? (*derived from F2*)
CQ11	Which fields of science use methods from science X to do Y? (*derived from F6 and F3*)

and Validation by test cases. After identifying motivational scenarios in a use case fashion, the next step is to derive a set of competency questions (CQs) from these scenarios.

Competency Questions. Competency questions serve as a kind of functional requirements specification for an ontology. They help the ontology engineer to identify relevant concepts, instances, object and datatype properties, as well as constraints. They also help to define the scope of knowledge which the ontology encapsulates effectively. Table 7.2 presents the main set of competency questions we have defined for ModSci. In the end, the ontology should be able to answer the competency questions. CQs have been derived from a set of facts, either collected from interviewing researchers from various fields of science or collected from scientific articles. Some of these facts are listed below:

- F1: Biology applies natural physical laws since all living matter is composed of atoms.
- F2: Organic chemistry has a close relationship to biology since it supplies its substances.
- F3: Biology requires mathematics to process, analyze, and report experimental research data and to represent relationships between some biological phenomena.
- F4: Biology requires history (particularly Phylogeny) to address the evolutionary process of species.
- F5: The production of *psychiatric drugs* is a result of studying the relationship between chemistry and psychology.
- F6: Astrometry uses statistics to calculate error correction. The results are then analyzed using statistical methods to compute data estimates and error ranges.

Verification by human. This evaluation has been conducted by means of expert judgment (ontology engineering experts), in which the concepts, relations and axioms defined in the ontology

Table 7.3: **Matched entities.** A part of the verification process of ModSci.

CQ	Matched entities
CQ1	(AppliedScience, subClassOf, ModernScience) (HealthSciences, subClassOf, ModernScience) (ComputerScience, subClassOf, AppliedScience)
CQ3	(Thermometer, instrumentUsedInScience, Studying_biochemical_reactions) (Telescope, instrumentUsedInScience, Light_magnification)
CQ5	(BioChemistry, subClassOf, Biology and Chemistry) (Semiotics, subClassOf, SocialScience and InterdisciplinaryScience)

Table 7.4: **Test cases.** Sample test cases.

Id	CQ	Inputs	Expected Results
T01	CQ01.01	Social Sciences	Linguistics, Natural Language Processing Anthropology, (no sub-classes)
T02	CQ01.02	Astronomy	Astrometry, Cosmology
T03	CQ02.01	Light magnification, Astronomy	Telescope
T04	CQ04.01	Physics	Conservation_of_energy

have been checked regarding whether they are able to answer the predefined competency questions (cf. Table 7.2) [223]. This approach enabled us not only to check whether the ontology can answer the CQs but also whether there are irrelevant (should be removed) or missing (should be added) terms in the ontology. Therefore, we performed this evaluation step in parallel with the ontology development in an iterative manner, which significantly helped in improving the ontology. After five iterations (development-to-evaluation and vice versa), we obtained the final version of ModSci. Table 7.3 illustrates a part of the verification process of ModSci, showing matched entities corresponding to the CQs. This verified that ModSci is able to answer *all* competency questions defined.

Validation by test cases. The aim of this evaluation is the validation of SPARQL queries over possible evaluation scenarios or test cases. Several evaluation scenarios have been prepared, from the predefined competency questions, in a competency question-driven approach for ontology testing. Some of the competency questions are defined at a high level of abstraction to help determining the scope of the ontology and its potential uses. Thus, in order to design test cases, we derived more specific questions from them. For example, we rewrote CQ01 more specifically as: “CQ01.01: What are the main branches of *Social Sciences* and their sub-branches?” and “CQ01.02: What are the sub-fields of *Astronomy*?”. In addition, we rewrote CQ9 more specifically as: “CQ09.01: List all phenomena discovered by *Physics* along with the scientists who discovered them?”. We implemented SPARQL queries corresponding to each of these questions, thus enabling queries of instances of the ontology. In Table 7.4, we present a sample of test cases. To query the ontology and execute test cases, an instance of the Virtuoso SPARQL endpoint[15] has been set up. Listing 7.1 shows the SPARQL query corresponding to CQ09.01, which is used in T04. After executing each test case, the returned results have been compared with the expected results and compute the recall. If the recall was less than 1.0, which means that not all required

[15] `https://virtuoso.openlinksw.com/`

results (identified by experts) were returned, we analyzed the reason and iteratively adapted the ontology and re-executed the test case until all expected results were returned, i.e., we obtained a recall of 1.0. In this case, we marked the test case as *passed*.

```
SELECT DISTINCT ?phenom  ?scientist
WHERE {
  ?phenom     modsci:discoveredByScientist  ?scientist.
  ?scientist  modsci:doingStudies           ?studies.
  ?studies    rdf:type                      ?science.
  FILTER(?science = modsci:Physics)
}
```

Listing 7.1: **SPARQL query** corresponding to CQ09.01, which is used in T04.

Summary

The Modern Science Ontology models relationships between modern science branches and related entities, such as scientific discoveries, renowned scientists, instruments, phenomena, etc. ModSci has been made publicly available in standard formats, at a permanent URL, following best practices for ontology publication and vocabulary metadata completion. Several design principles have been taken into consideration in the development of ModSci, such as publication and configuration to support semantic web applications, registration in online services for ontology visualization and exploration, validation, creation of human-readable documentation, and sustainability. To maximize reasoning capability, 1) several property characteristics, such as reflexivity, symmetry, and transitivity, have been asserted, 2) disjointness of roles, and 3) several logic rules have been added to the ontologies. Motivating examples affirmed the usefulness and potential uses of ModSci ontologies. The verification by expert evaluation and the ontology development have been performed in parallel in an iterative manner, which significantly helped in improving the ontology. ModSci closes an important gap between scientists in different fields of science, where, for example, a computer scientist can easily know about other fields of science (e.g., scientific methods, instruments, applications, etc.) when there is a plan for future collaboration.

7.2.3 Ontology for Computer Science Research Data Modeling

In this section, we present the latest version of the Semantic Survey Ontology (SemSur), which is used to describe how research data in Computer Science can be represented semantically, resulting in a knowledge graph that describes the individual research problems, approaches, implementations and evaluations in a structured and comparable way. In 2017, an initial version of SemSur [6] has been created in order to support our approach towards representing research findings as a knowledge graph and has now been substantially improved and expanded.

Overview

The traditional way of providing researchers an overview of a particular research problem and enabling them to compare state-of-the-art is via *survey articles*. However, preparing such articles and studying individual articles requires a significant amount of time, often several months of work. The major drawback from the reader's perspective is that most survey articles are published in printed form or as semi-structured digital (e.g., PDF) documents, which does not make them efficiently accessible for comparative or other analyses. A striking feature of

SemSur is that it supports retrieving, exploring, and comparing research findings based on an explicit semantic representation of the knowledge contained in scientific publications. If applied widely, SemSur can have a significant impact on scholarly communication. Specifically, it addresses researchers who want to compare their research with related works, get an overview on contributions in a particular research topic, or find research contributions addressing a specific problem or having certain characteristics.

Most of the prior work on the semantic representation of scholarly communication focused on either describing the document structure or bibliographic information rather than the scientific content itself. In the initial version of SemSur in 2017, we identified the high level of abstraction of concepts, which lack the alignment with other related ontologies. This version also suffers from the absence of rules for the elicitation of implicit relations, such as co-authorship. The ontology is structured around five core concepts, as shown in Figure 7.6:

- *Research problem* – represents a challenge in a particular field, possibly hierarchically decomposed into sub-problems, which have related problems, a motivation and possible approaches addressing the problem,
- *Approach* – comprises research methods and procedures to address a particular research problem,
- *Implementation* – describes the implementation of an approach in a concrete technical environment,
- *Evaluation* – describes how an implementation is evaluated using an evaluation method in a defined scenario,
- *Publication* – refers to an article and accompanying bibliographic metadata, including authors, title, keywords and abstract.

This section gives a comprehensive overview on SemSur, focusing on the main new features compared to the initial version, including broader coverage of the domain, better alignment with external ontologies, and rules for eliciting implicit knowledge. The open W3C standards RDF and OWL were used to develop SemSur, SWRL for maintenance and quality checks as well as SPARQL for querying data adhering to SemSur. The ontology is publicly available at `http://purl.org/semsur/owl/`, subject to the Creative Commons Attribution license. The documentation is available through its URL.

Motivating Scenario

We present a motivating scenario in order to motivate our approach and to illustrate how SemSur can be used to complement review papers for comparative evaluation of research findings. Two kinds of researchers can benefit from SemSur 1) researchers who want to publish their research findings as a knowledge graph to be FAIR (findable, accessible, interoperable, reusable) for other researchers, and 2) researchers who want to get an overview of the research efforts related to a particular research problem, e.g., to write a survey paper or a related work analysis.

For more illustration, suppose that *Alice* conducts research about "SPARQL query Federation, and she wants to publish her research findings as a knowledge graph. *Bob* wants to reuse Alice's research findings or get an overview of the state of the art of research on the same topic in terms of research motivations, currently published approaches, implemented frameworks/tools, and challenges faced. *Bob* can then construct a new experiment by replacing the dataset used or modifying the approach, and then republish the new findings as a continuation of Alice's original, including a link to her paper. Thus, *Alice* is the research producer, while *Bob* is the

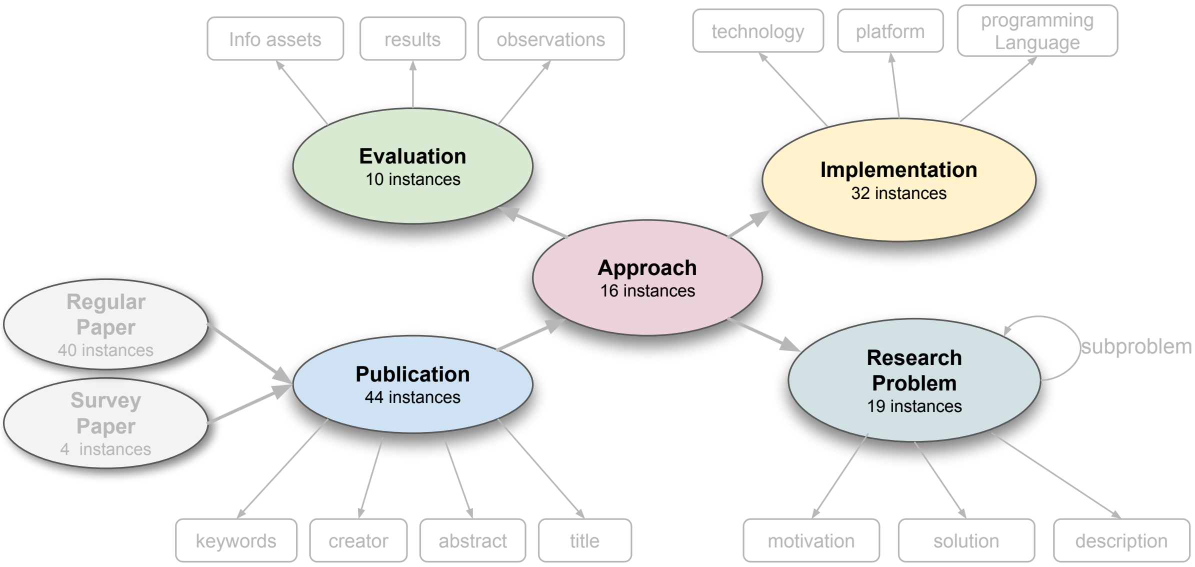


Figure 7.6: **SemSur core concepts.** The five core concepts in the SemSur domain.

consumer. To support this scenario, a comprehensive ontology for describing research findings and their relationships, and a platform for adding and retrieving them are needed. Listing 7.2 shows the corresponding SPARQL query to achieve the goal. The output of the query includes the approach implementation, published article, research problem and motivation, and results of the experiments along with experimental requirements and goals.

Methodology

In this section, we describe the methodology we have followed in the development process of SemSur. Ontology development is an iterative process that aims at producing an efficient and well-formed ontology. We have iteratively interviewed several experts, including ontology engineering experts and domain experts, during the whole process in order to improve the quality of the resultant ontology. The following steps drive the development of SemSur:

- *Exploring the domain* – every research domain has its own culture, requirements, and findings, e.g., technical fields such as computer science have implementations, and other fields such as agriculture have different concepts like machines, etc. In this step, experts in a research domain explore the culture, needs, and findings of that domain and define concepts based on that. Some concepts, such as *research problem*, are needed in all research domains.
- *Asking experts* – brainstorming with other domain experts on the concepts defined in the previous step in order to validate, remove or update them as well as add missing ones.
- *Reusing ontologies* – in the ontology refinement process, we explore the terms of already existing vocabularies to select the best matches and reuse or align with them.
- *Adding missing concepts* – if existing vocabularies comprise the identified concepts, then we reuse these directly, specialize them or add a property restriction; otherwise, we define them appropriately.
- *Defining inference rules* – define inference rules, mostly, for properties which can be defined

implicitly.

- *Implementing the ontology* – using an ontology editor as well as adding labels, descriptive comments, and metadata.

```
SELECT DISTINCT ?Motivation  ?Approach  ?expGoal  ?pos  ?chall
WHERE {
  or:SPARQL_query_Federation  or:hasMotivation      ?Motivation .
  ?Motivation         or:hasDescription     ?MotivationDescription .
  ?Motivation         or:motivatesApproach  ?Approach .
  ?Approach           or:hasImplementation  ?Framework .
  ?Approach           or:hasPositiveAspects ?pos .
  ?Approach           or:hasChallenges      ?chall .
  ?Approach           or:hasEvaluation      ?eval .
  ?eval               or:run                ?exp .
  ?exp                expo:has_experimental_requirements ?expGoal .
}
```

Listing 7.2: **SPARQL Query of Q10.** Corresponding SPARQL query for *"What are the motivations, approaches, experiment goal and frameworks for "SPARQL query Federation" and possible challenges and positive aspects?"*

Reuse of External Ontologies

It is known that when developing an ontology-based application, it is important to reuse and integrate existing ontologies to provide the background knowledge required by the application [224]. According to best practices (e.g., [225]), developing ontologies becomes easy and efficient by reusing existing ontologies. Ontology reuse should begin by identifying candidate ontologies to be reused, having them evaluated by domain experts, and choosing the adequate ontologies to be reused among the candidate ontologies having the highest quality [225].

Pinto and Martins [226] proposed two different ontology reuse processes: fusion/merging and composition/integration. *Fusion* means building ontologies by unifying knowledge from source ontologies in the same domain as the target ontology [227]. *Composition* means building ontologies by assembling two or more ontologies that might come from domains different from the domain of the target ontology [227]. An ontology fusion process has been performed by re-using seven ontologies from the scholarly communication domain. We have selected the most closely related ontologies listed in the Linked Open Vocabularies (LOV) directory. The vocabularies reused by SemSur are listed in Table 7.5. SemSur is aligned and linked with the following related ontologies from three categories for:

- *describing scholarly articles*, we reused the DC, SWRC, DoCO, EXPO and FOAF ontologies,
- *describing the inner structure of a scientific article* independently of the field of research, we reused DEO and LSC, and
- *describing concepts of specific fields of research*, we reused MLS and DOAP.

Core Concepts

SemSur classes are divided into three groups: 1) classes from the imported ontologies, 2) classes from previously-imported[16] ontologies, and 3) newly defined classes, which are not exist in any

[16] Ontologies imported in SemSur 1.0.

Table 7.5: **Reused vocabularies.** Prefixes and namespace URIs of the reused vocabularies.

Prefix	Vocabulary	URI
dc	Dublin Core Metadata Initiative	`http://purl.org/dc/terms/`
swrc	Semantic Web for Research Communities	`http://swrc.ontoware.org/ontology#`
foaf	Friend of a Friend ontology	`http://xmlns.com/foaf/0.1/`
mls	Machine Learning Schema	`https://www.w3.org/ns/mls#`
deo	The Discourse Elements Ontology	`http://purl.org/spar/deo/`
lsc	Linked Science Core Vocabulary	`http://linkedscience.org/lsc/ns#`
doap	Description of a Project	`http://usefulinc.com/ns/doap#`
doco	Document Components Ontology	`http://purl.org/spar/doco`
expo	Scientific EXPeriments Ontology	`http://www.hozo.jp/owl/EXPOApr19.xml/`

Table 7.6: **SemSur core classes.** Main classes defined by SemSur and reused from other ontologies.

Group	Classes	Source
reused	Abstract, Appendix	doco
	ScientificExperiment, ExperimentResult, ExperimentDesign, ExperimentRequirements, ExperimentHypothesis, Model, ExperimentalModel, DomainModel	expo
	Questionnaire	fabio
	Organization, Employee, Person	foaf
	ResearchProject, DevelopmentProject, ResearchTopic, Manual, Book, MasterThesis, PhDThesis	swrc
newly defined	SurveyPaper, RegularPaper, InformationAsset, Complexity, PositiveAspects, Limitations, Documentation, Toolbox, Challenges, SimulationSoftware, SingleAuthorPublication, Toolbox, Benchmark	semsur

of the reused ontologies. Some of these classes need more specialization, so we created respective subclasses. For instance, we added three subclasses `Mathematical Model`, `ArchitecturalModel` and `PipelineModel` for the `Model` class inherited from the MLS ontology. Another concern is the integration of imported ontologies. In other words, classes imported from an ontology should have a proper relation with related classes found in the other ontologies. For instance, the `Article` class (from SWRC) should have a relation with the `Conclusion` class (from LSC) with the relation *produces* (from LSC). Newly added classes in SemSur 2.0 are highlighted in light-blue in Figure 7.8.

Class specialization. We created specializations of some imported classes from these ontologies, such as `SurveyPaper` and `RegularPaper` as specializations of the `swrc:Publication` class. Another concern is the integration of the reused ontologies, e.g., the `swrc:Publication` class was put into a relationship with `lsc:Conclusion` via the `lsc:produces` property. Table 7.6 covers the main classes defined by SemSur and reused from other ontologies. Figure 7.7 shows the class hierarchy of the Publication class from general to specific classes, i.e., from `SurveyPaper` to `Publication`.

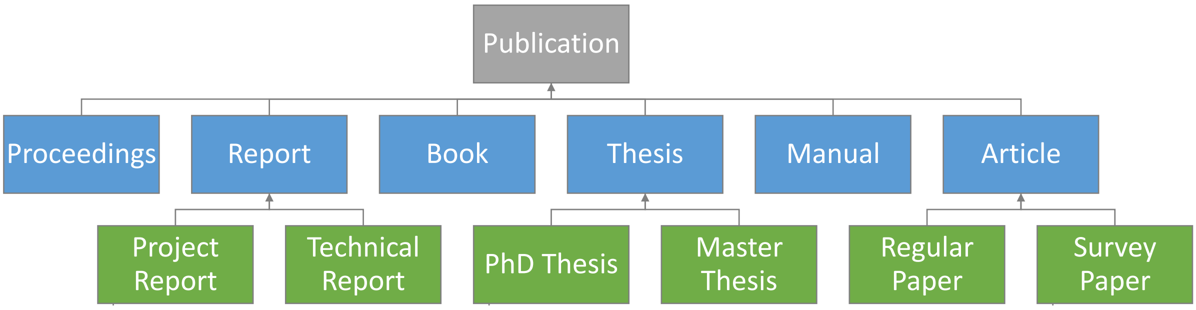


Figure 7.7: **Publication class** hierarchy in SemSur.

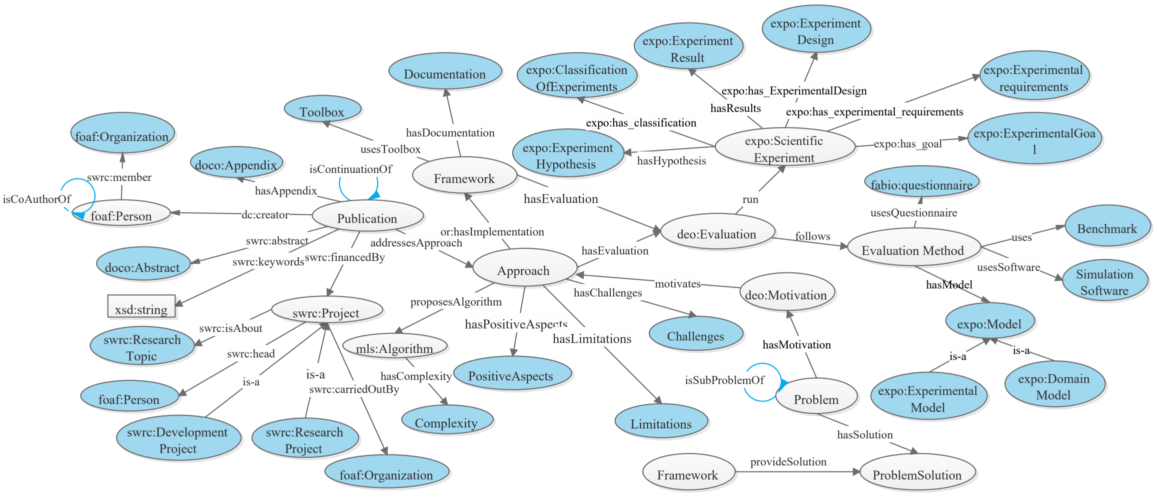


Figure 7.8: **Overview of the SemSur 2.0 graph.** New classes in SemSur 2.0 are highlighted. Some classes from SemSur 1.0 were omitted for better visualization.

Relations

Similarly, Table 7.7 lists the main relations defined by SemSur as well as the reused ones. SemSur has two transitive relations: for representing co-authorship between researchers and for representing that a research problem has sub-problems. Co-authorship means that there is cooperation between two or more authors in a publication. The relation `isCoAuthorOf` is a transitive relation, when considered in the restricted scope of one publication, that represents the co-authorship between two `Persons`. Furthermore, it is also a symmetric relation. In addition, new transitive relations have been defined such as `isContinuationOf` and `isSubProblemOf`. Also new symmetric relations have been defined, such as `isCoAuthorOf` and `hasRelatedProblem`.

Supporting Inference

A standard way to infer new information on the Semantic Web is to define inference rules [228]. Inference on the Semantic Web is used to improve the quality of data integration in the ontology by combining rules and ontologies to discover new relationships, detect possible inconsistencies and infer logical consequences from a set of asserted facts or axioms. Several languages and

Table 7.7: **SemSur Relations.** Main relations defined by SemSur and the reused ones.

Group	Relations	Source
reused	creator, title, hasVersion	dc
	has_experimental_requirements, has_classification, has_ExperimentalDesign, has_goal	expo
	vendor, OS, platform	doap
	name	foaf
	carriedOutBy, head, isAbout, abstract, member, financedBy, keywords	swrc
newly defined	hasLimitations, hasPositiveAspects, proposesAlgorithm, addressesApproach, isContinuationOf, hasAppendix, addressesApproach, hasChallenges, motivates, hasMotivation, provideSolution, hasSolution, hasEvaluation, hasHypothesis, hasResults, usesQuestionnaire, hasDocumentation, usesToolbox, isCoAuthorOf	semsur

standards have been proposed for writing rules for ontologies, including RuleML[17] (Rule Markup Language), Jess [229] (Java expert system shell) and SWRL (Semantic Web Rule Language) [230]. Our goal is to define a rule set for discovering new relationships and inferring new knowledge from instance data and class descriptions, which did not explicitly exist [9]. SWRL is designed based on a combination of the OWL DL and OWL Lite sublanguages of OWL. It allows users to write Horn-like rules expressed in terms of OWL classes and properties to reason about OWL individuals [182]. These rules are used to infer new knowledge from the existing OWL knowledge bases. According to O'Connor et al. [183], SWRL rules are written as antecedent (body)/consequent (head) pairs. The antecedent is the rule's body, and the consequent is the head. Head and body consist of a conjunction of one or more atoms. As an example for inference, the `SingleAuthorPublication` class represents publications with only a single author. Individuals of this class can be inferred by Equation 7.5. We have defined the following SWRL rules[18]:

$$SingleAuthorPublication(p) \leftarrow Publication(p) \wedge creator(p,x) \wedge creator(p,y) \wedge (x=y) \quad (7.5)$$

$$\textit{swrc}:\textit{Publication}(?p) \wedge \textit{dc}:\textit{creator}(?p,?x) \ \wedge \ \textit{swrlb}:\textit{equal}(?x,1) \\ \rightarrow \textit{SingleAuthorPublication}(?p) \quad (7.6)$$

$$\textit{swrc}:\textit{Publication}(?p) \wedge \textit{dc}:\textit{creator}(?p,?x) \wedge \textit{dc}:\textit{creator}(?p,?y) \wedge \textit{owl}:\textit{differentFrom}(?x,?y) \\ \rightarrow \textit{isCoAuthor}(?y,?x) \quad (7.7)$$

[17] `http://wiki.ruleml.org`

[18] Every formula is assumed to be universally quantified over all its free variables. The equality symbol = denotes primitive logical equality, and ≠ denotes its negation.

Table 7.8: **SemSur Versions Statistics**. A comparison between SemSur versions statistics.

Metrics	**SemSur 1.0**	**SemSur 2.0**
Axioms	6,161	16,880
Logical axiom	1,696	11,260
Declaration axioms	1,076	1,867
Class	294	876
Object property	281	341
Data property	109	140
Individual	354	415
Annotation property	113	98

$$\begin{aligned} &\mathit{Problem}(?x) \;\wedge\; \mathit{Problem}(?y) \;\wedge\; \mathit{isSubProblem}(?x, ?y) \;\wedge\; \mathit{hasMotivation}(?y, ?m) \\ &\qquad \rightarrow \mathit{hasMotivation}(?x, ?m) \end{aligned} \tag{7.8}$$

$$\begin{aligned} &\mathit{swrc} : \mathit{Publication}(?x) \wedge \mathit{swrc} : financedBy(?x, ?y) \;\wedge\; \mathit{swrc} : \mathit{isAbout}(?y, ?z) \\ &\qquad \rightarrow \mathit{swrc} : \mathit{isAbout}(?x, ?z) \end{aligned} \tag{7.9}$$

In order to express the co-authorship between authors, we introduce the rule in Equation 7.7. After running Drools reasoner on the ruleset, a significant number of new (ABox) axioms have been inferred. Table 7.8 shows the ontology statistics after running the rule engine and the successful transformation of the new knowledge into OWL. SQWRL (Semantic Query-Enhanced Web Rule Language) is an SWRL-based query language that provides SQL-like operators for extracting information from OWL ontologies [231]. The following SQWRL query is used to retrieve all single-author publications along with the author name ordered by author names.

$$\begin{aligned} &\mathit{SingleAuthorPublication}(?p) \wedge \mathit{dc} : \mathit{creator}(?p, ?x) \\ &\qquad \rightarrow \mathit{sqwrl} : \mathit{select}(?p) \wedge \mathit{sqwrl} : \mathit{orderBy}(?x) \end{aligned} \tag{7.10}$$

Individuals

The final step in ontology engineering is the creation of instances/individuals of classes [73]. To better understand the domain of SemSur and to build test cases, we have created several individuals representing the scientific content in a sample of four survey articles, which published in high-quality venues.

- Bringing Relational Databases into the Semantic Web: A Survey. [232]
- A Survey of Current Link Discovery Frameworks. [233]
- Querying over Federated SPARQL Endpoints –A State of the Art Survey. [16]
- Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. [86]

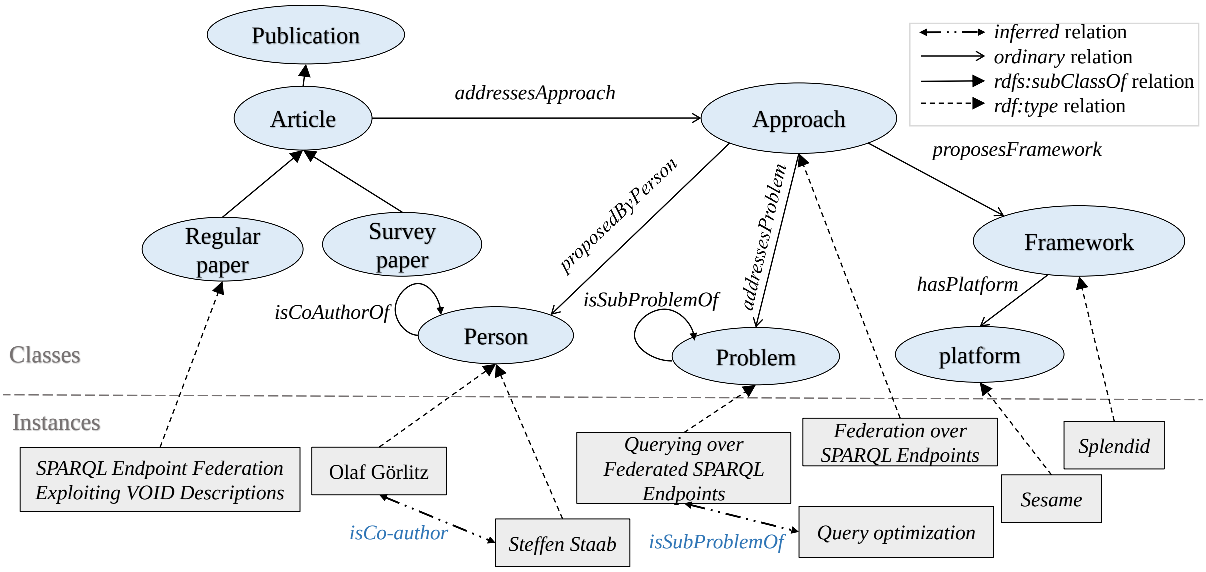


Figure 7.9: **Sample Instances**. Instantiation of key SemSur classes.

Currently, SemSur contains a total of 415 instances of different classes extracted from these survey articles, including `Person` (126), `Publication` (4 survey papers and 40 regular papers covered by the survey papers), `implementations` (32), `approach` (16) and `ResearchProblem` (19). Figure 7.9 presents the core SemSur classes with one instance of each. The dashed arrow indicates the inferred statement that *Olaf Görlitz* and *Steffen Staab* are co-authors of the publication entitled "*SPARQL Endpoint Federation Exploiting VOID Descriptions*" [234].

Use case

SemSur is currently used and maintained in OpenResearch.org, which is a semantic crowd-sourcing platform to collect metadata about several scholarly artifacts, mainly survey papers and scientific events, using ontologies for each such entity type. A sample wiki page[19], representing a scientific article, is added to OpenResearch. On the right-hand side, an information box exists in which the semantic representation of the metadata of the article is presented in accordance with SemSur. Listing 7.3 shows an example[20] of an individual scientific paper created on OpenResearch.org using SemSur.

Evaluation

This section describes the procedure of evaluating SemSur. We followed two strategies: a *satisfaction questionnaire* and an *expert assessment*. We divided the participants who participated in the evaluation (all are computer scientists) into two groups: experts in ontology engineering, who are aware of the challenges in this area, and researchers in other fields of computer science. To ensure the quality of SemSur, it should be able to answer a number of competency queries listed in Table 7.9.

[19] https://www.openresearch.org/wiki/Towards_a_Knowledge_Graph_Representing_Research_Findings_by_Semantifying_Survey_Articles

[20] http://openresearch.org/wiki/ANAPSID:_An_Adaptive_Query_Processing_Engine_for_SPARQL_Endpoints

```
{{Paper
    |Title= ANAPSID: An Adaptive Query Processing Engine for SPARQL Endpoints
    |Authors= Maribel Acosta, Maria-Esther Vidal, Tomas Lampo, Julio Castillo,
    |Series= ISWC                                           |Year= 2011
    |Keywords=Adaptive Query Processing, ANAPSID, Linked Data
    |Abstract=Following the design rules of Linked Data...  |Future work= we plan to...
    |Approach= Querying Distributed RDF Data Sources        |Problem= SPARQL Query Federation
    |Implementation=ANAPSID                                 |Model= Architectural Model
    |PositiveAspects= Decompose the query into simple sub-plans...
    |Challenges= Query Decomposition, Query Optimization... |InfoRepresentation= RDF
    |Methodology= Lightweight wrappers translate SPARQL sub-queries into calls to endpoints...

    |Download-page= https://github.com/anapsid/anapsid      |OS= Linux CentOS
    |Framework= Twisted Network framework                   |GUI= No
    |DocumentationURL= https://github.com/anapsid/anapsid   |ProgLang= Python 2.6.5
    |Motivation= Distribution of RDF datastores             |Benchmark= FedBench
    |Subproblem= query processing on Linked Data            |Version= 1.0
    |RelatedProblem= decompose queries into sub-queries...
}}
```

Listing 7.3: **Paper Description on OpenResearch.org.** Description of a scientific paper and its segments on OpenResearch.org.

Satisfaction questionnaire evaluation. A total of 18 researchers, with Semantic Web background, were recruited for this questionnaire[21]. At the beginning of the evaluation, we made sure that all participants understood the approach by giving them: 1) a presentation about the ontology, the methodology, and the domain, 2) a demonstration of a case study illustrating the potential benefits of using SemSur by trying to answer predefined queries (in Table 7.9), and 3) the results of a set of 20 predefined queries to measure their satisfaction with the results compared to those of other tools. The queries were chosen in increasing order of complexity, to cover SemSur's capabilities.

Expert assessment. Hlomani and Stacey [235] proposed a metric suite for ontology evaluation, including accuracy, adaptability, clarity, cohesion, completeness, conciseness, consistency, and coverage. Each of these metrics is defined in the questionnaire to provide a clear description of it and to avoid ambiguity or misunderstanding. A total of 10 ontology engineering experts were recruited for this questionnaire[22]. Each expert had to give a percentage for each one of these metrics for the ontology.

Observations and Results. All participants used digital libraries such as ACM DL or Web of Science and also used web search engines that index the full text or metadata of scholarly literature, such as Google Scholar, and sometimes using advanced search options and filters. The results retrieved this way were either out of scope for the query or merely related to the keywords, i.e., the search lacked semantics. All subjects unanimously agreed that the current way would not help them unless they explored the respective field more deeply, e.g., by reading survey articles on the topic. From analyzing the responses of the questionnaire participants, we made the following observations:

- Two-thirds of the participants found it difficult to obtain precise information using traditional tools,

[21] https://goo.gl/4SX2Bb
[22] https://goo.gl/8ybDf4

Table 7.9: **Competency questions.** A variable like X is a placeholder for any suitable text.

No.	Query text
Q1	List the publications (title, year, keywords, authors and abstract) financed by organization X.
Q2	List the survey papers addressing problem X and its related problems.
Q3	List the evaluation metrics, information assets, results and benchmarks that are used to evaluate knowledge graph refinement frameworks.
Q4	List the Single-Author publications by a person X in which X proposed approach Y and addressed research problem Z.
Q5	List the implementations for an approach X along with the platform, addressed problems, running OS, and programming language.
Q6	List the co-authors of a person X sharing publications about research topic Y along with their other publications.
Q7	List the experiments design, goal and hypothesis used to evaluate implementations addressing research problem X.
Q8	List the publications that tackle the problem of generating RDF data from existing large quantities of data residing in relational databases.
Q9	List the platforms used to implement approach X and information assets used for evaluation.
Q10	List the motivations, approaches, experiment goal and frameworks addressing a research problem X as well as possible challenges and positive aspects.

- Around 72% of them believe that it is hard to get answers of queries like "*what are the approaches proposed to solve research problems using the traditional tools?*",
- Eight participants pointed out that SemSur would be very helpful for new researchers while ten agreed that SemSur would be *very helpful* for experienced researchers,
- Almost all participants believe that SemSur can help in either the decision of selecting relevant articles for a survey or finding the state-of-the-art approaches to compare their research with.
- About 94% of the participants indicate that the proposed approach saves a lot of time and effort.
- It is worth mentioning that 12 participants are *very happy* to reuse or query SemSur in the future, four are *happy* and only one is *neutral*.

Regarding expert assessment, the highest average percentage by experts is given to the *Clarity* (83%), which reflects how effectively the ontology communicates the intended meaning of the defined terms. The *Consistency*, i.e., are there any conflicts between ontology elements, has the second-largest percentage of 81% which demonstrates that the ontology does not include or allow for both *Conciseness*, and *Accuracy* amount to 78%, which means that the asserted knowledge in the ontology agrees with the expert's knowledge about the domain. The average percentage of Adaptability, Cohesion, Completeness, and Coverage are 67%, 77%, 74%, and 74%, respectively.

Overall, eight participants taking the satisfaction questionnaire are strongly satisfied with SemSur, and eight are satisfied. As anticipated, the evaluation results emphasize the validity of SemSur and show its potentially significant impact on scholarly communication.

Summary

We believe that SemSur breaks new ground in the studies of the transition from document-based to knowledge-based scholarly communication. We have defined an accompanying rule set for discovering new relationships and inferring new knowledge that does not explicitly exist in the knowledge graph. The results of our evaluation show that researchers agree that the traditional way of gathering an overview on a particular research topic is cumbersome and time-consuming. We have created instances of research findings of four survey articles on different fields of research to be able to determine whether SemSur can answer typical information needs and give precise information. As evidence, interviewed domain experts mentioned that it might be necessary to read and understand 30 to 100 scientific articles to get a proper level of understanding or an overview of a topic or sub-topics. As anticipated, SemSur enables successful retrieval of precise and accurate information about particular research aspects, which potentially saves a lot of time and effort compared to traditional ways. The SemSur ontology is considered as a step in a long-term research agenda to create a paradigm shift from document-based to knowledge-based scholarly communication. In conclusion, SemSur can have a significant influence on the scientific community for both new and experienced researchers who want to write a survey or a literature review on a particular research topic.

7.2.4 Ontologies for Natural Sciences

In this section, we present an initiation of developing a set of ontologies for modeling research data in various natural sciences, involving Physics and Pharmaceutical Science. The integration of these ontologies with other vocabularies is also explained in the next subsections. We aim at capturing knowledge inside scientific publications, thus supporting knowledge discovery and reducing the time and effort for the exploration of related content inside publications by linking scientific concepts. The proposed ontologies specifically address high-level abstractions of concepts, such as investigations, theories, methods, experiments, and findings as first-class entities in scholarly communities.

PharmSci Ontology

PharmSci ontology (with the namespace prefix `pharmsci`) is an ontology for describing scientific knowledge in Pharmaceutical sciences publications. Pharmaceutical sciences combine a broad range of scientific disciplines related to drug development, including drug discovery and design, drug analysis, and drug action [236]. PharmSci helps to shape the pharmaceutical science research using a knowledge-based representation. PharmSci has been developed using OWL. PharmSci resources are identified using persistent identifiers (`https://w3id.org/skgo/pharmsci#`); future versions can be collaboratively revised on SKGO Git repository.

The primary intention behind the development of the PharmSci ontology is to support pharmaceutical research exploration and analysis. PharmSci is primarily focusing on modeling the research data required to facilitate experimental and non-experimental clinical studies. We examined these studies with particular attention to all procedures involved in the pharmaceutical research papers. Since the key factor to develop cost-effective and high-quality ontologies is reusing existing ontologies, several concepts have been reused from well-defined ontologies in the pharmaceutical science domain, including:

- Core Ontology of the National Cancer Institute (NCI)[23] – for representing physical entities, which can be stored in digital systems, documents, and events that happen within the publishing process, for example, the acceptance of a paper,
- BioAssay Ontology (BAO) [237] – for representing drugs, and bio-assay components,
- Semanticscience Integrated Ontology (SIO) [238] – for representing objects, processes and their attributes in the biomedical research,
- Friend Of A Friend (FOAF) ontology – for representing agents,
- The Modern Science Ontology (ModSci) [239] – for representing research fields and related discoveries, and
- Geo-spatial (GEO)[24] ontology – for representing representing latitude, longitude and altitude information in the WGS84[25] geodetic reference datum.

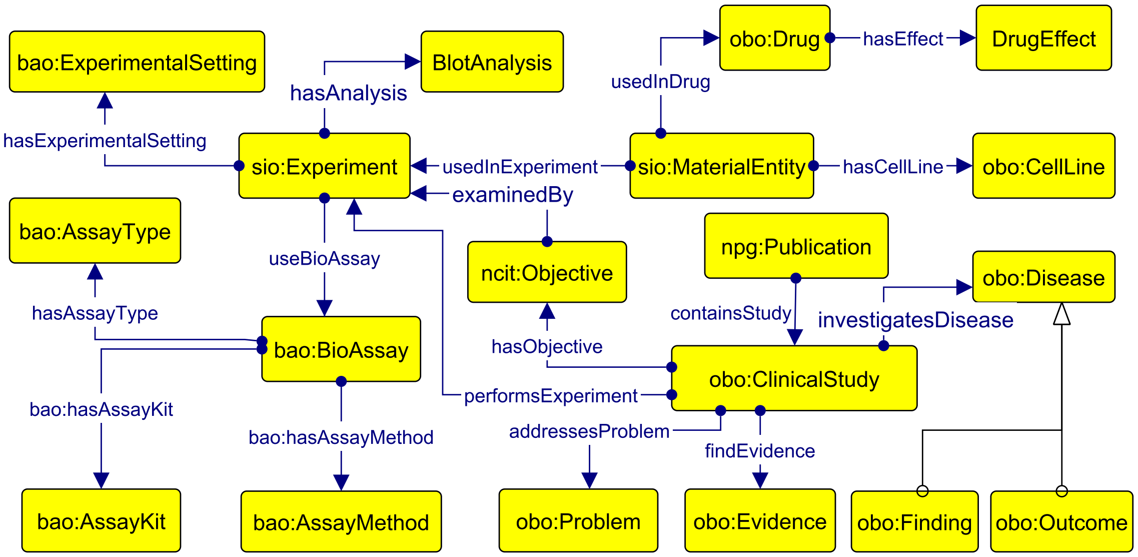


Figure 7.10: **Core Concepts of PharmSci.** A Graffoo diagram illustrating the core concepts of PharmSci and their interlinking relationships. Classes without prefixes are newly defined.

The focal concepts in PharmSci are `Clinical Study`, `Drug`, `Experiment`, `Material`, `DrugEffect`, and `Disease`, as shown in Figure 7.10. Concretely, these entities are represented in PharmSci as `owl:Class`. For instance, the objective of a particular clinical study is represented by connecting `obo:ClinicalStudy` to `ncit:Objective` via the relation `hasObjective`. A notable characteristic of PharmSci is the integration between state-of-the-art ontologies by linking them through the newly defined object properties. For instance, `useBioAssay` property relates `Experiment` and `BioAssay` classes in SIO and BAO ontologies, respectively. Concepts

[23] `http://ns.nature.com/terms/`

[24] `https://www.w3.org/2005/Incubator/geo/XGR-geo/`

[25] `https://www.linz.govt.nz/data/geodetic-system/datums-projections-and-heights/geodetic-datums/world-geodetic-system-1984-wgs84`

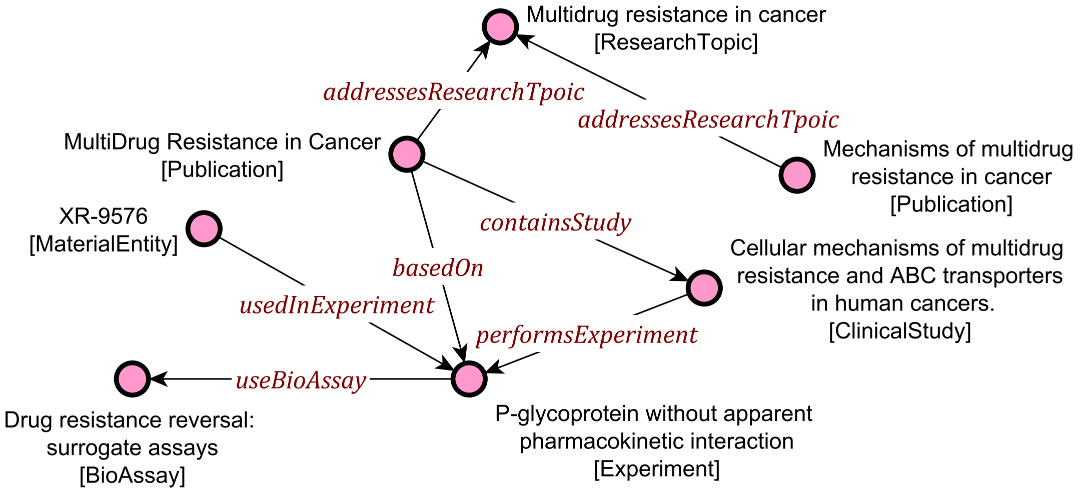


Figure 7.11: **PharmSci instantiation.** A Graffoo diagram illustrating a part of the scientific knowledge presented in [240]. Classes of these instances are placed between square brackets.

and relations have been captured from interviewing researchers in the domain as well as journal publications.

One of the research articles investigated is the one written by Gottesman, Fojo and Bates [240], who propose a clinical study for "*MultiDrug Resistance in Human Cancer*". The scientific knowledge presented in this article can be represented as a knowledge graph, Figure 7.11 depicts a part of it. This type of knowledge representation can help to answer various queries that are of particular interest to many pharmaceutical scientists. Some of these queries are listed below. The SPARQL query corresponding to one of these queries, for example (Q1.3), is shown in Listing 7.4.

(Q1.1) What are the objectives of a clinical study S?

(Q1.2) Which drugs are used in a therapeutic procedure P and a clinical study S for investigating a disease D?

(Q1.3) What are the experiment settings and material entities used in an experiment E contained in a clinical study S?

(Q1.4) What is the type of Blot Analysis used for Experiment E in Clinical Study S?

(Q1.5) What are the publications investigating a disease D with an objective O? and

(Q1.6) What are the assay type, methods, and kits used in experiment E?.

```
SELECT DISTINCT ?expSettings  ?material  ?clinicalstudy
WHERE  {
?clinicalstudy     pharmsci:performsExperiment        ?experiment .
?experiment        pharmsci:hasExperimentalSetting    ?expSettings.
?material          pharmsci:usedInExperiment          ?experiment .
?clinicalstudy     rdfs:label    'Cellular mechanisms of multidrug resistance'
}
```

Listing 7.4: A SPARQL query corresponding to Q1.3.

PhySci Ontology

PhySci ontology (with the namespace prefix `physci`) is an ontology for describing scientific knowledge in Physics publications. It is considered as a step in the direction of constructing an infrastructure that supports the semantic representation of scientific knowledge found in Physics publications, which in turn facilitates the exploration and the reusability of such data. PhySci is available using persistent identifiers (`https://w3id.org/skgo/physci#`). Following best practices, PhySci emphasizes the reuse of state-of-the-art vocabularies (listed below) and the alignment with concepts between them.

- Semantic Web for Earth and Environmental Technology ontology (SWEET) [212] – for representing top-level concepts include math, space, science, physical phenomena, and physical processes,
- The Springer Nature Core Ontology (NPG)[26] – for providing definitions for the fundamental concepts of interest to content publishing,
- The Dublin Core Metadata Element Set [241] – for representing publications metadata, such as abstract, creator, language, date accepted, etc.
- The Semantic Sensor Network Ontology (SSN) [242] – for representing sensors-related concepts, including observations, input, output, and device, etc.
- The Sensor, Observation, Sample, and Actuator (SOSA) Ontology [243] – for representing the interaction between the entities involved in the acts of observation, actuation, and sampling,
- The Modern Science Ontology (ModSci) [239] – for representing research fields and related discoveries,
- The Extensible Observation Ontology (OBOE) [244] – for representing scientific observations and measurements, and
- Friend Of A Friend (FOAF) ontology – for representing agents.

Domain conceptualization has been made by reading publications as well as interviewing researchers in the domain, asking about research methodologies and unfamiliar concepts found in the publications. The core concepts in PhySci are `ResearchWork`, `sosa:Results`, `Theorem`, `sosa:Observation`, `sweet:ScientificModel`, and `Law`. Concretely, these entities are represented in PharmSci as `owl:Class`. As shown in Figure 7.12, PhySci scientific knowledge is linked via object properties, such as `establishesTheorem`, and data type properties, such as `hasPurpose`. For instance, the scientific problem (`physci:ScientificProblem`) addressed by a particular research work (`physci:ResearchWork`) is linked through the relation `addressesProblem`.

One of the research articles investigated is the one that entitled "*Glass processing with pulsed CO_2 laser radiation*", which is written by Weingarten et al. [245]. The scientific knowledge provided in this article can be represented as a knowledge graph, as depicted in Figure 7.13.

A number of queries that are particularly useful for many researchers in the physics research community are listed below. The SPARQL query corresponding to one of these queries, for example (Q2.3), is shown in Listing 7.5. This query can be of particular interest to researchers who are interested in the research topic "Laser ablation".

[26] `https://scigraph.springernature.com/ontologies/core/`

(Q2.1) List published articles that use scientific method X to solve a scientific problem Y,

(Q2.2) List published articles belonging to high energy physics that use a particular scientific model X,

(Q2.3) List the scientific problems that belong to a research topic X along with related publications,

(Q2.4) List theories and scientific laws are used in a research work X that argue the scientific argument Y,

(Q2.5) List the observations reported by publication X,

(Q2.6) List the applications of a research topic X along with the published articles about these applications.

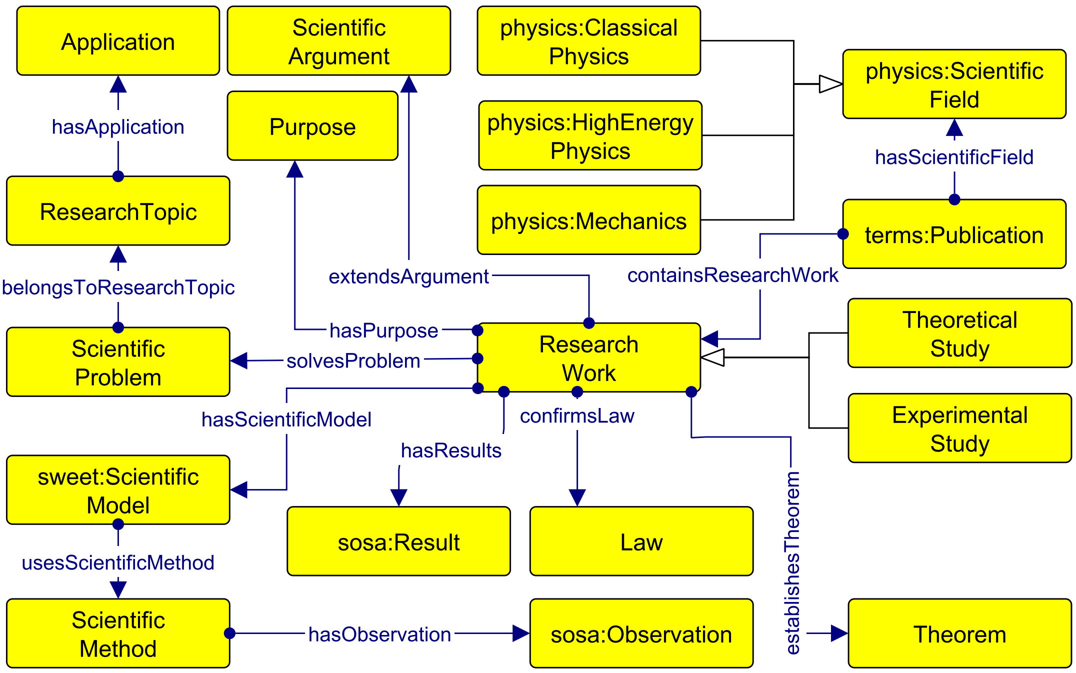


Figure 7.12: **Core Concepts of PhySci.** A Graffoo diagram illustrating the core concepts of PhySci and their interlinking relationships. Classes without prefixes are newly defined.

```
SELECT DISTINCT ?prob ?topic ?work ?pub
WHERE  {
  ?work     physci:solves                       ?prob .
  ?prob     physci:belongsToResearchTopic       ?topic .
  ?pub      physci:containsResearchWork         ?work .
  ?topic    rdfs:label                          'laser ablation'
}
```

Listing 7.5: A SPARQL query corresponding to Q2.3.

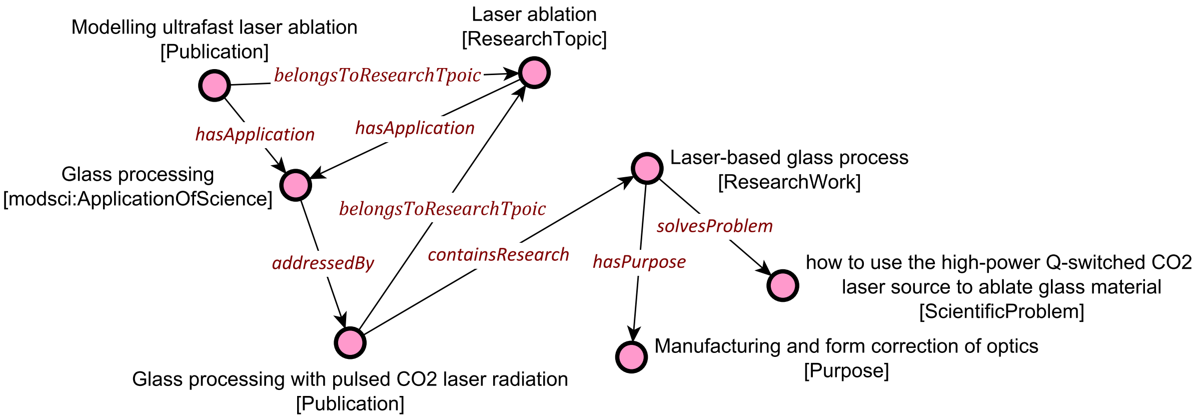

Modelling ultrafast laser ablation
[Publication]
Laser ablation
[ResearchTopic]
belongsToResearchTpoic
hasApplication
hasApplication
Glass processing
[modsci:ApplicationOfScience]
belongsToResearchTpoic
Laser-based glass process
[ResearchWork]
solvesProblem
addressedBy
containsResearch
hasPurpose
how to use the high-power Q-switched CO2
laser source to ablate glass material
[ScientificProblem]
Manufacturing and form correction of optics
[Purpose]
Glass processing with pulsed CO2 laser radiation
[Publication]

Figure 7.13: **PhySci instantiation.** A Graffoo diagram illustrating a part of the scientific knowledge presented in [240]. Classes of these instances are placed between square brackets.

7.3 Exploring Scholarly Knowledge Graphs

Researchers are increasingly exposed to an extraordinary amount of knowledge reported in scientific publications, usually created following a traditional document-oriented workflow. Albeit extensively followed, the document-oriented scholarly communication hinders knowledge extraction and search, as well as reduces connectivity among related publications. Survey articles are a particular type of scholarly publications, in which researchers provide an overview of approaches tackling a particular research problem or area. However, preparing survey articles takes a significant amount of time and effort, because, when following the document-oriented scholarly communication, relationships among approaches are human-produced and presented in various human-readable formats. Consequently, the published research articles are mostly in an unstructured format (e.g., PDF), which does not make them efficiently accessible for evaluations, comparisons, or other analyses.

In this work, we tackle the problem of generating comprehensive overviews of research findings in a structured and comparable way. We present Aurora, a semi-automatic crowd-sourcing approach that captures such information into the *OpenResearch.org* semantic wiki. Aurora provides a structured representation of research articles, using SemSur ontology (see subsection 7.2.3), and brings structure into the knowledge represented inside them in order to enable researchers to explore domain overviews as well as support analysing academia dynamics. It further supports the studying of scholarly literature in a certain research topic, in particular comparing research contributions, which facilitates writing literature surveys and related work sections in regular papers. The observed results suggest that structured representation of research artifacts in Aurora provides better domain overviews for researchers.

7.3.1 Motivating Example: A Producer-Consumer Scenario

In this section, we motivate the problem of automating the generation of the knowledge that is embedded in survey papers. Researchers are increasingly exposed to an incredible amount of information that is created by the traditionally document-oriented workflow. The key problem from the reader's perspective is that the time taken in reviewing the literature to write a survey article, which sometimes takes several years. Most survey articles are published in printed form or as semi-structured digital (e.g., PDF) documents, which do not make their contents efficiently

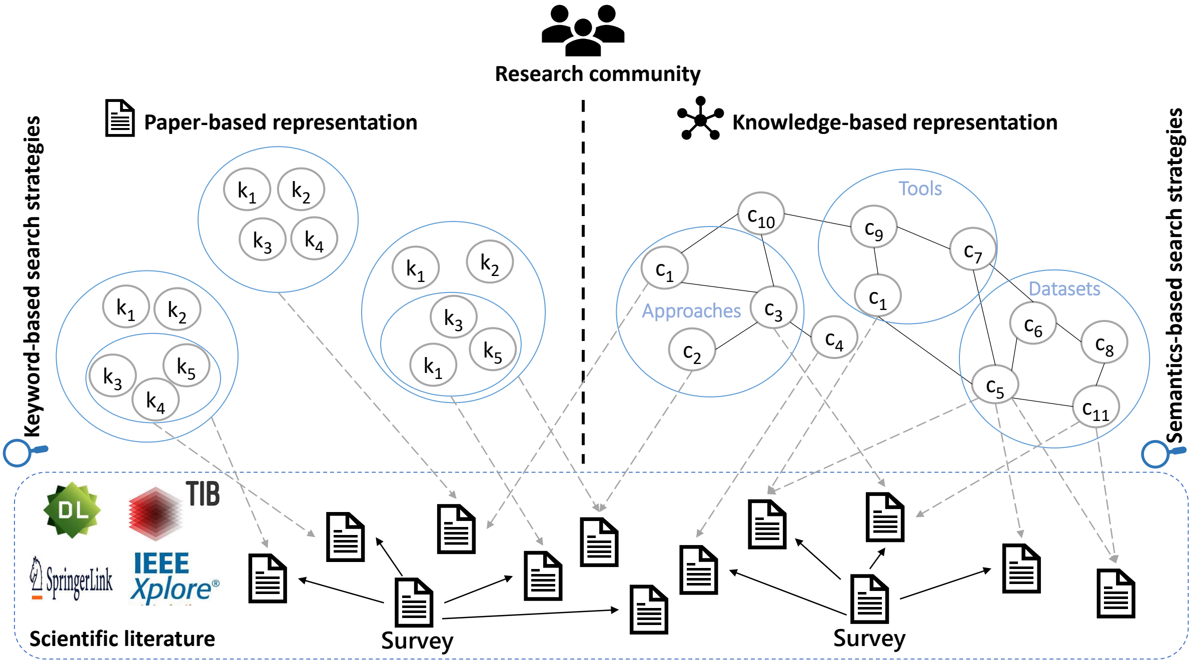

Research community
Paper-based representation
Knowledge-based representation
Keyword-based search strategies
Semantics-based search strategies
k1 k2 k3 k4 k5
c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 c11
Tools
Approaches
Datasets
TIB
DL
SpringerLink
IEEE Xplore
Scientific literature
Survey
Survey

Figure 7.14: **Motivating Example**. A comparison between paper-based (i.e., unstructured) and knowledge-based (i.e., structured) representation of the scientific literature and corresponding search strategies used.

accessible for comparative or other analyses. In this context, the exchange of scholarly knowledge still follows the document-based publishing of results. Instead of reading a large number of documents and exploring further information from each article, it is better for researchers to directly get the requested information in a structured form (more details in subsection 2.2.5). Figure 7.14 depicts a comparison between paper-based (i.e., unstructured) and knowledge-based (i.e., structured) representation of the scientific literature and corresponding search strategies used. In paper-based representation, the scientific contributions are presented in an unstructured form, and a set of keywords (k_i) is used by keyword-based search techniques to find documents that are relevant to users' query. In contrast, in knowledge-based representation, the contents of the scientific papers are represented in a knowledge graph as a set of interconnected concepts (c_i), which are used by semantic search techniques to get more precise information. The latter is used by Aurora to annotate the contents of the scientific papers with the SemSur ontology into the *OpenResearch.org*'s knowledge graph. Details about how Aurora explores this knowledge graph are described in section 6.3.

Let us assume two groups of researchers, one generating an overview of relevant information about the research topic of *Link Discovery* and the other one seeking information to have an overview of particular interest in a research domain. Consider **Alice**, a researcher from the *Data Integration* community, who has little knowledge about *link discovery* and is in need of getting an overview about developments and status of this domain. In contrast, **Axel** is a senior researcher, who, together with three of his students, created a survey paper on this topic entitled *A Survey of Current Link Discovery Frameworks*. At the time of writing our paper, by using the keyword "Link discovery survey" on Google Scholar, Axel's survey paper is the second hit with 71 citations; thus, this is one of the relevant survey papers that Alice would analyze and compare.

However, there are at least ten more survey papers that look relevant, and Alice would fact the challenge of having to read all of them in detail or making an informed selection. On the other hand, it took Axel and his group considerable effort and time to conduct their comprehensive survey on the topic of *link discovery.* This survey paper covers ten different linked discovery tools and performs a functionality-based comparison based on a common set of criteria. Axel's group needed to establish an in-depth undertaking of each framework and come up with a significant list of comparisons. Despite all these efforts, Alice might need a different set of comparisons that requires herself tracing some of the original descriptions of those ten frameworks, yet not gaining a comprehensive overview.

An approach that can automatically generate overviews of the state-of-the-art can allow for the identification of the relevant related work, thus minimizing the effort and time of scholars in providing textual surveys of the topic and maximizing the comprehensiveness of such knowledge for researchers in need of gaining it. Aurora is a framework developed for the purpose of facilitating scholarly communication by generating such overviews. Aurora employs a semantic representation and exploits the wisdom the community has about developments. Thus, it enables comprehensive domain overviews by exploring the underlying knowledge graph.

7.3.2 Architecture

In this section, we present an overview of the architecture of Aurora (Figure 7.15) by explaining the input, workflow and the output.

Input. Aurora receives a set of scholarly articles and annotations provided by the crowd.

Workflow. Appropriate semantic forms have been developed to ease the process of *OpenResearch.org*'s Knowledge graph population. *Semantic Forms* are used for the creation of the instances of the classes. They extend a Semantic MediaWiki[27] in order to allow users to easily create and edit pages that exhibit structured data as so-called info-box-style templates. We used this extension (i.e., *Create Paper*[28]) to create several forms for various categories of the papers metadata, including Bibliographical Metadata, Research Problem, and Evaluation. These forms allow to instantiate the SemSur classes, thus establishes the interlinking between research elements (such as the ones found in the 36 articles used to evaluate in this work). The results of this instantiation correspond to annotations, which are used to create an overview of the paper. Semantic templates, the markup that Semantic MediaWiki [171] introduces through MediaWiki templates [246], are used to specify annotations without the need for researchers to learn the syntax. In addition, semantic templates are used to enable users to edit wiki pages without selecting the right properties or categories. Listing 7.6 shows a snippet of the predefined semantic templates that are used for populating *OpenResearch.org*'s knowledge graph following SemSur ontology.

Output. 1) Aurora populates *OpenResearch.org*'s knowledge graph with the instance data representing the data contained in the provided articles. Traversing and exploring the generated knowledge graph facilitates to analysis and comparison of the existing approaches in a particular domain. 2) Aurora creates wiki pages for each research article, thus linking individual research content, such as publishing venue, authors profile, datasets used in the evaluation, to the corresponding wiki-page. For instance, a wiki page represents the article entitled "*Analysing Scholarly Communication Metadata of Computer Science Events*" written by Fathalla et al.[104]

[27] `https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki`

[28] `https://www.openresearch.org/wiki/Special:FormEdit/Paper#Bibliographical_Metadata`

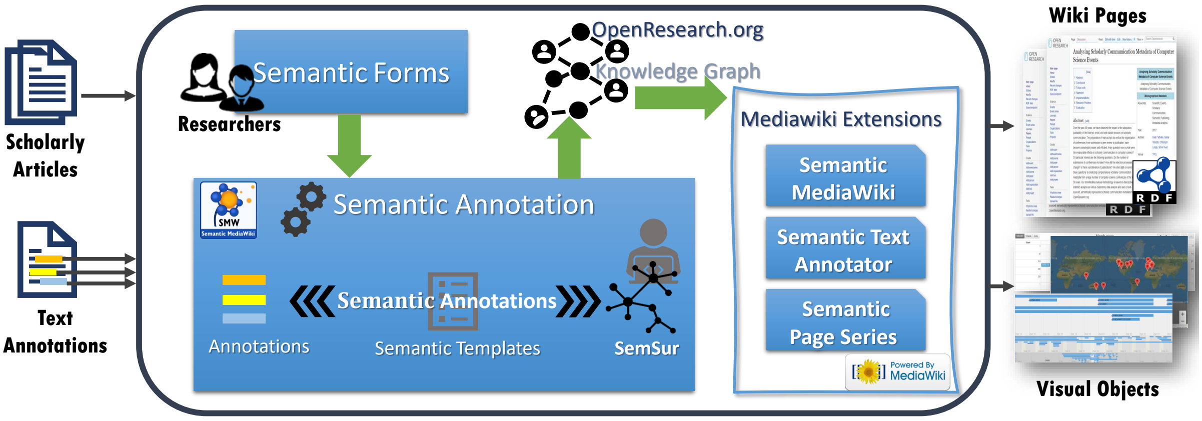


Figure 7.15: **Aurora Architecture**. Aurora receives a set of documents and attributes representing different elements of scientific papers. By extracting and mapping knowledge from documents, an RDF dataset composed of RDF triples is created that describes the scientific documents semantically.

is available[29]. 3) several visual objects, such as timeline, maps and calendar, are used to visualize the data being queried using *Result formats*[30] extension.

```
{{#ifeq:{{{Abstract|}}}|||{{Tablerow|Label=Abstract:|Value=[[has abstract:={{{Abstract}}}]]}}}}
|-
{{#ifeq:{{{Subject|}}}|||{{Tablerow|Label=Subject:|Value=[[has subject:={{{Subject}}}]]}}}}
|-
{{#ifeq:{{{Keywords|}}}|||{{Tablerow|Label=Keywords:|Value=[[has keywords:={{{Keywords}}}]]}}}}
|-
{{#ifeq:{{{Year|}}}|||{{Tablerow|Label=Year:|Value=[[has year:={{{Year}}}]]}}}}
|-
{{#ifeq:{{{Fund|}}}|||{{Tablerow|Label=Fund:|Value=[[has fund:={{{Fund}}}]]}}}}
```

Listing 7.6: A part of the semantic template used for populating *OpenResearch.org*'s knowledge graph.

7.3.3 Implementation

Aurora has been implemented within openresearch.org platform using various Semantic MediaWiki (SMW)[31] extensions. These extensions extend various MediaWiki extensions in order to allow users to easily create, display and edit wiki pages that exhibit structured data as so-called info-box-style templates. A variety of MediaWiki extensions can further extend Semantic MediaWiki with additional functionality; similarly, several extensions of MediaWiki can be appropriate for Semantic MediaWiki. Many Semantic MediaWiki extensions (the complete set is available at SMW website[32]) have been concretely developed for various purposes for Semantic MediaWiki. These extensions are categorized into various categories, including:

- *Adding and modifying data* extensions, such as *AutoFillFormField*, *Semantic Forms*, *HierarchyBuilder* and *Semantic Page Series*,

[29] https://www.openresearch.org/wiki/Analysing_Scholarly_Communication_Metadata_of_Computer_Science_Events
[30] https://www.semantic-mediawiki.org/wiki/Help:Result_formats
[31] https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki
[32] https://www.semantic-mediawiki.org/wiki/Help:Semantic_MediaWiki_extensions

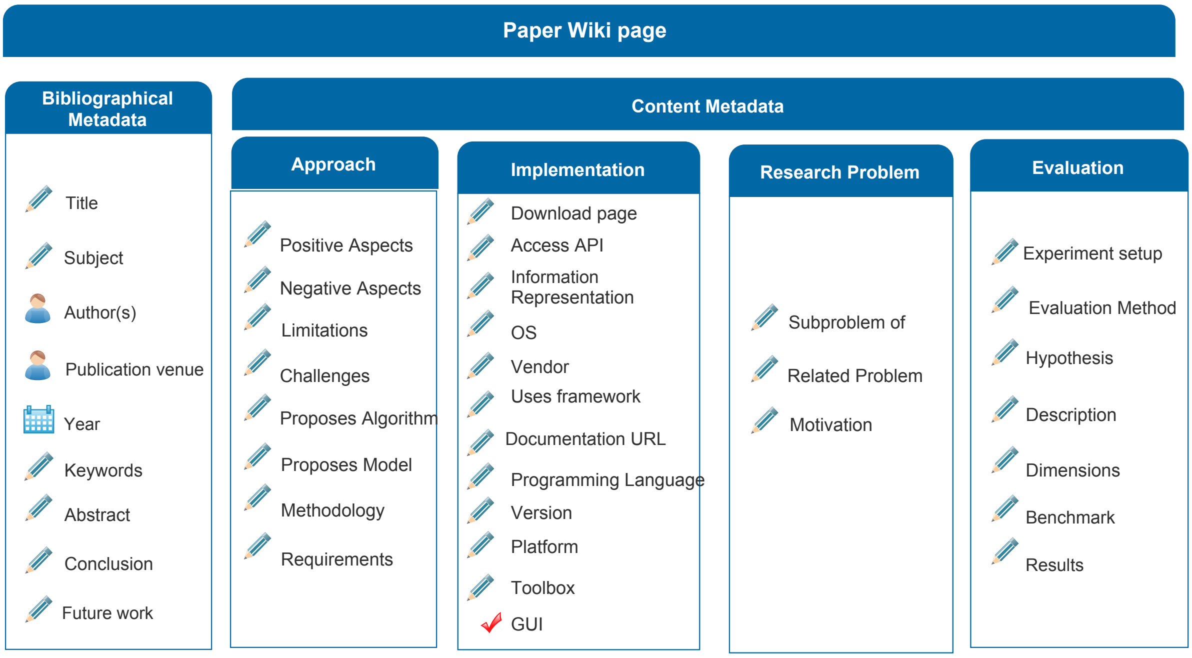


Figure 7.16: **Description Forms.** Used to describe a scientific paper in terms of the main characteristics of the paper approach. Results of this description correspond to annotations, which are used to create an overview of the paper.

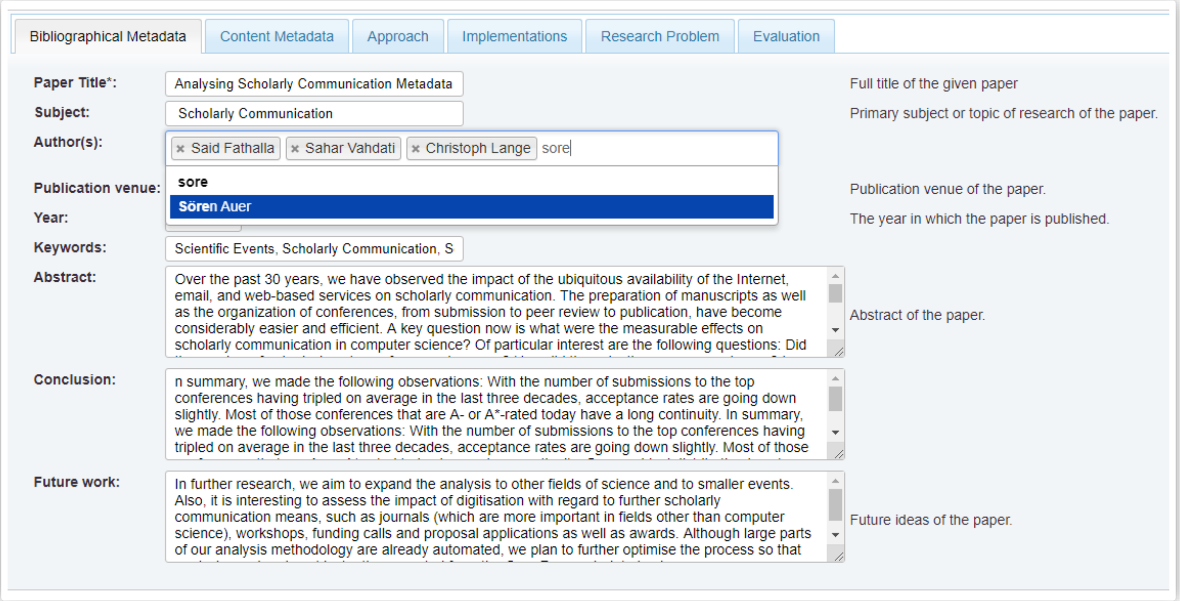
- *Displaying data* extensions, such as *Semantic Result Formats*, *Maps*, *Semantic Drilldown* and *Semantic MediaWiki Graph*,
- *Storing data via an RDF triplestore* extensions, such as *LinkedWiki*, and
- *Importing data* extensions, such as *External Data*.

A brief description about the main extensions used by Aurora is listed below.

Semantic Forms (knows as Page Forms)[33] are used for creating and editing wiki pages, as well as querying data, without any programming experience. A prominent feature of the semantic forms is the auto-completion of the fields of the form, which significantly facilitates avoiding issues of naming ambiguity, spelling mistakes (particularly when referring to other wiki pages), etc. In SMW, semantic forms exists in a separate namespace, "*Form:*". A form enables wiki users to populate a pre-defined set of semantic templates. We employed this extension to create the forms for *Create Paper*[34], which allow for populating SemSur classes and thus establishes the interlinking between research elements (such as the ones found in the 36 articles used for evaluation). Figure 7.16 presents the attributes of research papers involved in each form. The form for entering the bibliographical metadata of the research paper [104] showing the auto-completion feature for linking Sören Auer's to his wiki page is shown in Figure 7.17

Semantic Result Formats[35] is an extension that is used to display query results in various result formats, including tree and Tagcloud for qualitative data, timelines and calendar for time, and D3 charts and Google charts for quantitative data as well as mathematical functions.

[33] https://www.mediawiki.org/wiki/Extension:Page_Forms
[34] http://OpenResearch.org/wiki/Special:FormEdit/Paper
[35] https://www.semantic-mediawiki.org/wiki/Extension:Semantic_Result_Formats



Paper Title*: Analysing Scholarly Communication Metadata — Full title of the given paper

Subject: Scholarly Communication — Primary subject or topic of research of the paper.

Author(s): × Said Fathalla × Sahar Vahdati × Christoph Lange sore

Publication venue: — Publication venue of the paper.

Year: — The year in which the paper is published.

Keywords: Scientific Events, Scholarly Communication, S

Abstract: Over the past 30 years, we have observed the impact of the ubiquitous availability of the Internet, email, and web-based services on scholarly communication. The preparation of manuscripts as well as the organization of conferences, from submission to peer review to publication, have become considerably easier and efficient. A key question now is what were the measurable effects on scholarly communication in computer science? Of particular interest are the following questions: Did — Abstract of the paper.

Conclusion: n summary, we made the following observations: With the number of submissions to the top conferences having tripled on average in the last three decades, acceptance rates are going down slightly. Most of those conferences that are A- or A*-rated today have a long continuity. In summary, we made the following observations: With the number of submissions to the top conferences having tripled on average in the last three decades, acceptance rates are going down slightly. Most of those

Future work: In further research, we aim to expand the analysis to other fields of science and to smaller events. Also, it is interesting to assess the impact of digitisation with regard to further scholarly communication means, such as journals (which are more important in fields other than computer science), workshops, funding calls and proposal applications as well as awards. Although large parts of our analysis methodology are already automated, we plan to further optimise the process so that — Future ideas of the paper.

Figure 7.17: **Semantic Form.** For the research paper [104] showing the paper's Bibliographical Metadata with the auto-completion feature linking to Sören Auer's wiki page.

Besides, this extension can be used to export query results in various formats, such as BibTeX, vCard, and spreadsheet formats. It also can be used in inline queries and other semantic searches.

Semantic Page Series[36] is an extension that is used to generate a series of pages from one Semantic Form. For instance, all research papers have been created using the aforementioned semantic form belong to the same page series, i.e., Paper Series.

Semantic Templates[37] is an extension that is used for handling semantic annotations by including the markup that Semantic MediaWiki introduces through MediaWiki templates. The prominent feature of using templates is that users can specify annotations without learning the syntax. For collecting data about papers, a new semantic template for scientific papers (`https://www.openresearch.org/wiki/Template:Paper`) has been created.

7.3.4 Data-driven Evaluation

In order to obtain sample queries to be implemented by Aurora, five researchers brainstormed (authors of this thesis–three senior and two junior) from the domain of Linked Data and Knowledge Engineering and Data Management. This evaluation aims at illustrating how Aurora enables the execution of complex queries that requires information from different sources about various artifacts. A set of ten predefined natural language queries (listed in Table 7.10) has been finalized to be implemented as *ASK* queries. We chose these queries to cover essential findings presented in the four survey articles. These queries were chosen in increasing order of complexity. To answer such queries, typically, researchers are required to perform a comprehensive literature review on a specific topic by being able to conclude such comparison tables. Figure 7.18 shows a comparison between a table in the survey paper [247] and the corresponding table generated by

[36] `https://www.mediawiki.org/wiki/Extension:Semantic_Page_Series`
[37] `https://www.semantic-mediawiki.org/wiki/Help:Semantic_templates`

Aurora. To demonstrate the efficiency of our proposed approach, we used a set of four survey papers and extracted information of 36 papers surveyed, as a seed in order to create a knowledge graph representing research findings in these papers referenced by these four survey papers. Assuming that a researcher wants to answer these queries using the current scholarly search engines, the question is how much time and effort does it take to answer these queries.

Table 7.10: **List of Questions**. Aurora competency questions for generating research overviews. Questions are presented in increasing order of complexity.

Id	Query in natural language
Q1	List the tools, including programming language, developer, download URL, documentation URL, Data Catalogue and Platform, addressing the problem of "*SPARQL Query Federation*" along with the articles describing them.
Q2	List the resources (either a tool or an ontology) which have been developed for the topic *Semantic web* or *Machine learning*?
Q3	Recommend papers, including paper title, authors, and year, addressing the problem of "*SPARQL Query Federation* " and being published in the ISWC conference.
Q4	Recommend an appropriate venue (with CORE ranking A or B) in which I can publish my work with a research problem of "*Transforming Relational Databases into Semantic Web*"?
Q5	Which evaluation dimensions, evaluation methods, and benchmarks are used to evaluate *"LOD Link Discovery"* tools along with the tools being evaluated and the title of the published article and the year of publication?
Q6	What experiment setups should be considered for evaluating a federated query engine with a brief description of the evaluation procedure and in which tool (including a download link) these setups are used?
Q7	What are the motivations, approaches, challenges, and frameworks for the problem of "*converting relational database to RDF triples*" and its subproblems and related problems?
Q8	List related work for the topic "*SPARQL Query Federation"* or list the papers addressing any of the related research problems of this topic?
Q9	Who is working on "*Semantifying scholarly artifacts*" ? and what papers (with abstract and future work) did these authors publish on this topic?
Q10	List the ontologies representing scholarly artifacts along with their respective creators and maintainers?

Query Implementation and Execution

Semantic MediaWiki includes a query language for semantic search called "*ASK*". This kind of query enables users to directly retrieve certain information from the knowledge graph underlying the wiki pages. The "ASK" API module allows executing queries against Semantic MediaWiki using the MediaWiki API and get results back serialized in one of the supported formats. ASK queries are simpler and more limited than SPARQL queries, but since the SMW knowledge graph can also be exported and loaded into a fully-fledged triple store, this does not pose a significant limitation. We used ASK query API in order to implement the queries. ASK is the query language for semantic search on SMW.

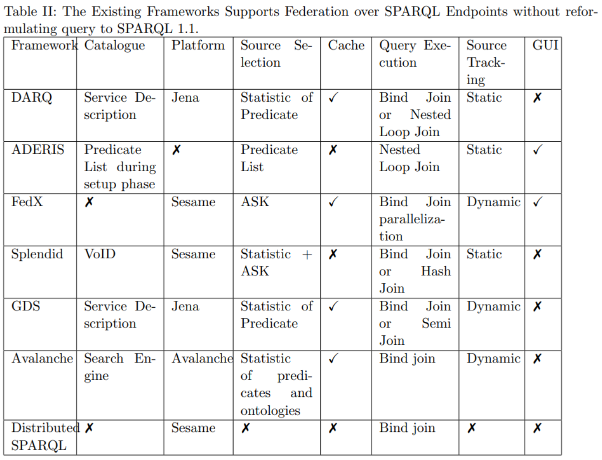
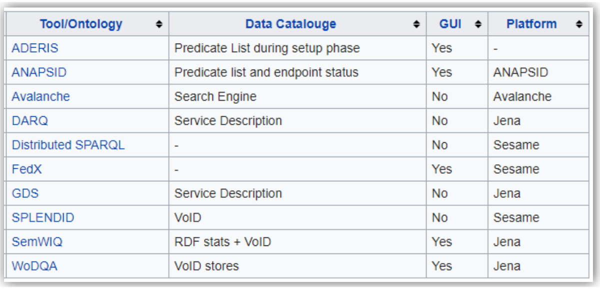


Table II: The Existing Frameworks Supports Federation over SPARQL Endpoints without reformulating query to SPARQL 1.1.

Framework	Catalogue	Platform	Source Selection	Cache	Query Execution	Source Tracking	GUI
DARQ	Service Description	Jena	Statistic of Predicate	✓	Bind Join or Nested Loop Join	Static	✗
ADERIS	Predicate List during setup phase	✗	Predicate List	✗	Nested Loop Join	Static	✓
FedX	✗	Sesame	ASK	✓	Bind Join parallelization	Dynamic	✓
Splendid	VoID	Sesame	Statistic + ASK	✗	Bind Join or Hash Join	Static	✗
GDS	Service Description	Jena	Statistic of Predicate	✓	Bind Join or Semi Join	Dynamic	✗
Avalanche	Search Engine	Avalanche	Statistic of predicates and ontologies	✓	Bind join	Dynamic	✗
Distributed SPARQL	✗	Sesame	✗	✗	Bind join	✗	✗



Tool/Ontology	Data Catalouge	GUI	Platform
ADERIS	Predicate List during setup phase	Yes	-
ANAPSID	Predicate list and endpoint status	Yes	ANAPSID
Avalanche	Search Engine	No	Avalanche
DARQ	Service Description	No	Jena
Distributed SPARQL	-	No	Sesame
FedX	-	Yes	Sesame
GDS	Service Description	No	Jena
SPLENDID	VoID	No	Sesame
SemWIQ	RDF stats + VoID	Yes	Jena
WoDQA	VoID stores	Yes	Jena

Figure 7.18: **Knowledge Extraction**. A table included in the survey paper [247] (left-side) is compared to the table generated (right-side) by querying the knowledge graph. A fine-grained description can be generated.

Getting an overview of a particular topic

We shed light on Q1 and Q2 from Table 7.9 in order to demonstrate how researchers can obtain an overview of a particular topic by querying the OpenResearch.org knowledge graph.

Q1 provides an overview of the tools addressing the "SPARQL Query Federation" research problem. Listing 7.7 shows the ASK query used to answer this query. The results of this query are listed in Figure 7.19. Users can go further by exploring, for example, ISWC to get more details about the *ISWC* conference series, such as acceptance rate, CORE 2017 Rank, and geographical distribution since its beginning.

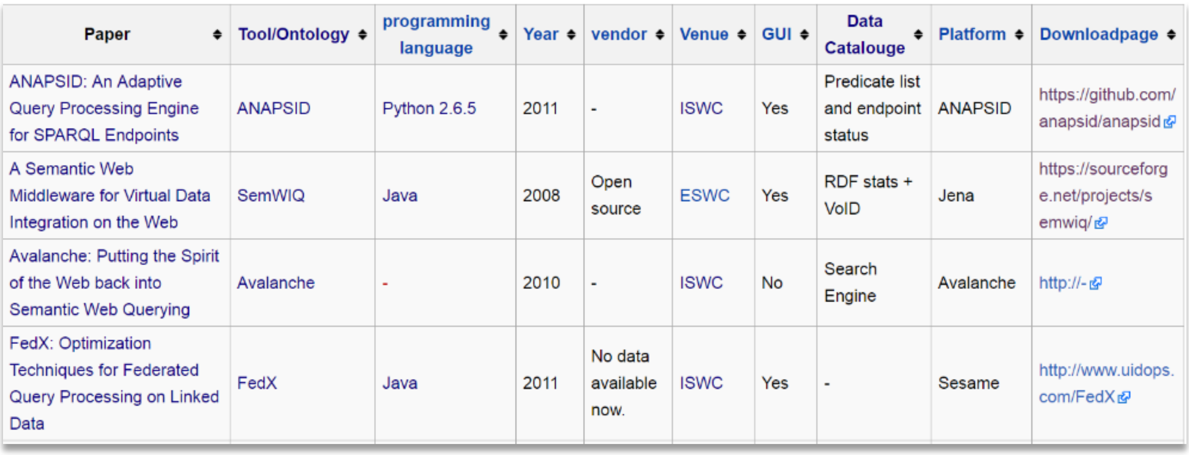
```
{{#ask: [[Category:Paper]] [[Has problem::SPARQL Query Federation]]
| ?Has Implementation = Tool/Ontology
| ?implementedIn ProgLang = programming language
| ?Has year = Year          | ?Has vendor = vendor
| ?Event in series = Venue  | ?Has GUI = GUI
| ?Has DataCatalogue = Data Catalogue
| ?Has platform = Platform  | ?Has Downloadpage = Downloadpage
| mainlabel = Paper         | format = broadtable
| sort = Title
}}
```

Listing 7.7: The ASK query to answer query 1 (Q1).

Q2 provides an overview of the resources (either a tool or an ontology) developed on topics of Semantic Web or machine learning in the last decade. Listing 7.8 shows the ASK query used to answer this query. The results of this query are listed in Figure 7.20.

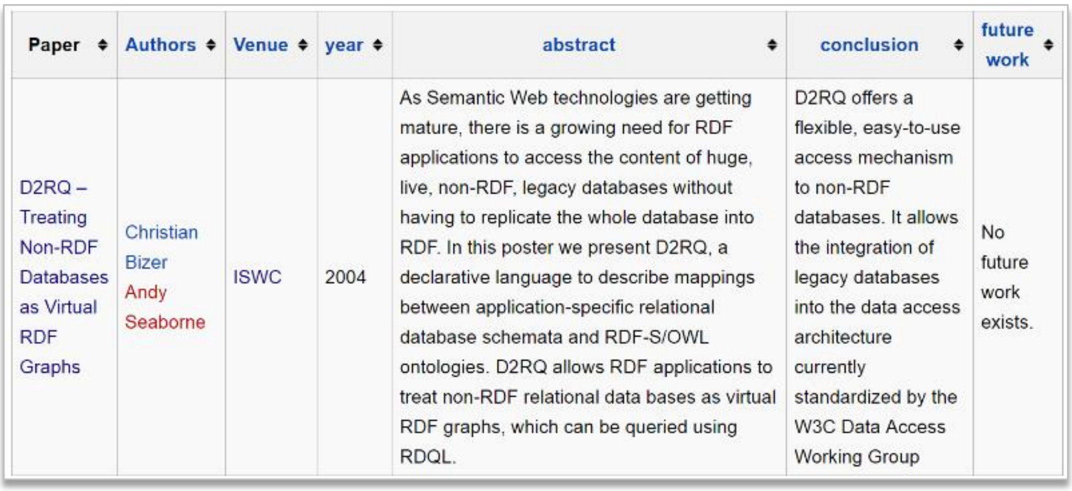
Academic recommendations

We shed light on Q3 and Q4 in Table 7.9 to illustrate how our approach can recommend articles for studying the literature of a certain topic or venue. Q3 intends to recommend papers, including paper title, authors, and publication year, which addresses the problem of "SPARQL Query Federation" and being published in the ISWC conference. Listing 7.9 shows the ASK query used to answer this query. The results of this query are listed in Figure 7.21.



Paper	Tool/Ontology	programming language	Year	vendor	Venue	GUI	Data Catalouge	Platform	Downloadpage
ANAPSID: An Adaptive Query Processing Engine for SPARQL Endpoints	ANAPSID	Python 2.6.5	2011	-	ISWC	Yes	Predicate list and endpoint status	ANAPSID	https://github.com/anapsid/anapsid
A Semantic Web Middleware for Virtual Data Integration on the Web	SemWIQ	Java	2008	Open source	ESWC	Yes	RDF stats + VoID	Jena	https://sourceforge.net/projects/semwiq/
Avalanche: Putting the Spirit of the Web back into Semantic Web Querying	Avalanche	-	2010	-	ISWC	No	Search Engine	Avalanche	http://-
FedX: Optimization Techniques for Federated Query Processing on Linked Data	FedX	Java	2011	No data available now.	ISWC	Yes	-	Sesame	http://www.uidops.com/FedX

Figure 7.19: **Results** of the ASK query (`http://openresearch.org/wiki/Papers_query1`) with fine-grained description of surveyed papers.



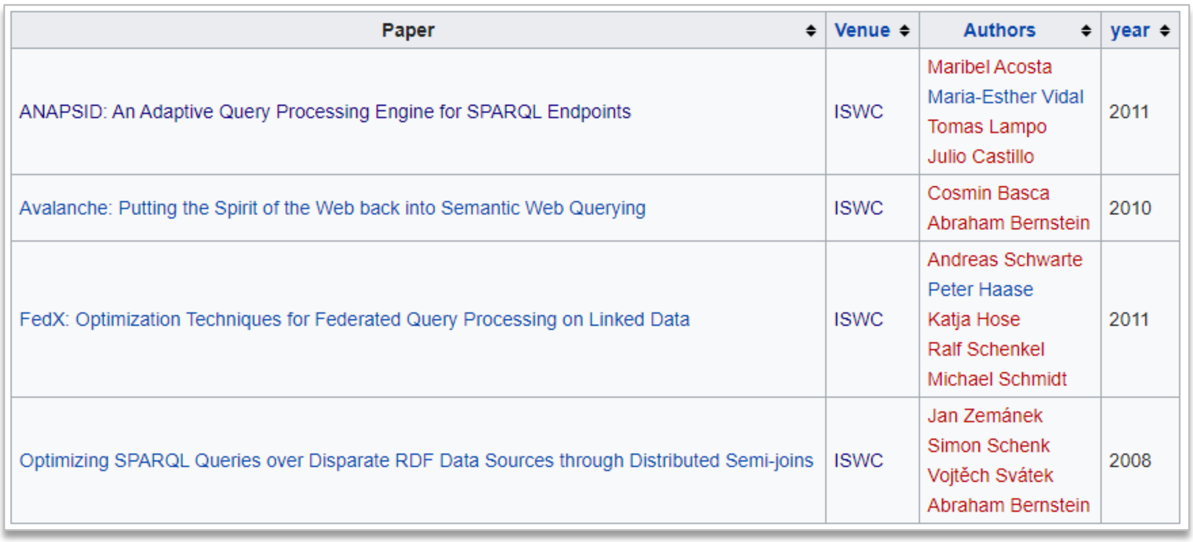
Paper	Authors	Venue	year	abstract	conclusion	future work
D2RQ – Treating Non-RDF Databases as Virtual RDF Graphs	Christian Bizer Andy Seaborne	ISWC	2004	As Semantic Web technologies are getting mature, there is a growing need for RDF applications to access the content of huge, live, non-RDF, legacy databases without having to replicate the whole database into RDF. In this poster we present D2RQ, a declarative language to describe mappings between application-specific relational database schemata and RDF-S/OWL ontologies. D2RQ allows RDF applications to treat non-RDF relational data bases as virtual RDF graphs, which can be queried using RDQL.	D2RQ offers a flexible, easy-to-use access mechanism to non-RDF databases. It allows the integration of legacy databases into the data access architecture currently standardized by the W3C Data Access Working Group	No future work exists.

Figure 7.20: **Results** of the ASK query (`http://openresearch.org/wiki/Papers_query2`) with resources developed in the last decade.

```
{{#ask: [[Category:Tool]] [[Category:Semantic Web]] OR [[Category:Machine learning]]
    [[release date::>{{2008}}-{{January}}-{{01}}]]
    | ?Programming language = developed in
    | ?Category                | ?Maintainer=Contact person
    | ?release date=release Date | format= broadtable
}}
```

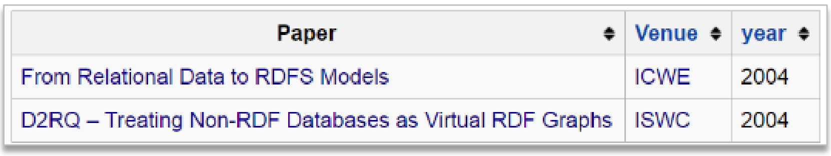
Listing 7.8: The ASK query to answer query 2 (Q2).

Q4 intends to recommend A- or B- ranked conferences that published research papers related to “Transforming Relational Databases into Semantic Web” and corresponding relevant articles. Listing 7.10 shows the ASK query used to answer this query. The results of this query are listed in Figure 7.22. The ranking system of OpenResearch.org is currently based on the CORE2017 and CORE2018 rankings. Note that the event series are selected and filtered for A or B ranked conferences with a subquery indicated by `<q>...</q>`.



Paper	Venue	Authors	year
ANAPSID: An Adaptive Query Processing Engine for SPARQL Endpoints	ISWC	Maribel Acosta Maria-Esther Vidal Tomas Lampo Julio Castillo	2011
Avalanche: Putting the Spirit of the Web back into Semantic Web Querying	ISWC	Cosmin Basca Abraham Bernstein	2010
FedX: Optimization Techniques for Federated Query Processing on Linked Data	ISWC	Andreas Schwarte Peter Haase Katja Hose Ralf Schenkel Michael Schmidt	2011
Optimizing SPARQL Queries over Disparate RDF Data Sources through Distributed Semi-joins	ISWC	Jan Zemánek Simon Schenk Vojtěch Svátek Abraham Bernstein	2008

Figure 7.21: **Result** of the ASK query (`http://openresearch.org/wiki/Papers_query3`). This facilitates the exploration of a paper characteristics.



Paper	Venue	year
From Relational Data to RDFS Models	ICWE	2004
D2RQ – Treating Non-RDF Databases as Virtual RDF Graphs	ISWC	2004

Figure 7.22: **Results** of the ASK query (`http://openresearch.org/wiki/Papers_query4`) over the knowledge graph; they include description of the related surveyed approaches in terms of the venues where they are published.

```
{{#ask: [[Category:Paper]] [[Has problem::SPARQL Query Federation]]  [[Event in series::ISWC]]
| ?Event in series = Venue      | mainlabel = Paper
| ?has authors = Authors      | ?Has year = year
| format = broadtable         | sort = Has problem
}}
```

Listing 7.9: The ASK query to answer query 3 (Q3).

```
{{#ask: [[Category:Paper]]
[[Has problem::Transforming Relational Databases into Semantic Web]]
[[Event in series::<q> [[Category:Event series]] [[Has CORE2017 Rank::A||B]] </q>]]
| ?Event in series = Venue      | ?Has year = year
| mainlabel = Paper             | format = broadtable
| sort = Title
}}
```

Listing 7.10: The ASK query to answer query 4 (Q4).

Usability Testing

Usability testing is a non-functional testing technique that is used in user-centered interaction design to measure how easy the system is by letting real users use the system [193]. The objective is to optimize the usability of Aurora. In this evaluation, a total of 19 end-users (from various

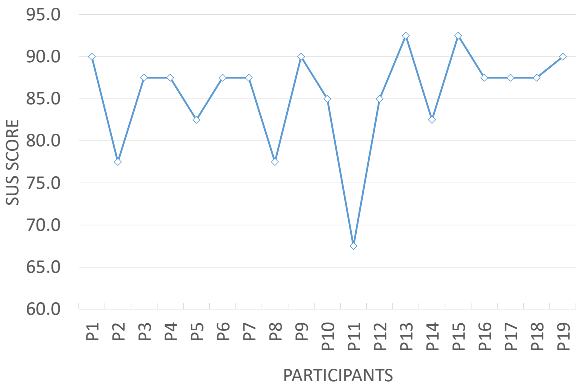


Figure 7.23: **SUS scores** for each participant involved in the test.

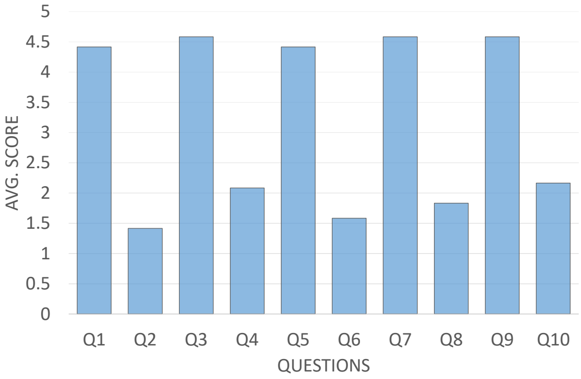


Figure 7.24: **Average scores** for each question in the SUS questionnaire.

fields and professions, such as computer scientists, dentists, and engineers) were recruited in order to test and assess the usability of Aurora. Participants were explicitly asked to fill in the System Usability Scale (SUS) questionnaire [194] (available at `https://forms.gle/LisXbKdmug3fgkBAA`). SUS is a standard usability test that comprises ten questions with five response options each, ranging from strongly agree (score of 5) to strongly disagree (score of 1). We promoted the usability study on several mailing lists and between colleagues to acquire participants as many as possible. *Results.* For the analysis of results, the SUS scoring method [195] is used. The average SUS score falls into seven adjective ratings, ranging from *Best Imaginable* (above 90.9) to *Worst Imaginable* (below 12.5) [196]. Most strikingly, findings showed that Aurora scored a high average SUS score of 83.8, i.e., *Excellent*, which corresponds to grade `A`, thus achieving exceptional usability. As shown in Figure 7.23, the SUS scores for all participants fall into the range of 68–93, which is relatively high. Figure 7.24 presents the average score for each question in the SUS questionnaire. It is worth noting that the questions in SUS questionnaire are designed in a way so that odd-numbered questions measure positive agreement while even-numbered ones measure negative agreement. Notably, the average scores for odd-numbered questions are relatively high, while the ones for even-numbered questions are relatively low, which reflects a high agreement with the usability of Aurora.

7.3.5 Summary

We presented Aurora for representing research findings in computer science in a semantic way and crowd-sourcing the creation of these semantic representations using a semantic wiki. We evaluated the approach with a number of competency questions, which can now be answered using Aurora and simple queries, instead of long-term survey compilation. It serves not only as a means for representing research findings but also as a new way of how survey articles can be created. We are aware that this work can be one of the initial steps in the broader research and development plan for transforming the traditional document-based information flows in scholarly communication into knowledge-based ones, e.g., using ontologies and knowledge graphs for representation. In particular, the work needs to be advanced along three avenues: 1) Concerning the representation of scholarly communication in knowledge graphs, we need to cater for other disciplines, which follow different methodologies. Aurora follows the problem-approach-implementation-evaluation methodological pattern, which is widespread in computer science and other engineering fields. 2) Scholarly communication with respect to a particular field is initially often based on fuzzy conceptualizations, which gradually evolve into more formal, solid ones. Also, scholarly discourse involves disagreements, discussions, and controversies, which need to be represented, e.g., using argumentation ontologies. 3) Moreover, to successfully master the transition and elicit a critical mass of crowd-sourced contributions, we need to develop means to contribute with minimal effort and provide instant benefits for contributors.

CHAPTER 8

Conclusion and Future Work

In this thesis, we investigated the problem of supporting scholarly communication by means of scholarly data management, analysis, and retrieval.

- *Towards facilitating scholarly data management*, an ontology (OR-SEO) for describing scholarly events metadata and their related entities has been developed as well as the Science Knowledge Graph Ontologies (SKGO) suite. This suite captures the knowledge of scientific information typically presented in publications in various fields of science, including Computer Science (SemSur), Physics (PhySci), and Pharmaceutical science (PharmSci). Besides, the scholarly events dataset (EVENTSKG), which contains historical data about the publications, submissions, start date, end date, location, and homepage for the most renowned event series belonging to eight Computer Science communities, have been published. EVENTSKG enables sharing machine-readable interlinked data on the Web, which enable data publishers to link their data to linked open data sources to provide context.
- *Towards facilitating scholarly data analysis*, we devise a lightweight methodology (SEMA) and a metrics suite (SEQA) for facilitating scholarly data curation and analysis, respectively. Specifically, SEQA aims at assessing the quality and identifying the characteristics of high-quality scholarly events within scientific communities, while SEMA targets facilitating building knowledge graphs of scholarly events with the purpose of identifying the characteristics of renowned events and providing a recommendation to various stakeholders in the scholarly communication domain.
- *Towards facilitating scholarly data retrieval*, we developed a front-end service (SPARQL-AG) for SPARQL query generation and execution to be used by researchers belonging to all computer science communities, i.e., not restricted to whom aware of Semantic Web technologies. SPARQL-AG queries several data sources when generating queries, including EVENTSKG, OR-SEO, and DBpedia.

In chapter 1, we presented the research problem and the challenges to be addressed. Preliminary information about the terminologies, data models, tools, and technologies used in this work presented in this thesis are provided in chapter 2. An overview of the related work is presented in chapter 3. In chapter 4, we investigated the problem of study the various characteristics of scholarly events in different fields of science to assess their impact. In chapter 5, we tackled the problem of semantically represent scholarly events metadata by developing the Scientific Events Ontology as well as a Linked Open Dataset (EVENTSKG). Therefore, we provide a comprehensive semantic description of scientific events in four fields of science as well as a historical semantic description of renowned events in eight computer science fields over the past five decades. chapter 6 presents a front-end (SPARQL-AG) that automatically generates and executes SPARQL queries for querying EVENTSKG. In chapter 7, we proposed a set of ontologies for modeling the research findings in various fields of science, resulting in a knowledge graph of the scientific findings in modern sciences. Finally, the thesis is concluded by revisiting the research questions in section 8.1. Section 8.2 presents the future work that can expand the work presented in this thesis.

8.1 Revisiting the Research Questions

The research problem guiding the work of this thesis can be expressed by the question (i.e., the main research question): How can scholarly data be understood by machines making data representation, data analytics, and information retrieval more efficient? Along the road to answering this question, four challenges have been faced: 1) Exploring the characteristics of the renowned scholarly events in four fields of science, 2) Integrating heterogeneous scholarly events metadata, 3) Representing knowledge about entities involved in the scholarly events domain, and 4) Representing scientific knowledge in various fields of science using semantic technologies. In order to conduct the work of this thesis, the main research question is divided into four research questions. In this section, we revisit these research questions, which were raised in chapter 1.

RQ1: How can the characteristics of renowned scholarly events in different fields of science be utilized to assess their impact?

This research question is addressed in chapter 4, in which we investigate the problem of studying the various characteristics of scholarly events in different fields of science to assess their impact. We proposed the Scholarly Events Quality Assessment metrics suite (SEQA), which contains ten metrics, with different granularity levels, that can be used for assessing the quality of scholarly events. This suite enables the analytical study of the evolution of key characteristics of the renowned scholarly events in the respective communities. In addition, we present a novel methodology, i.e., Scholarly Events Metadata Analysis Methodology (SEMA). SEMA is a reproducible build methodology for building knowledge graphs of scholarly events with the purpose of identifying the characteristics of renowned events and providing a recommendation to various stakeholders in the scholarly communication domain. Consequently, we applied the proposed methodology, i.e., SEQA, and conducted a study for identifying the various characteristics of scholarly events in four fields of science, namely Computer Science, Physics, Engineering, and Mathematics. We analyzed the metadata of approximately 2,000 events that took place in the last five decades. Out of this study, we recorded the most noteworthy findings, and a set of recommendations has been concluded to different stakeholders, involving event organizers, potential authors, and sponsors. An exploratory data analysis is performed, aiming at inferring facts and figures about renowned scholarly events since their beginning. Overall, this work helps to shed light on the evolving and different publishing practices in various communities and helps to identify novel ways for scholarly communication, such as the blurring of journals and conferences or open-access overlay-journals as they already started to emerge.

RQ2: How can we represent and integrate heterogeneous scholarly event metadata in knowledge graphs to facilitate scholarly data management and retrieval?

This research question is addressed in chapter 5 and chapter 6. In these chapters, we developed a Linked Open Dataset (EVENTSKG), which offers a comprehensive semantic description of scholarly events of 73+ renowned Computer Science events over the last five decades as well as a front-end that automatically generates and executes SPARQL queries for querying EVENTSKG. EVENTSKG contains historical data about the publications, submissions, start date, end date, location, and homepage for the most renowned event series belonging to eight Computer Science communities. We published three releases of EVENTSKG. The latest release contains metadata of approximately 2,000 events belonging to eight Computer Science communities; 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). EVENTSKG stores data relevant to these events in RDF, and each event's metadata is described appropriately through employing the data and object properties in the Scientific Events Ontology (RQ3). We believe that EVENTSKG can bridge the gap between stakeholders involved in the scholarly events life cycle, starting from event establishment through paper submission till proceedings publishing, including event's organizers, potential authors, publishers, and sponsors. Besides, sharing and reusing scholarly datasets have become a new form of scholarly communication. Concerning facilitating scholarly retrieval, SPARQL-AG offers the research community a familiar and intuitive interface for querying scholarly data without the need to know neither the schema of the knowledge graph being queried nor the SPARQL syntax. The former can widen the accessibility of semantic-based representation of scholarly data to researchers who do not know Semantic Web technologies. To the best of our knowledge, this is the first web-based service that allows end-users to create and execute both simple and complex SPARQL queries over scholarly knowledge bases.

RQ3: How can ontologies represent semantics encoded in entities involved in scholarly events and relationships among them?

This research question is addressed in chapter 5, in which we tackle the problem of representing scientific events metadata semantically, i.e., integrating existing events vocabularies and making explicit the relationships and interconnections between event data. Accordingly, we support the transformation from a “Web of documents” into a “Web of data” in the scientific domain and making it easier to efficiently query and process the data. We engineered the Scientific Events Ontology (OR-SEO), which enables a semantically enriched representation of scholarly event metadata, interlinked with other datasets and knowledge graphs. OR-SEO is in use as the schema of *OpenResearch.org*, which contains thousands of event wiki pages, linked by several entities such as event series, location, etc. OR-SEO emphasizes the reuse of events-related vocabularies, the alignment with concepts between them, as well as the design and visualization patterns. OR-SEO is maintained and used by (but not limited to) Said Fathalla, an editor of OR, to represent metadata of scientific events so far, mainly in Computer Science but also some other fields, including physics and chemistry.

RQ4: How can published research in various fields of science be understood by machines, making information retrieval, analysis, and scholarly data management more efficient?

This research question is addressed in chapter 7, in which we investigate the problem of providing a semantic representation for scientific knowledge belonging to various fields in modern science, namely Computer Science, Physics and Pharmaceutical Science, as well as modeling various branches in modern science along with related concepts, such as Phenomena, Scientific Discoveries, Instruments, etc. First, we outline the limitations of the traditional representation, i.e., PDF format, of scientific knowledge. Then, we present our vision of enabling researchers to work on a standard knowledge base comprising comprehensive descriptions of their research, thus making research contributions transparent and comparable. In the direction to this goal, we propose the Science Knowledge Graph Ontologies (SKGO) suite, which comprises four OWL ontologies for representing the scientific knowledge in various fields of science, including Computer Science (SemSur), Physics (PhySci), and Pharmaceutical science (PharmSci) as well as an upper ontology on top of them called Modern Science Ontology (ModSci). The development methodology can be easily applied to other fields of science. Finally, we affirm the applicability of developing a knowledge graph-based approach for exploring the Knowledge Graph of Science using a Semantic Mediawiki platform. We reveal the advantages of utilizing knowledge graphs for semantically describing and interlinking the scientific knowledge belonging to various branches in modern science.

8.2 Future Work

There are numerous directions for extending the work conducted in this thesis. In this section, we outline the possible directions that can be considered for the further improvement of the various contributions offered by this thesis.

Concerning facilitating scholarly data management, we list the following directions:

- Elaborating on the set of features, such as acceptance rate, h-index, and organizers' reputation (can be identified in terms of their h-index and i10-index) that can be used to efficiently compare events in the same community. Besides, adding more events attributes to EVENTSKG, such as hosting university or organization, sponsors, and event steering committees or program committee chairs as well as adding events from other fields of science, including Physics, Mathematics, and Engineering, in addition to events belonging to other Computer Science communities, such as computer vision, data management, and computational learning.
- Adopting 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.
- Automating SEMA subtasks; i.e., data cleansing and completion, data structure transformation, and name unification using NLP techniques.
- Developing machine learning techniques for anomaly detection[1], and Named-entity recognition (seeks to locate and classify named entities, such as location, important dates, etc., in events websites).

[1] Anomaly detection is referred to the identification of data values or events that do not conform to an expected pattern or to other items in the dataset

- Adopting OR-SEO ontology to cover authors, affiliations, titles, and keywords as well as to model event evolution considering property changes such as type, e.g., from a symposium to a conference, or events re-scheduled, or events whose chairs changed.
- Developing a smart data analytics tool in order to assess events' progress and recommend relevant events to potential authors.
- Implementing an algorithm, inspired by the one proposed in [8], to identify relationships between individual research elements, e.g., benchmarks, and between authors, i.e., co-authorship.
- Regarding the formal representation of the scientific process and its entities, we aim at aligning ModSci's own model with existing formal models of a science whose processes and structures have already been investigated in depth, i.e., Mathematics.
- Refining the formal representation of science in SKGO, covering further fields of science, such as Earth sciences, Biology, as well as Mathematics, by dedicated ontologies, and realizing services on top of them.
- To boost real-world applications of SKGO, we are planning to realize knowledge management and e-research services on top of the Open Research Knowledge Graph (ORKG), into which we will integrate the SKGO ontologies.

Concerning facilitating scholarly data analysis, we list the following directions:

- Assessing the impact of digitization regarding further scholarly communication means, such as journals (which are more important in fields other than computer science), workshops, funding calls, and proposal applications as well as awards.
- Optimizing the process so that analysis can be almost instantly generated from the OpenResearch.org data basis.
- Extending the analysis to other fields of science and applying more metrics such as author and paper affiliation analysis, sponsorship and co-authorship analysis, and awards.
- SEQA metrics can be used in providing new and innovative venue rankings for different research fields, thus allowing in particular younger researchers without a long-term experience to identify better publication strategies and consequently contribute more productively to the advancement of research.
- Systematically investigate review quality in the future for assessing the peer review process of scholarly events.
- Multidisciplinary data harvesting services, for example, metadata of OpenAIRE [248] project[2], are planned to be used in future work.

Concerning facilitating scholarly visualization and retrieval, we list the following directions for improving SPARQL-AG and Aurora:

- Regarding data visualization, enabling instantaneously generated interactive charts based on the results retrieved by the SPARQL-AG endpoints.
- Regarding expressiveness, we are planning to cover almost all SPARQL 1.1 features, in particular, subqueries, multidimensional queries, nested aggregations for rich analytics, and graphs as results for CONSTRUCT queries and updates, since not all specifications are covered.

[2] http://openaire.eu

- Regarding interface robustness, we are planning to adopt the interface to let the user select the knowledge base to be queried and improve the interface based on participants' feedback.
- Developing a service to automatically provide, with minimal effort, an overview of a particular research topic. For example, a survey of related work could be automatically generated once a researcher described his contribution using a form-based widget integrated into a submission system.

Bibliography

[1] T. Berners-Lee and R. Cailliau, *WorldWideWeb: Proposal for a HyperText Project*, (1990), URL: `http://www.w3.org/Proposal` (cit. on p. 1).

[2] S. Lawrence and C. L. Giles, *Searching the world wide web*, Science **280**.5360 (1998) 98 (cit. on p. 1).

[3] T. Berners-Lee, J. Hendler and O. Lassila, *The semantic web*, Scientific american **284**.5 (2001) 34 (cit. on p. 1).

[4] T. Heath and C. Bizer, *Linked data: Evolving the web into a global data space*, Synthesis lectures on the semantic web: theory and technology **1**.1 (2011) 1 (cit. on pp. 1, 20, 97, 116, 119).

[5] S. M. Fathalla, Y. F. Hassan and M. El-Sayed, "A hybrid method for user query reformation and classification", *Computer Theory and Applications (ICCTA), 2012 22nd International Conference on*, IEEE, 2012 132 (cit. on p. 2).

[6] S. Fathalla, S. Vahdati, S. Auer and C. Lange, "Towards a Knowledge Graph Representing Research Findings by Semantifying Survey Articles", *International Conference on Theory and Practice of Digital Libraries*, Springer, 2017 315 (cit. on pp. 2, 35, 46, 137, 150).

[7] S. Fathalla, S. Vahdati, S. Auer and C. Lange, *SemSur: a core ontology for the semantic representation of research findings*, Procedia Computer Science **137** (2018) 151 (cit. on pp. 2, 11).

[8] S. Fathalla and Y. Kannot, "A Bidirectional-Based Spreading Activation Method for Human Diseases Relatedness Detection Using Disease Ontology", *Conference on Computational Collective Intelligence Technologies and Applications*, Springer, 2017 14 (cit. on pp. 2, 183).

[9] S. Fathalla, *Detecting Human Diseases Relatedness: A Spreading Activation Approach Over Ontologies*, International Journal on Semantic Web and Information Systems (IJSWIS) **14**.3 (2018) 120 (cit. on pp. 2, 25, 156).

[10] T. Yu, J. Li, Q. Yu, Y. Tian, X. Shun, L. Xu, L. Zhu and H. Gao, *Knowledge graph for TCM health preservation: Design, construction, and applications*, Artificial intelligence in medicine **77** (2017) 48 (cit. on p. 2).

[11] S. Sundararajan and S. V. Nitta, *Designing engaging intelligent tutoring systems in an age of cognitive computing*, IBM Journal of Research and Development **59**.6 (2015) 10 (cit. on p. 2).

[12] M.-C. Müller, "Semantic author name disambiguation with word embeddings", *International Conference on Theory and Practice of Digital Libraries*, Springer, 2017 300 (cit. on p. 2).

[13] Z. Wu, J. Wu, M. Khabsa, K. Williams, H.-H. Chen, W. Huang, S. Tuarob, S. R. Choudhury, A. Ororbia, P. Mitra et al., "Towards building a scholarly big data platform: Challenges, lessons and opportunities", *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, IEEE Press, 2014 117 (cit. on p. 2).

[14] S. Auer, V. Kovtun, M. Prinz, A. Kasprzik, M. Stocker and M. E. Vidal, "Towards a knowledge graph for science", *8th International Conference on Web Intelligence, Mining and Semantics*, ACM, 2018 (cit. on pp. 2, 96, 136).

[15] S. Peroni, "The semantic publishing and referencing ontologies", *Semantic web technologies and legal scholarly publishing*, Springer, 2014 (cit. on pp. 3, 46, 96, 99).

[16] N. A. Rakhmawati, J. Umbrich, M. Karnstedt, A. Hasnain and M. Hausenblas, *Querying over Federated SPARQL Endpoints - A State of the Art Survey.*, CoRR **abs/1306.1723** (2013) (cit. on pp. 5, 157).

[17] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne et al., *The FAIR Guiding Principles for scientific data management and stewardship*, Scientific data **3** (2016) (cit. on pp. 5, 19, 138, 141, 142).

[18] D. Castelli, P. Manghi and C. Thanos, *A vision towards scientific communication infrastructures*, International Journal on Digital Libraries **13**.3-4 (2013) 155 (cit. on p. 5).

[19] S. Khan, X. Liu, K. A. Shakil and M. Alam, *A survey on scholarly data: From big data perspective*, Information Processing & Management **53**.4 (2017) 923 (cit. on pp. 5, 18).

[20] T. J. Ameloot, M. Marx, W. Martens, F. Neven and J. van Wees, *30 Years of PODS in facts and figures*, SIGMOD Record **40**.3 (2011) (cit. on pp. 5, 37, 39, 40).

[21] D. Aumüller and E. Rahm, *Affiliation analysis of database publications*, SIGMOD Record **40**.1 (2011) (cit. on pp. 5, 37, 38, 40).

[22] D. Hiemstra, C. Hauff, F. De Jong and W. Kraaij, "SIGIR's 30th anniversary: an analysis of trends in IR research and the topology of its community", *ACM SIGIR Forum*, vol. 41, 2, ACM, 2007 18 (cit. on pp. 5, 35, 37, 40, 41).

[23] M. A. Nascimento, J. Sander and J. Pound, *Analysis of SIGMOD's co-authorship graph*, ACM Sigmod record **32**.3 (2003) 8 (cit. on pp. 5, 37, 38, 40).

[24] S. Agarwal, N. Mittal and A. Sureka, *A scientometric analysis of 9 ACM SIGWEB cooperating conferences*, ACM SIGWEB Newsletter Autumn (2016) 6 (cit. on pp. 5, 37).

[25] S. D. J. Barbosa, M. S. Silveira and I. Gasparini, *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 (2017) 275 (cit. on pp. 5, 35, 37, 39, 40).

[26] S. Jeong and H.-G. Kim,
Intellectual structure of biomedical informatics reflected in scholarly events,
Scientometrics **85**.2 (2010) 541 (cit. on pp. 5, 37).

[27] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil and D. Barton,
Big data: the management revolution, Harvard business review **90**.10 (2012)
(cit. on p. 6).

[28] D. Shotton, *Semantic publishing: the coming revolution in scientific journal publishing*,
Learned Publishing **22**.2 (2009) 85 (cit. on pp. 6, 15).

[29] S. Fathalla and C. Lange, "EVENTS: a dataset on the history of top-prestigious events in five computer science communities", *Semantics, Analytics, Visualization*,
Springer, 2017 110 (cit. on p. 10).

[30] A. Gyrard, M. Serrano and G. A. Atemezing, "Semantic web methodologies, best practices and ontology engineering applied to Internet of Things",
Internet of Things (WF-IoT), 2015 IEEE 2nd World Forum on, IEEE, 2015
(cit. on pp. 10, 117).

[31] S. Fathalla, S. Vahdati, S. Auer and C. Lange,
"The scientific events ontology of the openresearch. org curation platform",
Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, 2019
(cit. on p. 10).

[32] S. Fathalla, S. Vahdati, C. Lange and S. Auer, "SEO: A Scientific Events Data Model",
International Semantic Web Conference, Springer, 2019 In Press (cit. on p. 10).

[33] A. S. C. Committee et al.,
Scholarly Communication: Principles and strategies for the reform of scholarly communication: Issues related to the formal system of scholarly communication,
College & Research Libraries News **64**.8 (2019) 526 (cit. on p. 15).

[34] J.-C. Guédon, B. Kramer, M. Laakso, B. Schmidt, E. Šimukovič, J. Hansen, R. Kiley, A. Kitson, W. van der Stelt, K. Markram et al., *Future of scholarly publishing and scholarly communication: Report of the Expert Group to the European Commission*,
Report of the Expert Group to the European Commission (2019) (cit. on pp. 15, 17, 18).

[35] J. Priem, *Scholarship: Beyond the paper*, Nature **495**.7442 (2013) 437 (cit. on p. 15).

[36] A. S. C. Committee et al.,
Principles and Strategies for the reform of Scholarly Communication,
American Library Association (2006), URL:
`http://www.ala.org/acrl/publications/whitepapers/principlesstrategies`
(cit. on p. 15).

[37] M. Świgoń, *Knowledge sharing practices in informal scholarly communication amongst academics in Poland*,
Malaysian Journal of Library & Information Science **22**.2 (2017) 101 (cit. on p. 15).

[38] M. Ware and M. Mabe,
The STM report: An overview of scientific and scholarly journal publishing, (2015)
(cit. on p. 15).

[39] K. Shearer and W. F. Birdsall, *A researcher's research agenda for scholarly communication in Canada*, New Review of Information Networking **11**.1 (2005) 99 (cit. on p. 16).

[40] S. Sawant, *Transformation of the scholarly communication cycle*, Library Hi Tech News **29**.10 (2012) 21 (cit. on p. 16).

[41] M. Ernst, "Choosing a venue: Conference or journal", *Retireved from https://homes. cs. washington. edu/˜ mernst/advice/conferences-vs-journals. html*, 2006 (cit. on p. 17).

[42] *What is peer-review?*, Online; accessed 15 November 2019, URL: `https://www.elsevier.com/reviewers/what-is-peer-review` (cit. on p. 17).

[43] D. R. De Vries, E. A. Marschall and R. A. Stein, *Exploring the peer review process: what is it, does it work, and can it be improved?*, Fisheries **34**.6 (2009) 270 (cit. on p. 17).

[44] D. Starovoytova, *Scientific Research, Writing, and Dissemination (Part 4/4): Dissemination of Scholarly Publications*, Journal of Education and Practice (USA) (2017) 2222 (cit. on p. 17).

[45] J. Arora, *Digital preservation: an overview*, (2009) (cit. on p. 18).

[46] B. Houghton, *Preservation challenges in the digital age*, D-lib magazine **22**.7/8 (2016) 1 (cit. on p. 18).

[47] A. Freitas and E. Curry, "Big data curation", *New horizons for a data-driven economy*, Springer, Cham, 2016 87 (cit. on p. 18).

[48] M. Khabsa and C. L. Giles, *The number of scholarly documents on the public web*, PloS one **9**.5 (2014) e93949 (cit. on p. 18).

[49] D. Shotton, *Semantic publishing: the coming revolution in scientific journal publishing*, Learned Publishing **22**.2 (2009) 85 (cit. on p. 18).

[50] D. Shotton, K. Portwin, G. Klyne and A. Miles, *Adventures in semantic publishing: exemplar semantic enhancements of a research article*, PLoS computational biology **5**.4 (2009) e1000361 (cit. on p. 18).

[51] R. Johnson, A. Watkinson and M. Mabe, *The STM Report: An overview of scientific and scholarly publishing*, International Association of Scientific, Technical and Medical Publishers (2018) (cit. on p. 18).

[52] *FAIR Principles*, Online; accessed 30 October 2019, URL: `https://www.go-fair.org/fair-principles/` (cit. on p. 19).

[53] K. M. Albert, *Open access: implications for scholarly publishing and medical libraries*, Journal of the Medical Library Association **94**.3 (2006) 253 (cit. on p. 19).

[54] S. Harnad, T. Brody, F. Vallières, L. Carr, S. Hitchcock, Y. Gingras, C. Oppenheim, C. Hajjem and E. R. Hilf, *The access/impact problem and the green and gold roads to open access*, Serials review **30**.4 (2004) 310 (cit. on p. 19).

[55] J. Beall, *Predatory publishers are corrupting open access*, Nature News **489**.7415 (2012) 179 (cit. on p. 19).

[56] W. W. W. Consortium et al., *Semantic Web, 2001*, (2001), Online; accessed 11 November 2019, URL: `https://www.w3.org/standards/semanticweb` (cit. on p. 20).

[57] T. Berners-Lee, *Linked data-design issues*, (2006), Online; accessed 31 October 2019, URL: `http://www.w3.org/DesignIssues/LinkedData.html` (cit. on p. 20).

[58] C. Bizer, T. Heath and T. Berners-Lee, "Linked data: The story so far", *Semantic services, interoperability and web applications: emerging concepts*, IGI Global, 2011 205 (cit. on p. 20).

[59] L. Masinter, T. Berners-Lee and R. T. Fielding, *Uniform resource identifier (URI): Generic syntax*, (2005), URL: `https://tools.ietf.org/html/rfc3986` (cit. on p. 20).

[60] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach and T. Berners-Lee, *Hypertext transfer protocol–HTTP/1.1*, Retrieved 31 October 2019, 1999, URL: `http://www.w3.org/Protocols/rfc2616/rfc2616.html` (cit. on pp. 20, 21).

[61] C. Bizer and R. Cyganiak, *Quality-driven information filtering using the WIQA policy framework*, J. Web Sem. **7**.1 (2009) (cit. on p. 21).

[62] T. Berners-Lee, *Is your linked open data 5 star*, BERNERS-LEE, T. Linked Data. Cambridge: W3C (2010) (cit. on pp. 21, 113, 116).

[63] T. Berners-Lee, *5-star deployment scheme*, Retrieved 31 October 2019, 2012, URL: `http://5stardata.info/en` (cit. on p. 21).

[64] E. Miller, *An introduction to the resource description framework*, Bulletin of the American Society for Information Science and Technology **25**.1 (1998) 15 (cit. on p. 22).

[65] G. Antoniou and F. v. Harmelen, *A Semantic Web Primer, 2Nd Edition (Cooperative Information Systems)*, 2nd ed., The MIT Press, 2008 (cit. on p. 22).

[66] M. Arenas, C. Gutierrez and J. Pérez, "Foundations of RDF databases", *Reasoning Web International Summer School*, Springer, 2009 158, URL: `https://doi.org/10.1007/978-3-642-03754-2_4` (cit. on p. 22).

[67] R. Cyganiak, D. Wood, M. Lanthaler, G. Klyne, J. J. Carroll and B. McBride, *RDF 1.1 concepts and abstract syntax*, W3C recommendation (2014), [Online; accessed 22-October-2019], URL: `https://www.w3.org/TR/rdf11-concepts/` (cit. on p. 22).

[68] A. Zimmermann, *RDF 1.1 concepts and abstract syntax*, W3C recommendation (2014), [Online; accessed 22-October-2019], URL: `https://www.w3.org/TR/rdf11-datasets/` (cit. on p. 22).

[69] D. Beckett, *Rdf 1.1 n-triples*, W3C recommendation (2014), [Online; accessed 28-October-2019], URL: `https://www.w3.org/TR/n-triples/` (cit. on p. 24).

[70] D. Beckett, T. Berners-Lee, E. Prud'hommeaux and G. Carothers, *RDF 1.1 Turtle*, World Wide Web Consortium (2014), [Online; accessed 29-October-2019], URL: `https://www.w3.org/TR/turtle/` (cit. on p. 24).

[71] P. J. Landin, *The mechanical evaluation of expressions*, The computer journal **6**.4 (1964) 308 (cit. on p. 24).

[72] M. Sporny, D. Longley, G. Kellogg, M. Lanthaler and N. Lindström, *JSON-LD 1.0*, W3C recommendation (2014), [Online; accessed 29-October-2019], URL: https://www.w3.org/TR/json-ld11/ (cit. on p. 25).

[73] G. Antoniou and F. Van Harmelen, *A semantic web primer*, MIT press, 2004 (cit. on pp. 25, 27, 28, 157).

[74] P. M. Simons, *Ontology in metaphysics*, Encyclopædia Britannica, inc. Encyclopædia Britannica, 2019, URL: https://www.britannica.com/topic/ontology-metaphysics (cit. on p. 25).

[75] T. R. Gruber, *A translation approach to portable ontology specifications*, Knowledge acquisition **5**.2 (1993) 199 (cit. on p. 25).

[76] N. Guarino, D. Oberle and S. Staab, “What is an ontology?”, *Handbook on ontologies*, Springer, 2009 1 (cit. on p. 25).

[77] D. Oberle, *How ontologies benefit enterprise applications*, Semantic Web **5**.6 (2014) 473 (cit. on p. 26).

[78] N.-W. Chi, K.-Y. Lin and S.-H. Hsieh, *Using ontology-based text classification to assist job hazard analysis*, Advanced Engineering Informatics **28**.4 (2014) 381 (cit. on p. 26).

[79] S. Jeong and H.-G. Kim, *SEDE: An ontology for scholarly event description*, Journal of Information Science **36**.2 (2010) 209 (cit. on pp. 26, 51).

[80] S. Vahdati, N. Arndt, S. Auer and C. Lange, “OpenResearch: Collaborative Management of Scholarly Communication Metadata”, *EKAW*, Springer, 2016 (cit. on p. 26).

[81] R. Stevens, C. A. Goble and S. Bechhofer, *Ontology-based knowledge representation for bioinformatics*, Briefings in bioinformatics **1**.4 (2000) 398 (cit. on p. 26).

[82] D. Brickley, R. V. Guha and B. McBride, *RDF Schema 1.1*, W3C recommendation **25** (2014) (cit. on p. 27).

[83] W. O. W. Group, *OWL 2 Web Ontology Language*, (2012), [Online; accessed 16-November-2019], URL: https://www.w3.org/TR/owl2-overview/ (cit. on p. 28).

[84] H. Wu and A. Yamaguchi, *Semantic Web technologies for the big data in life sciences*, Bioscience trends **8**.4 (2014) 192 (cit. on p. 28).

[85] S. Ronzhin, E. Folmer, P. Maria, M. Brattinga, W. Beek, R. Lemmens and R. van’t Veer, *Kadaster Knowledge Graph: Beyond the Fifth Star of Open Data*, Information **10**.10 (2019) 310 (cit. on p. 29).

[86] H. Paulheim, *Knowledge graph refinement: A survey of approaches and evaluation methods*, Semantic web **8**.3 (2017) 489 (cit. on pp. 30, 157).

[87] K. Bollacker, C. Evans, P. Paritosh, T. Sturge and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge", *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, AcM, 2008 1247 (cit. on p. 30).

[88] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak and S. Hellmann, *DBpedia-A crystallization point for the Web of Data*, Web Semantics: science, services and agents on the world wide web **7**.3 (2009) 154 (cit. on p. 30).

[89] D. Vrandečić and M. Krötzsch, *Wikidata: a free collaborative knowledge base*, (2014) (cit. on p. 30).

[90] J. Hoffart, F. M. Suchanek, K. Berberich and G. Weikum, *YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia*, Artificial Intelligence **194** (2013) 28 (cit. on p. 30).

[91] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault and J. Morissette, *Bio2RDF: towards a mashup to build bioinformatics knowledge systems*, Journal of biomedical informatics **41**.5 (2008) 706 (cit. on p. 30).

[92] A. Singhal, *Introducing the knowledge graph: things, not strings*, Official google blog **5** (2012) (cit. on p. 30).

[93] G. A. Miller, *WordNet: An electronic lexical database*, MIT press, 1998 (cit. on p. 30).

[94] F. M. Suchanek, G. Kasneci and G. Weikum, "Yago: a core of semantic knowledge", *Proceedings of the 16th international conference on World Wide Web*, ACM, 2007 697 (cit. on p. 30).

[95] T. Rebele, F. Suchanek, J. Hoffart, J. Biega, E. Kuzey and G. Weikum, "YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames", *International Semantic Web Conference*, Springer, 2016 177 (cit. on p. 30).

[96] S. Harris and A. Seaborne, *SPARQL 1.1 Query Language*, W3C Recommendation, World Wide Web Consortium (W3C), 2013, URL: `http://www.w3.org/TR/sparql11-query/` (visited on 26/10/2019) (cit. on pp. 31, 33).

[97] R. Cyganiak, *A relational algebra for SPARQL*, Digital Media Systems Laboratory HP Laboratories Bristol. HPL-2005-170 **35** (2005) 9 (cit. on p. 31).

[98] J. Pérez, M. Arenas and C. Gutiérrez, *Semantics and complexity of SPARQL*, ACM Trans. Database Syst. **34**.3 (2009) 16:1, URL: `http://doi.acm.org/10.1145/1567274.1567278` (cit. on p. 32).

[99] S. Agarwal, N. Mittal and A. Sureka, *A glance at seven acm sigweb series of conferences*, ACM SIGWEB Newsletter Summer (2016) 5 (cit. on pp. 35, 38, 40).

[100] M. Biryukov and C. Dong, "Analysis of computer science communities based on DBLP", Springer, 2010 (cit. on pp. 35, 38).

[101] F. Osborne, E. Motta and P. Mulholland, "Exploring scholarly data with Rexplore", *ISWC*, Springer, 2013 (cit. on p. 35).

[102] B. González-Pereira, V. P. Guerrero-Bote and F. Moya-Anegón, *A new approach to the metric of journals' scientific prestige: The SJR indicator*, Journal of informetrics **4**.3 (2010) 379 (cit. on pp. 37, 61).

[103] S. Fathalla and C. Lange, "EVENTSKG: a knowledge graph representation for top-prestigious computer science events metadata", *International Conference on Computational Collective Intelligence*, Springer, 2018 53 (cit. on pp. 37, 124).

[104] S. Fathalla, S. Vahdati, C. Lange and S. Auer, "Analysing scholarly communication metadata of computer science events", *TPDL*, Springer, 2017 (cit. on pp. 37, 168, 170, 171).

[105] F. M. Wolf, *Meta-analysis: Quantitative methods for research synthesis*, vol. 59, Sage, 1986 (cit. on p. 38).

[106] L. V. Hedges, *Advances in statistical methods for meta-analysis*, New Directions for Evaluation **1984**.24 (1984) 25 (cit. on p. 38).

[107] G. Guilera, M. Barrios and J. Gómez-Benito, *Meta-analysis in psychology: a bibliometric study*, Scientometrics **94**.3 (2013) 943 (cit. on p. 38).

[108] H. M. A. El-Din, A. S. Eldin and A. M. Hanora, *Bibliometric analysis of Egyptian publications on Hepatitis C virus from PubMed using data mining of an in-house developed database (HCVDBegy)*, Scientometrics **108**.2 (2016) 895 (cit. on p. 38).

[109] V. Bakare and G. Lewison, *Country over-citation ratios*, Scientometrics **113**.2 (2017) 1199 (cit. on p. 38).

[110] S. Yan and D. Lee, "Toward alternative measures for ranking venues: a case of database research community", *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, ACM, 2007 235 (cit. on p. 38).

[111] C. Bartneck and J. Hu, "Scientometric analysis of the CHI proceedings", *Proceedings of the SIGCHI conference on human factors in computing systems*, ACM, 2009 699 (cit. on pp. 39, 40).

[112] L. Barkhuus and J. A. Rode, "From mice to men-24 years of evaluation in CHI", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2007 1 (cit. on pp. 39, 40).

[113] S. Greenberg and B. Buxton, "Usability evaluation considered harmful (some of the time)", *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 2008 111 (cit. on pp. 39, 40).

[114] Y. Liu, J. Goncalves, D. Ferreira, B. Xiao, S. Hosio and V. Kostakos, "CHI 1994-2013: mapping two decades of intellectual progress through co-word analysis", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2014 3553 (cit. on pp. 39, 40).

[115] M. P. Rombach, M. A. Porter, J. H. Fowler and P. J. Mucha,
Core-periphery structure in networks,
SIAM Journal on Applied mathematics **74**.1 (2014) 167 (cit. on p. 39).

[116] I. Gasparini, S. D. J. Barbosa, M. S. Silveira and F. C. de Mendonça,
"Self-knowledge: reflecting on the influence of IHC publications on its own event",
Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems,
ACM, 2016 20 (cit. on p. 40).

[117] A. Correia, H. Paredes and B. Fonseca,
Scientometric analysis of scientific publications in CSCW,
Scientometrics **114**.1 (2018) 31 (cit. on pp. 40, 41).

[118] J. Wainer and C. Barsottini, *Empirical research in CSCW—a review of the ACM/CSCW conferences from 1998 to 2004*,
Journal of the Brazilian Computer Society **13**.3 (2007) 27 (cit. on p. 40).

[119] A. Kienle and M. Wessner,
The CSCL community in its first decade: development, continuity, connectivity,
International Journal of Computer-Supported Collaborative Learning **1**.1 (2006) 9
(cit. on p. 40).

[120] C. Bartneck,
The end of the beginning: a reflection on the first five years of the HRI conference,
Scientometrics **86**.2 (2010) 487 (cit. on p. 40).

[121] D. M. Nichols and S. J. Cunningham,
"A scientometric analysis of 15 years of CHINZ conferences", *CHINZ 2015*, ACM, 2015
73 (cit. on p. 40).

[122] A. F. Smeaton, G. Keogh, C. Gurrin, K. McDonald and T. Sødring,
"Analysis of papers from twenty-five years of SIGIR conferences: what have we been doing for the last quarter of a century?", *ACM SIGIR Forum*, vol. 37, 1, ACM, 2003 49
(cit. on pp. 40, 41).

[123] A. G. Nuzzolese, A. L. Gentile, V. Presutti and A. Gangemi,
"Semantic web conference ontology-a refactoring solution",
International Semantic Web Conference, Springer, 2016 (cit. on pp. 42, 99).

[124] K. Möller, T. Heath, S. Handschuh and J. Domingue,
"Recipes for semantic web dog food—The ESWC and ISWC metadata projects",
The Semantic Web, Springer, 2007 (cit. on pp. 42, 50).

[125] B. Vasilescu, A. Serebrenik and T. Mens,
"A historical dataset of software engineering conferences",
10th Working Conference on Mining Software Repositories, IEEE Press, 2013
(cit. on p. 42).

[126] D. Luo and K. Lyons, *CASCONet: A Conference dataset*, arXiv (2017) (cit. on p. 42).

[127] S. Gottschalk and E. Demidova,
"EventKG: A Multilingual Event-Centric Temporal Knowledge Graph",
Extended Semantic Web Conference (ESWC 2018), Springer, 2018 (cit. on p. 43).

[128] W. R. Van Hage, V. Malaisé, R. Segers, L. Hollink and G. Schreiber, *Design and use of the Simple Event Model (SEM)*, Web Semantics: Science, Services and Agents on the World Wide Web **9**.2 (2011) (cit. on pp. 43, 49).

[129] M. Färber, "The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data", *International Semantic Web Conference*, Springer, 2019 113 (cit. on p. 43).

[130] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. P. Hsu and K. Wang, "An overview of microsoft academic service (mas) and applications", *Proceedings of the 24th international conference on world wide web*, ACM, 2015 243 (cit. on p. 44).

[131] C. Deliot, N. Wilson, L. Costabello and P.-Y. Vandenbussche, "The British national bibliography: who uses our linked data?", *International Conference on Dublin Core and Metadata Applications*, 2017 24 (cit. on p. 44).

[132] A. Di Iorio, C. Lange, A. Dimou and S. Vahdati, "Semantic publishing challenge–assessing the quality of scientific output by information extraction and interlinking", *Semantic Web Evaluation Challenge*, Springer, 2015 65 (cit. on p. 44).

[133] L. B. Sollaci and M. G. Pereira, *The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey*, Journal of the medical library association **92**.3 (2004) 364 (cit. on p. 45).

[134] S. Peroni and D. Shotton, "The SPAR ontologies", *International Semantic Web Conference*, Springer, 2018 119 (cit. on pp. 46, 48, 99).

[135] S. Peroni and D. Shotton, *FaBiO and CiTO: ontologies for describing bibliographic resources and citations*, Web Semantics: Science, Services and Agents on the World Wide Web **17** (2012) 33 (cit. on pp. 46–48).

[136] A. Di Iorio, A. G. Nuzzolese, S. Peroni, D. M. Shotton and F. Vitali, "Describing bibliographic references in RDF.", *SePublica*, 2014 (cit. on p. 47).

[137] A. Constantin, S. Peroni, S. Pettifer, D. Shotton and F. Vitali, *The document components ontology (DoCO)*, Semantic Web **7**.2 (2016) 167 (cit. on p. 47).

[138] S. Peroni, D. Shotton and F. Vitali, "Scholarly publishing and linked data: describing roles, statuses, temporal and contextual extents", *Proceedings of the 8th International Conference on Semantic Systems*, ACM, 2012 9 (cit. on p. 47).

[139] A. Gangemi, S. Peroni, D. Shotton and F. Vitali, "A pattern-based ontology for describing publishing workflows", *Proceedings of the 5th International Conference on Ontology and Semantic Web Patterns-Volume 1302*, CEUR-WS. org, 2014 2 (cit. on p. 47).

[140] F. Osborne, S. Peroni and E. Motta,
"Clustering citation distributions for semantic categorization and citation prediction",
Proceedings of the 4th International Conference on Linked Science-Volume 1282,
CEUR-WS. org, 2014 24 (cit. on p. 47).

[141] D.-L. Magazine,
The five stars of online journal articles—a framework for article evaluation,
D-Lib Magazine **18**.1/2 (2012) (cit. on p. 47).

[142] Y. Sure, S. Bloehdorn, P. Haase, J. Hartmann and D. Oberle,
"The SWRC ontology-semantic web for research communities", *EPIA*, Springer, 2005
218 (cit. on p. 47).

[143] L. N. Soldatova and R. D. King, *An ontology of scientific experiments*,
Journal of the Royal Society Interface **3**.11 (2006) 795 (cit. on pp. 48, 142).

[144] M. Dabrowski, M. Synak and S. R. Kruk, *Bibliographic ontology*, 2009 (cit. on p. 49).

[145] Y. Raimond and S. Abdallah, *The event ontology*, Accessed = 2017-11-30, 2007,
URL: `http://motools.sourceforge.net/event/event.html` (cit. on pp. 49, 51).

[146] R. Shaw, R. Troncy and L. Hardman, "Lode: Linking open descriptions of events",
Asian semantic web conference, Springer, 2009 (cit. on p. 49).

[147] V. Tomberg, D. Lamas, M. Laanpere, W. Reinhardt and J. Jovanovic,
"Towards a comprehensive call ontology for Research 2.0", *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*,
ACM, 2011 42 (cit. on p. 50).

[148] D. Brickley, *W3C basic geo vocabulary*, 2003 (cit. on p. 51).

[149] A. Isaac and E. Summers, *SKOS Simple Knowledge Organization System*,
Primer, World Wide Web Consortium (W3C) (2009) (cit. on p. 51).

[150] S. Bechhofer, R. Stevens, G. Ng, A. Jacoby and C. Goble,
"Guiding the user: An ontology driven interface",
User Interfaces to Data Intensive Systems, 1999. Proceedings, IEEE, 1999 (cit. on p. 51).

[151] E. Kaufmann, A. Bernstein and R. Zumstein,
"Querix: A natural language interface to query ontologies based on clarification dialogs",
ISWC, 2006 (cit. on pp. 51, 53).

[152] E. Kaufmann and A. Bernstein,
"How useful are natural language interfaces to the semantic web for casual end-users?",
ISWC/ASWC, Springer, 2007 (cit. on pp. 51, 53, 124, 131).

[153] V. Lopez, E. Motta and V. Uren, "Poweraqua: Fishing the semantic web", *ESWC*,
Springer, 2006 (cit. on pp. 51, 53).

[154] L. Clark, "SPARQL views: A visual SPARQL query builder for Drupal",
9th International Semantic Web Conference, ISWC, 2010 (cit. on pp. 51, 53, 131).

[155] V. Lopez, M. Pasin and E. Motta,
"Aqualog: An ontology-portable question answering system for the semantic web",
European Semantic Web Conference, Springer, 2005 (cit. on p. 53).

[156] E. Kaufmann, A. Bernstein and L. Fischer, “NLP-Reduce: A naive but domain-independent natural language interface for querying ontologies”, *ESWC*, 2007 (cit. on p. 53).

[157] S. Campinas, “Live SPARQL Auto-Completion.”, *International Semantic Web Conference (Posters & Demos)*, 2014 (cit. on p. 53).

[158] B. Balis, T. Grabiec and M. Bubak, “Domain-driven visual query formulation over RDF data sets”, *International Conference on Parallel Processing and Applied Mathematics*, Springer, 2013 (cit. on p. 53).

[159] F. Haag, S. Lohmann, S. Siek and T. Ertl, “QueryVOWL: A visual query notation for linked data”, *European Semantic Web Conference*, Springer, 2015 (cit. on p. 53).

[160] S. Lohmann, S. Negru, F. Haag and T. Ertl, “VOWL 2: User-oriented visualization of ontologies”, *EKAW*, Springer, 2014 (cit. on p. 53).

[161] J. Wilsdon, *The metric tide: Independent review of the role of metrics in research assessment and management*, Sage, 2016 (cit. on p. 56).

[162] B. Parhami, *Low Acceptance Rates of Conference Papers Considered Harmful*, Computer **49**.4 (2016) 70 (cit. on p. 57).

[163] A. Sabharwal, *Digital curation in the digital humanities: Preserving and promoting archival and special collections*, Chandos Publishing, 2015 (cit. on p. 62).

[164] G. Scimago, *SJR–SCImago Journal & Country Rank*, 2007 (cit. on p. 63).

[165] C. M. Judd, G. H. McClelland and C. S. Ryan, *Data analysis: A model comparison approach*, Routledge, 2011 (cit. on p. 64).

[166] M. G. Larson, *Descriptive statistics and graphical displays*, Circulation **114**.1 (2006) 76 (cit. on p. 64).

[167] J. T. Behrens and C.-H. Yu, *Exploratory data analysis*, Handbook of psychology (2003) (cit. on p. 64).

[168] J. Diederich, W.-T. Balke and U. Thaden, “Demonstrating the semantic growbag: automatically creating topic facets for faceteddblp”, ACM, 2007 (cit. on p. 65).

[169] F. Osborne, E. Motta and P. Mulholland, “Exploring scholarly data with rexplore”, *International semantic web conference*, Springer, 2013 (cit. on p. 65).

[170] V. Bryl, A. Birukou, K. Eckert and M. Kessler, “What’s in the proceedings? Combining publisher’s and researcher’s perspectives” (cit. on p. 65).

[171] M. Krötzsch, D. Vrandečić and M. Völkel, “Semantic mediawiki”, *International semantic web conference*, Springer, 2006 935 (cit. on pp. 67, 168).

[172] J. Huisman and J. Smits, *Duration and quality of the peer review process: the author’s perspective*, Scientometrics **113**.1 (2017) 633 (cit. on p. 74).

[173] S. Peroni, “A simplified agile methodology for ontology development”, *OWL:Experiences and Directions–Reasoner Evaluation*, Springer, 2016 (cit. on pp. 97, 100).

[174] D. Berrueta, J. Phipps, A. Miles, T. Baker and R. Swick, *Best practice recipes for publishing RDF vocabularies*, Working draft, W3C (2008) (cit. on pp. 97, 98, 112, 137, 138, 141).

[175] R. Lewis, *Dereferencing http uris*, Draft Tag Finding, http://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14.html (2007) (cit. on p. 98).

[176] D. Garijo and M. Poveda-Villalón, *A checklist for complete vocabulary metadata*, 2017, URL: `https://w3id.org/widoco/bestPractices` (cit. on pp. 98, 139).

[177] E. Simperl, *Reusing ontologies on the Semantic Web: A feasibility study*, Data & Knowledge Engineering **68**.10 (2009) 905 (cit. on p. 99).

[178] D. Lonsdale, D. W. Embley, Y. Ding, L. Xu and M. Hepp, *Reusing ontologies and language components for ontology generation*, Data & Knowledge Engineering **69**.4 (2010) 318 (cit. on p. 99).

[179] J. G. Breslin, S. Decker, A. Harth and U. Bojars, *SIOC: an approach to connect web-based communities*, International Journal of Web Based Communities **2**.2 (2006) (cit. on p. 99).

[180] A. Gangemi and V. Presutti, “Ontology design patterns”, *Handbook on ontologies*, Springer, 2009 (cit. on pp. 101, 139, 142, 145).

[181] M. Proctor, “Drools: a rule engine for complex event processing”, *Proceedings of the 4th international conference on Applications of Graph Transformations with Industrial Relevance*, Springer-Verlag, 2011 2 (cit. on pp. 102, 146).

[182] D. Plinere and A. Borisov, *SWRL: Rule acquisition using ontology*, Scientific Journal of Riga Technical University. Computer Sciences **40**.1 (2009) 117 (cit. on pp. 102, 146, 156).

[183] M. O’connor, H. Knublauch, S. Tu, B. Grosof, M. Dean, W. Grosso and M. Musen, *Supporting rule system interoperability on the semantic web with SWRL*, The Semantic Web–ISWC 2005 (2005) 974 (cit. on pp. 103, 156).

[184] A. Gómez-Pérez, *Evaluation of ontologies*, International Journal of intelligent systems **16**.3 (2001) (cit. on p. 104).

[185] J. Yu, J. A. Thom and A. Tam, “Ontology evaluation using wikipedia categories for browsing”, *16th ACM conference on Conference on information and knowledge management*, 2007 (cit. on p. 105).

[186] C. Brewster, H. Alani, S. Dasmahapatra and Y. Wilks, “Data Driven Ontology Evaluation”, *4th International Conference on Language Resources and Evaluation (LREC’04)*, 2004 (cit. on p. 105).

[187] S. Tartir, I. B. Arpinar, M. Moore, A. P. Sheth and B. Aleman-Meza, *OntoQA: Metric-Based Ontology Quality Analysis*, IEEE ICDM Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources (2005) (cit. on p. 105).

[188] D. L. Rubin, H. Knublauch, R. W. Fergerson, O. Dameron and M. A. Musen, "Protege-owl: Creating ontology-driven reasoning applications with the web ontology language", *AMIA Annual Symposium Proceedings*, vol. 2005, American Medical Informatics Association, 2005 1179 (cit. on p. 112).

[189] C. Bizer, R. Cyganiak, T. Heath et al., *How to publish linked data on the web*, (2007) (cit. on p. 112).

[190] A. Bhardwaj, S. Bhattacherjee, A. Chavan, A. Deshpande, A. J. Elmore, S. Madden and A. G. Parameswaran, *Datahub: Collaborative data science & dataset version management at scale*, arXiv preprint arXiv:1409.0798 (2014) (cit. on p. 112).

[191] World Wide Web Consortium, *SPARQL 1.1 overview*, 2013 (cit. on p. 124).

[192] L. Feigenbaum, G. T. Williams, K. G. Clark and E. Torres, *SPARQL 1.1 Protocol*, Recommendation, W3C, 2013 (cit. on p. 127).

[193] J. Nielsen and T. K. Landauer, "A mathematical model of the finding of usability problems", *INTERACT/CHI*, ACM, 1993 (cit. on pp. 131, 175).

[194] J. Brooke et al., *SUS-A quick and dirty usability scale*, Usability evaluation in industry **189**.194 (1996) (cit. on pp. 132, 176).

[195] J. Sauro and J. R. Lewis, "When designing usability questionnaires, does it hurt to be positive?", *SIGCHI*, ACM, 2011 (cit. on pp. 132, 176).

[196] A. Bangor, P. Kortum and J. Miller, *Determining what individual SUS scores mean: Adding an adjective rating scale*, Journal of usability studies **4**.3 (2009) (cit. on pp. 132, 176).

[197] J. C. R. Licklider, *Libraries of the Future.*, (1965) (cit. on p. 136).

[198] A. Hars, *Designing scientific knowledge infrastructures: the contribution of epistemology*, Information Systems Frontiers **3**.1 (2001) (cit. on p. 136).

[199] W. Pike and M. Gahegan, *Beyond ontologies: Toward situated representations of scientific knowledge*, International Journal of Human-Computer Studies **65**.7 (2007) 674 (cit. on p. 136).

[200] R. Falco, A. Gangemi, S. Peroni, D. Shotton and F. Vitali, "Modelling OWL ontologies with Graffoo", *European Semantic Web Conference*, Springer, 2014 320 (cit. on pp. 138, 142).

[201] R. de Almeida Falbo, "SABiO: Systematic Approach for Building Ontologies.", *1st Joint Workshop Onto.Com/ODISE on Ontologies in Conceptual Modeling and Information Systems Engineering*, 2014 (cit. on pp. 138, 147).

[202] I. Jacobs, *Architecture of the world wide web, volume one*, http://www. w3. org/TR/webarch/ (2004) (cit. on p. 138).

[203] D. Garijo, "WIDOCO: a wizard for documenting ontologies", *International Semantic Web Conference*, Springer, 2017 94 (cit. on p. 139).

[204] J. Hartmann, R. Palma, Y. Sure, M. C. Suárez-Figueroa, P. Haase, A. Gómez-Pérez and R. Studer, "Ontology metadata vocabulary and applications", *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, Springer, 2005 906 (cit. on p. 139).

[205] B. C. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider and U. Sattler, *OWL 2: The next step for OWL*, Web Semantics: Science, Services and Agents on the World Wide Web **6**.4 (2008) (cit. on p. 139).

[206] V. Presutti and A. Gangemi, "Content ontology design patterns as practical building blocks for web ontologies", *International Conference on Conceptual Modeling*, Springer, 2008 128 (cit. on pp. 139, 142).

[207] K. Boyack, D. Klavans, W. Paley and K. Börner, *Scientific method: Relationships among scientific paradigms*, Seed Magazine **9** (2007) 36 (cit. on p. 140).

[208] F. Osborne, A. Salatino, A. Birukou and E. Motta, "Automatic classification of springer nature proceedings with smart topic miner", *International Semantic Web Conference*, Springer, 2016 383 (cit. on p. 141).

[209] R. Rousseau and F. O. ECOOM, *The Australian and New Zealand's Fields of Research (FoR) codes*, ISSI Newsletter **14**.3 (2018) 59 (cit. on p. 142).

[210] J. S. Mitchell, "Relationships in the Dewey Decimal Classification System", ed. by C. A. Bean and R. Green, Dordrecht: Springer Netherlands, 2001 211 (cit. on p. 142).

[211] Library of Congress contributors, *Library of Congress Classification*, [Online; accessed 7-May-2020], 2014, URL: `https://www.loc.gov/catdir/cpso/lcc.html` (cit. on p. 142).

[212] R. G. Raskin and M. J. Pan, *Knowledge representation in the semantic web for Earth and environmental terminology (SWEET)*, Computers & geosciences **31**.9 (2005) 1119 (cit. on pp. 142, 164).

[213] Wikipedia contributors, *Science — Wikipedia, The Free Encyclopedia*, [Online; accessed 7-October-2019], 2019, URL: `https://en.wikipedia.org/w/index.php?title=Science&oldid=918085492` (cit. on p. 143).

[214] J. Schickore, *Scientific discovery*, [Online; accessed 7-October-2019], 2014, URL: `https://plato.stanford.edu/entries/scientific-discovery/` (cit. on p. 143).

[215] M. Musen, B. Neumann and R. Studer, *Intelligent Information Processing: IFIP 17th World Computer Congress—TC12 Stream on Intelligent Information Processing August 25–30, 2002, Montréal, Québec, Canada*, vol. 93, Springer Science & Business Media, 2002 (cit. on p. 143).

[216] J. Völker, D. Fleischhacker and H. Stuckenschmidt, *Automatic acquisition of class disjointness*, Web Semantics: Science, Services and Agents on the World Wide Web **35** (2015) 124 (cit. on p. 143).

[217] N. Noy, A. Rector, P. Hayes and C. Welty, *Defining n-ary relations on the semantic web*, W3C working group note **12**.4 (2006) (cit. on p. 144).

[218] D. L. McGuinness, F. Van Harmelen et al., *OWL web ontology language overview*, W3C recommendation **10**.10 (2004) 2004 (cit. on p. 145).

[219] I. Horrocks, B. Parsia, P. Patel-Schneider and J. Hendler, "Semantic web architecture: Stack or two towers?", *International Workshop on Principles and Practice of Semantic Web Reasoning*, Springer, 2005 37 (cit. on p. 145).

[220] S. Fathalla, H. Mohamed and Y. Kannot, "Exploring Diseases Relationships: An Ontology-Based Spreading Activation Approach", *Computational Methods and Algorithms for Medicine and Optimized Clinical Practice*, IGI Global, 2019 133 (cit. on p. 145).

[221] 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", *Proceedings of the 10th International Conference on Knowledge Capture*, ACM, 2019 243 (cit. on p. 147).

[222] S. Vahdati, S. Fathalla, S. Auer, C. Lange and M.-E. Vidal, "Semantic Representation of Scientific Publications", *International Conference on Theory and Practice of Digital Libraries*, Springer, 2019 375 (cit. on p. 147).

[223] J. Brank, M. Grobelnik and D. Mladenic, "A survey of ontology evaluation techniques", *Proceedings of the conference on data mining and data warehouses*, Citeseer Ljubljana, Slovenia, 2005 166 (cit. on p. 149).

[224] M. Fernández-López, *Overview of methodologies for building ontologies*, (1999) (cit. on p. 153).

[225] H. S. Pinto and J. Martins, "Reusing ontologies", *AAAI 2000 Spring Symposium on Bringing Knowledge to Business Processes*, vol. 2, 000, Karlsruhe, Germany: AAAI, 2000 7 (cit. on p. 153).

[226] H. S. Pinto and J. P. Martins, *Ontologies: How can they be built?*, Knowledge and Information Systems **6**.4 (2004) 441 (cit. on p. 153).

[227] H. S. Pinto, A. Gómez-Pérez and J. P. Martins, "Some issues on ontology integration", IJCAI and the Scandinavian AI Societies. CEUR Workshop Proceedings, 1999 (cit. on p. 153).

[228] P. Hayes and B. McBride, *RDF semantics*, W3C recommendation **10** (2004) (cit. on p. 155).

[229] E. Friedman-Hill, *JESS in Action*, vol. 46, Manning Greenwich, 2003 (cit. on p. 156).

[230] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, M. Dean et al., *SWRL: A semantic web rule language combining OWL and RuleML*, W3C Member submission **21** (2004) 79 (cit. on p. 156).

[231] M. O'Connor and A. Das, "SQWRL: a query language for OWL", *Proceedings of the 6th International Conference on OWL: Experiences and Directions-Volume 529*, CEUR-WS. org, 2009 208 (cit. on p. 157).

[232] D. Spanos, P. Stavrou and N. Mitrou, *Bringing relational databases into the Semantic Web: A survey*, Semantic Web **3**.2 (2012) 169 (cit. on p. 157).

[233] M. Nentwig, M. Hartung, A. N. Ngomo and E. Rahm, *A survey of current Link Discovery frameworks*, Semantic Web **8**.3 (2017) 419 (cit. on p. 157).

[234] O. Görlitz and S. Staab, "Splendid: Sparql endpoint federation exploiting void descriptions", *Proceedings of the Second International Conference on Consuming Linked Data-Volume 782*, CEUR-WS. org, 2011 13 (cit. on p. 158).

[235] H. Hlomani and D. Stacey, *Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey*, Semantic Web Journal (2014) 1 (cit. on p. 159).

[236] I. Department of Pharmaceutical Sciences University of California, *WHAT ARE THE PHARMACEUTICAL SCIENCES?*, [Online; accessed 2-December-2019], 2019, URL: `http://pharmsci.uci.edu/about/what-are-the-pharmaceutical-sciences/` (cit. on p. 161).

[237] S. Abeyruwan, U. D. Vempati, H. Küçük-McGinty, U. Visser, A. Koleti, A. Mir, K. Sakurai, C. Chung, J. A. Bittker, P. A. Clemons et al., "Evolving BioAssay Ontology (BAO): modularization, integration and applications", *Journal of biomedical semantics*, vol. 5, S1, Springer, 2014 S5 (cit. on p. 162).

[238] M. Dumontier, C. J. Baker, J. Baran, A. Callahan, L. Chepelev, J. Cruz-Toledo, N. R. Del Rio, G. Duck, L. I. Furlong, N. Keath et al., *The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery*, Journal of biomedical semantics **5**.1 (2014) 14 (cit. on p. 162).

[239] S. Fathalla, S. Auer and C. Lange, "Towards the Semantic Formalization of Science", *Proceedings of the 35th ACM symposium on Applied computing*, ACM, 2020 In press (cit. on pp. 162, 164).

[240] M. M. Gottesman, T. Fojo and S. E. Bates, *Multidrug resistance in cancer: role of ATP–dependent transporters*, Nature Reviews Cancer **2**.1 (2002) 48 (cit. on pp. 163, 166).

[241] D. C. M. Initiative et al., *Dublin core metadata element set, version 1.1*, (2012) (cit. on p. 164).

[242] M. Compton, P. Barnaghi, L. Bermudez, R. GarciA-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog et al., *The SSN ontology of the W3C semantic sensor network incubator group*, Web semantics: science, services and agents on the World Wide Web **17** (2012) 25 (cit. on p. 164).

[243] K. Janowicz, A. Haller, S. J. Cox, D. Le Phuoc and M. Lefrançois, *SOSA: A lightweight ontology for sensors, observations, samples, and actuators*, Journal of Web Semantics **56** (2019) 1 (cit. on p. 164).

[244] J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington and F. Villa, *An ontology for describing and synthesizing ecological observation data*, Ecological informatics **2**.3 (2007) 279 (cit. on p. 164).

[245] C. Weingarten, E. Uluz, A. Schmickler, K. Braun, E. Willenborg, A. Temmler and S. Heidrich, *Glass processing with pulsed CO2 laser radiation*, Applied optics **56**.4 (2017) 777 (cit. on p. 164).

[246] D. J. Barrett, *MediaWiki: Wikipedia and beyond*, " O'Reilly Media, Inc.", 2008 (cit. on p. 168).

[247] N. A. Rakhmawati, J. Umbrich, M. Karnstedt, A. Hasnain and M. Hausenblas, *Querying over Federated SPARQL Endpoints—A State of the Art Survey*, arXiv preprint arXiv:1306.1723 (2013) (cit. on pp. 171, 173).

[248] G. Alexiou, S. Vahdati, C. Lange, G. Papastefanatos and S. Lohmann, "OpenAIRE LOD services: scholarly communication data as linked data", *International Workshop on Semantic, Analytics, Visualization*, Springer, 2016 45 (cit. on p. 183).

Appendix

APPENDIX A

List of Publications

- *Book Chapters*:

1. **Said Fathalla**, Heba Mohamed, and Yaman Kannot. *Exploring Diseases Relationships: An Ontology-Based Spreading Activation Approach.* Computational Methods and Algorithms for Medicine and Optimized Clinical Practice. IGI Global (2019): 133-159. DOI:10.4018/978-1-5225-8244-1.ch007

- *Journal Articles*:

2. **Said Fathalla**, and Yaman Kannot. *Bidirectional Spreading Activation Method for Finding Human Diseases Relatedness Using Well-Formed Disease Ontology.* International Journal of Computers in Clinical Practice (IJCCP) 2.1 (2017): 45-58. DOI:10.4018/978-1-7998-1204-3.ch090
3. **Said Fathalla**. *Detecting Human Diseases Relatedness: A Spreading Activation Approach Over Ontologies.* International Journal on Semantic Web and Information Systems (IJSWIS) 14.3 (2018): 120-133. DOI:10.4018/IJSWIS.2018070106
4. Mohamed Ali, **Said Fathalla**, Shimaa Ibrahim, Mohamed Kholief, and Yasser Hassan. *CLOE: A Cross-Lingual Ontology Enrichment Using Multi-Agent Architecture.* Enterprise Information Systems 13.7-8 (2019): 1002-1022. DOI:10.1080/17517575.2019.1592232
5. **Said Fathalla**, Sahar Vahdati, Sören Auer, and Christoph Lange. *Scholarly Event Characteristics in Four Fields of Science: A Metrics-based Analysis.* Scientometrics 123 (2020): 677–705. DOI:10.1007/s11192-020-03391-y

- *Conference Papers*:

6. **Said Fathalla**, Sahar Vahdati, Sören Auer, and Christoph Lange. *Towards a Knowledge Graph Representing Research Findings by Semantifying Survey Articles.* In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 315-327. Springer, Cham, 2017. DOI:10.1007/978-3-319-67008-9_25
7. **Said Fathalla**, Sahar Vahdati, Christoph Lange, and Sören Auer. *Analysing Scholarly Communication Metadata of Computer Science Events.* In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 342-354. Springer, Cham, 2017. DOI:10.1007/978-3-319-67008-9_27

8. Mohamed Ali, **Said Fathalla**, Mohamed Kholief, and Yasser F. Hassan. *The problem learning Non-Taxonomic Relationships of Ontologies from unstructured data sources.* In Proceedings of 23rd International Conference on Automation and Computing (ICAC), pp. 1-6. IEEE, 2017. DOI:10.23919/IConAC.2017.8082083
9. **Said Fathalla**, and Yaman Kannot. *A Bidirectional-Based Spreading Activation Method for Human Diseases Relatedness Detection Using Disease Ontology.* In International Conference on Computational Collective Intelligence (ICCCI), pp. 14-23. Springer, Cham, 2017. DOI:10.1007/978-3-319-67074-4_2
10. **Said Fathalla**, Sahar Vahdati, Sören Auer, and Christoph Lange. *SemSur: A Core Ontology for The Semantic Representation of Research Findings.* Procedia Computer Science 137 (SEMANTiCS), pp. 151-162, 2018. DOI:10.1016/j.procs.2018.09.015
11. Mohamed Ali, **Said Fathalla**, Shimaa Ibrahim, Mohamed Kholief, and Yasser Hassan. *Cross-Lingual Ontology Enrichment Based on Multi-Agent Architecture.* Procedia Computer Science 137 (SEMANTiCS), pp. 127-138, 2018. DOI:10.1016/j.procs.2018.09.013
12. **Said Fathalla**, and Christoph Lange. *EVENTSKG: A Knowledge Graph Representation for Top-Prestigious Computer Science Events Metadata.* In International Conference on Computational Collective Intelligence (ICCCI), pp. 53-63. Springer, Cham, 2018. DOI:10.1007/978-3-319-98443-8_6
13. **Said Fathalla**, Sahar Vahdati, Sören Auer, and Christoph Lange. *Metadata Analysis of Scholarly Events of Computer Science, Physics, Engineering, and Mathematics.* In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 116-128. Springer, Cham, 2018. DOI:10.1007/978-3-030-00066-0_10
14. **Said Fathalla**, Christoph Lange, and Sören Auer. *EVENTSKG: A 5-Star Dataset of Top-Ranked Events in Eight Computer Science Communities.* In European Semantic Web Conference (ESWC), pp. 427-442. Springer, Cham, 2019. DOI:10.1007/978-3-030-21348-0_28
15. **Said Fathalla**, Sahar Vahdati, Christoph Lange and Sören Auer, *SEO: A Scientific Events Data Model.* In International Semantic Web Conference (ISWC), pp. 79-95, Springer, 2019. DOI:10.1007/978-3-030-30796-7_6
16. **Said Fathalla**, Christoph Lange, and Sören Auer. *A Human-friendly Query Generation Frontend for a Scientific Events Knowledge Graph.* In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 200-214. Springer, Cham, 2019. DOI:10.1007/978-3-030-30760-8_18
17. Shimaa Ibrahim, **Said Fathalla**, Hamed Shariat Yazdi, Jens Lehmann, and Hajira Jabeen. *From Monolingual to Multilingual Ontologies: The Role of Cross-Lingual Ontology Enrichment.* Procedia Computer Science 137 (SEMANTiCS), 2019. DOI:10.1007/978-3-030-33220-4_16
18. Zeynep Say, **Said Fathalla**, Sahar Vahdati, Jens Lehmann and Sören Auer. *Ontology Design for Pharmaceutical Research Outcomes.* In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 119-132. Springer, 2020. DOI:10.1007/978-3-030-54956-5_9
19. Aysegul Say, **Said Fathalla**, Sahar Vahdati, Jens Lehmann and Sören Auer. *Semantic Representation of Physics Research Data.* In 12th International Joint Conference on

Knowledge Discovery, Knowledge Engineering and Knowledge Management, In press, SciTePress, 2020. DOI:10.5220/0010111000640075

- *Workshops and Posters Papers*:

  20. **Said Fathalla** and Christoph Lange. *EVENTS: A Dataset on The History of Top-Prestigious Events in Five Computer Science Communities.* In Proceedings of Semantics, Analytics, Visualization (SAVE-SD) at the World Wide Web conference. Springer, Cham, pp. 110-120, 2017. DOI:10.1007/978-3-030-01379-0_8
  21. **Said Fathalla**, Sahar Vahdati, Sören Auer, and Christoph Lange. *The Scientific Events Ontology of The Openresearch.org Curation Platform.* In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (SAC), pp. 2311-2313. ACM, 2019. DOI:10.1145/3297280.3297631
  22. Shimaa Ibrahim, **Said Fathalla**, Hamed Shariat Yazdi, Jens Lehmann, and Hajira Jabeen. *OECM: A Cross-Lingual Approach for Ontology Enrichment.* In European Semantic Web Conference (ESWC), pp. 100-104. Springer, Cham, 2019. DOI:10.1007/978-3-030-32327-1_20
  23. Sahar Vahdati, **Said Fathalla**, Sören Auer, Christoph Lange, and Maria-Esther Vidal. *Semantic Representation of Scientific Publications.* In International Conference on Theory and Practice of Digital Libraries (TPDL), pp. 375-379. Springer, Cham, 2019. DOI:10.1007/978-3-030-30760-8_37
  24. **Said Fathalla**, Sören Auer, and Christoph Lange. *Towards The Semantic Formalization of Science: The Science Knowledge Graph Ontologies Suite.* In Proceedings of the 35th ACM/SIGAPP Symposium on Applied Computing (SAC), pp. 2057-2059, 2020. DOI:10.1145/3341105.3374132
  25. Heba Mohamed, **Said Fathalla**, Jens Lehmann and Hajira Jabeen. *A Distributed Approach for Parsing Large-Scale OWL Datasets.* In 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, In press, SciTePress, 2020.

- *Working Draft and submitted papers*:

  26. **Said Fathalla**, Christoph Lange and Sören Auer. *ModSci: The Modern Science Ontology.* Submitted to Applied Ontology Journal, 2020.
  27. **Said Fathalla**, Sahar Vahdati, Mojtaba Nayyeri, Christoph Lange, Jens Lehmann, Sören Auer, Maria-Esther Vidal. *Exploring Scholarly Knowledge Graphs: A Wiki-based Approach.*
  28. Arthur Lackner, **Said Fathalla**, Sahar Vahdati, Andreas Behrend, Jens Lehmann and Rainer Manthey. *Semantic Representation of Common Characteristics of Scientific Events: Metadata Analysis of Renowned Computer Science Events.* Scientometrics Journal.

List of Figures

List of Tables