Through the Lens of the Web Conference Series: A Look Into the History of the Web

Damien Graux
Inria, Université Côte d'Azur, CNRS, I3S
Sophia Antipolis, France
damien.graux@inria.fr

Fabrizio Orlandi
ADAPT SFI Centre, Trinity College Dublin
Dublin, Ireland
orlandif@tcd.ie

ABSTRACT

During the last three decades, the Web has been growing considerably in terms of number of available resources, traffic, types of media, usages, etc. In parallel, with 30+ editions, the WebConf series (ex. WWW, soon-to-be ACM WebConf) has witnessed how academia has been dealing with the Web as an object of research. In this study, we focus on the *small* story within the *great* one of the Web. In particular, by analysing the WebConf's accepted papers and yearly events, we review how the conference has evolved across these decades and "driven" the evolution of the Web.

CCS CONCEPTS

• **General and reference** → **Surveys and overviews**; **General conference proceedings**; • **Information systems** → **World Wide Web**.

KEYWORDS

History, Web, WWW, WebConf series, Systematic Review, Conference Metadata

ACM Reference Format:
Damien Graux and Fabrizio Orlandi. 2022. Through the Lens of the Web Conference Series: A Look Into the History of the Web. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3485447.3512281

1 INTRODUCTION: A WEBCONF HISTORY

The First International WWW Conference was organised in 1994, in Geneva, by Robert Cailliau[1]. Only a year earlier, in 1993, CERN put the World Wide Web software in the public domain. The conference series has been organized by the International World Wide Web Conference Committee (IW3C2) every year since. Through its numerous editions (2022 will witness the 31st edition in Lyon, France), the conference series has been pushing the agenda for most of the main Web-associated technologies and W3C standards. For example, *XML* topics were heavily discussed at the venues during the 2000s and similarly have been the *Semantic Web* standards since the tenth edition. Another example could be the aspects of *security and privacy* on the Web, which have been included in almost every scientific program. In practice, the conference series published more than 7 000 articles over almost three decades, and has been a place where 15 000+ distinct authors exchanged ideas and solutions on Web-related topics. This illustrates that the Web is no longer exclusively an artifact but, as shown by [1], it is also a science object and an object for research.

In this article, we perform an analysis of all the WebConf articles and conference metadata (such as sessions, sponsors, proceedings, *etc.*) in order to have a better understanding of the Web evolution through the lens of the WebConf series. To do so, we crawled several Web sources[2] to collect all the necessary data. We then performed several aggregating tasks to enable comparisons between the conference editions.

The rest of the article describes the process of collecting these pieces of information; and then shows our findings for different aspects of the conference, to name a few: *who are the most prolific authors? where do they come from? what are the main topics?* After briefly reminding similar initiatives in other research communities, we conclude the article by describing the Web-research community.

2 DATA ACQUISITION

This section recaps how we collected the necessary data to conduct our analyses. Actually, gathering information on the Web is often a challenge, especially when the considered time window is almost three decades wide, meaning rolling back almost at the beginning of the Web itself.

In order to collect the bibliographic information, we reviewed several sources: DBLP[3] to have information about the 30 editions and the ACM[4] which gathers information about the conference series since 2001 (the Hong Kong 10th edition). Practically, crawling these sources is not a straightforward task as not everything is available through the APIs and as the ACM limits the number of requests considerably. In addition, there are restrictions for the download of the `pdf` files, even for the oldest editions (*e.g.* 2001, 2002, *etc.*). That is why, in order to ease the accessibility of the previous articles published, we share the information we gathered, curated and sorted to the community, making it available at the following repository:

<https://github.com/dgraux/webconf-history/>

This resource provides for instance, among other details, the DOIs (or HTTP links) pointing to the original articles.

[1] https://www.iw3c2.org/conferences/

[2] We sometimes had to search for old mirror versions of the websites as the original ones were no longer served online.
[3] https://dblp.org/
[4] https://www.acm.org/

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9096-5/22/04.
https://doi.org/10.1145/3485447.3512281

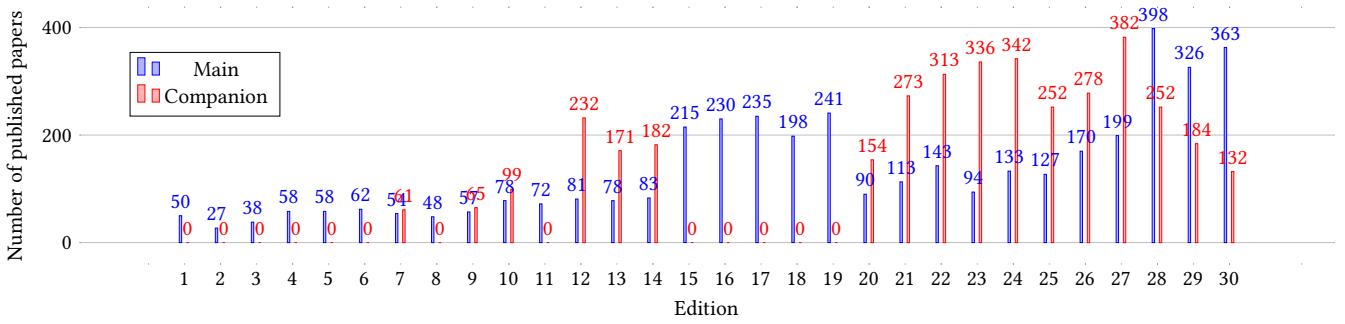

Figure 1: Number of articles published for each edition both as "main" paper or in the companion volume.

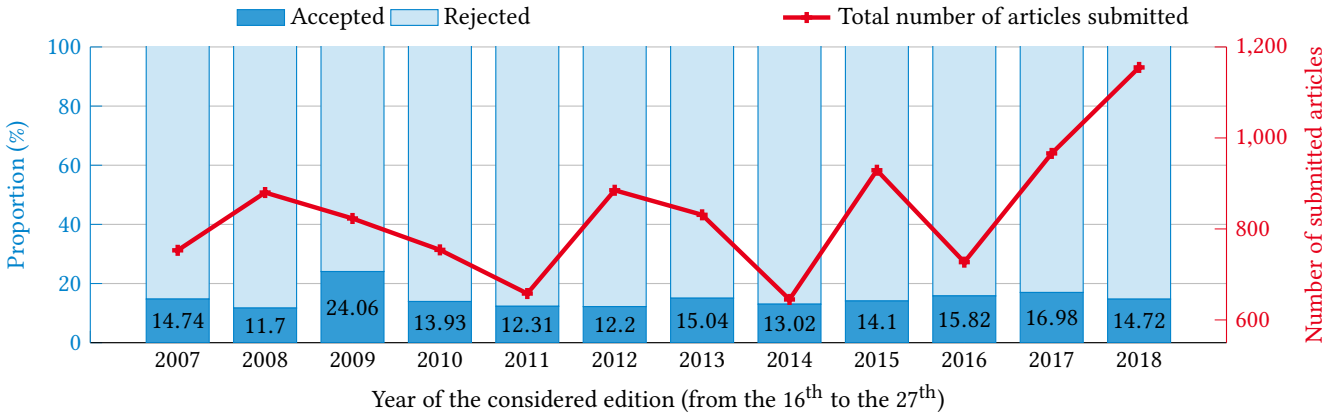

Figure 2: Acceptance rates from 2007 to 2018 and the total number of submitted articles.

As expected, the most complicated pieces of information to find were the conference metadata. It is indeed quite easy to find the list of accepted papers together with their titles, keywords, authors' names and affiliations. (There were, nonetheless, some exceptions for some older editions such as the third or the ninth ones.) However, obtaining the information such as sponsoring, acceptance rates, *etc.* requires to explore the online available sources such as the conference websites for each edition. These ones have often changed URL location and for most of them the International World Wide Web Conference Committee[5] (IW3C2) keeps a mirror accessible.

Ultimately, we had to "summon" the Internet Archive Wayback Machine[6] to exhume now-disappeared websites when the IW3C2 could not.

3 ANALYSIS & RESULTS

In this Section, we present the various findings we obtained after analysing systematically all the editions of the WebConf. In particular, we divide our results in several facets from statistical aspects to geographical notions passing by topics and focuses of the research efforts published by the conference series.

3.1 Scientific production overall

The WebConf conference series, previously named the International World Wide Web Conference, shortened www, has been running for 30 editions (since 1994) and the 31st one is scheduled in Lyon in 2022. As of today, thousands of research articles have been published and presented at the WebConf.

Number of published articles. We present in Figure 1 the number of published papers per edition. The obtained figures are based on the sources we presented in Section 2. Moreover, it is worth noting that during the first years, the publishing process was not always the same between editions. Actually, for some editions, only selected articles were published in Journals. Since the 10th edition, proceedings are more stable as the WebConf always goes through the ACM.

As showed on Figure 1, the number of accepted papers in the main track has been overall within the same order of magnitude for the first 11 editions: around 60 articles. From the seventh edition, companion proceedings started to be published too. Between the 12th and 20th editions, the volume of articles jumped between 200 and 300; in addition, Figure 1 also reflects the fact that main and companion volumes were joined between 15th and 19th. Later, the 300-papers bar was reached and the pace kept growing to pass the 500-papers bar in the 27th edition.

[5]https://thewebconf.org/
[6]https://web.archive.org/

#Paper	#Author	#Paper	#Author	#Paper	#Author
34	1	17	2	8	37
29	2	15	7	7	36
24	1	14	5	6	83
23	4	13	10	5	125
22	1	12	4	4	241
20	1	11	7	3	434
19	4	10	16	2	1 259
18	1	9	28	1	6 660

Table 1: Number of authors having a certain number of papers, since 2001 for the main proceedings.

Acceptance rate. As mentioned prior, finding conference metadata such as acceptance rates is not easy. Fortunately, we were lucky enough to find some information from the conferences' websites and the ACM portal. Figure 2 references the acceptance rates for twelve editions, from 2007 to 2018. We notice that during this period, the acceptance rate remains every year (but in 2009) under 17%. In parallel, Figure 2 displays also the total number of submitted papers for each year (the red curve). In line with the previous discussion about the number of accepted papers per edition (see Figure 1), the red curve of submissions particularly increases for the 26th and 27th editions, leading to an increasing number of accepted papers as the acceptance rates do not change drastically.

Overall. These high-level statistics show the WebConf series is attractive and vivid for the community. After establishing itself in a decade, it started to attract more submissions while keeping low acceptance rate.

3.2 The population of authors

Since the tenth edition in 2001, there has been 15 297 distinct authors who published an article at the WebConf either in the main proceedings or in associated companion volumes. More precisely, 8 969 distinct authors published in main volumes and 8 381 in companion ones. Among them, some people have been more prolific than others; for instance during the last two decades Jure Leskovec has been the top publisher in the main proceedings with 34 distinct papers, followed by Ravi Kumar and Jiawei Han having each 29 articles.

More generally, Table 1 lists the couples (number of paper, number of author) since 2001 for the main proceedings. It is interesting to see that the majority of the authors (6 660 over 8 969) have only one article published at a WebConf edition. Indeed, there are less than 1 000 people having strictly more than 2 papers. And to be in the top-10, one needs to have at least 20+ articles.

Figure 3 shows the number of distinct authors per year. We confirm our previous findings about the increasing number of submissions these last three years. The number of distinct authors is also jumping to the upper order of magnitude.

Finally, we have a look at the co-authorship distribution (Figure 4) *i.e.* the number of papers in function of their number of authors. We remark that the majority of the published papers had 2 to 4 authors, with a max for 3 authors. Few papers only have one author (148) and a dozen have more than 10 authors over two decades.

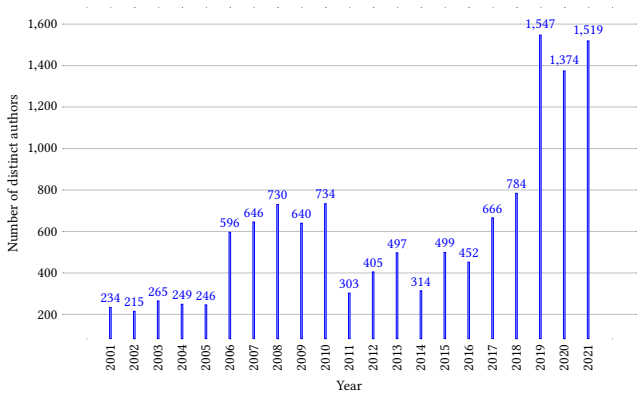

Figure 3: Number of distinct authors per year published in the main volume over two decades.

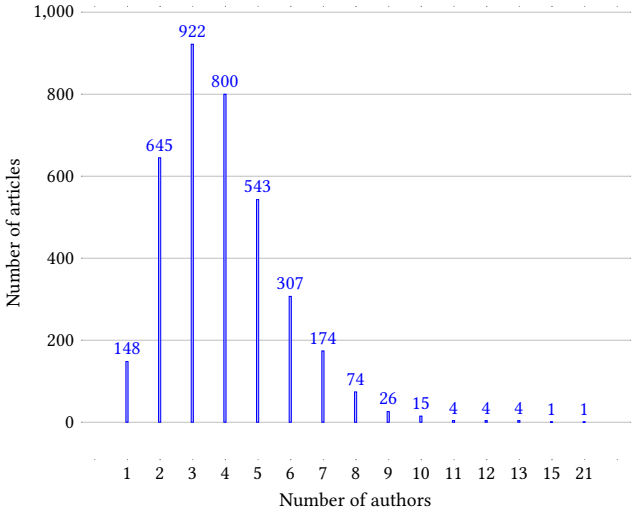

Figure 4: Co-authorship distribution in the main proceedings, since 2001.

3.3 Affiliations

As just showed, co-authorship is common in the WebConf community. And as we know, researchers regularly use conferences to publish joint efforts. Knowing which affiliation is listed by an author[7] helps understand how research is funded and how research strategies are done. In Table 2, we listed the top-50 affiliations as per the number of published papers they were referred in authors' signatures. We highlighted in gray the non-public affiliations. It occurs that, among the top-50, 13 are not publicly funded. More important, we discovered that the podium is composed of private labs: Yahoo Research Labs, Microsoft Research and Google LLC. Moreover, if we sum up the scores obtained by the various subsidiaries in the top-50, the podium changes to Microsoft, Yahoo and Google.

[7] An author may have several affiliations listed in her signature as funds might be coming from several bodies.

Affiliation	Paper Count
Yahoo Research Labs	224
Microsoft Research	220
Google LLC	211
Tsinghua University	185
Microsoft Corporation	170
Stanford University	152
Carnegie Mellon University	143
Microsoft Research Asia	107
Cornell University	103
Peking University	102
IBM Thomas J. Watson Research Center	92
University of Southampton	90
Pennsylvania State University	89
National University of Singapore	88
Chinese University of Hong Kong	82
Yahoo Inc.	80
Georgia Institute of Technology	79
Chinese Academy of Sciences	76
University of California, Berkeley	74
Max Planck Institute for Informatics	73
University of Oxford	68
IBM Research - Almaden	65
University of California, Santa Barbara	61
Massachusetts Institute of Technology	59
Zhejiang University	58
University College London	58
EPFL	58
Arizona State University	54
Kyoto University	54
Shanghai Jiao Tong University	53
University of Illinois at Chicago	53
National Taiwan University	52
University of Michigan, Ann Arbor	52
Hong Kong University of Science and Technology	51
IBM Research	51
Facebook, Inc.	51
Alibaba Group Holding Limited	51
Technion - Israel Institute of Technology	49
Politecnico di Milano	49
New York University	47
Yahoo Research Barcelona	47
The University of Tokyo	43
University of Southern California	43
University of Toronto	42
Qatar Computing Research Institute	42
Stony Brook University	41
Tencent Holdings Limited	40
Indian Institute of Technology Kharagpur	39
Federal University of Minas Gerais	39
University of California, San Diego	39

Table 2: Top-50 affiliations in terms of papers, since 2001.

From a public-funded point of view, universities and labs from North-America, Western Europe and Asia are present in the top-50. Tsinghua is even in the top-5.

More generally, even if a majority of the top-50 affiliations are publicly-funded, the fact that the top-5 contains 4 private labs suggests that the industry has been and is very interested in Web related topics.

3.4 The WebConf Venues

Since 1994, the WebConf has been organized at different places. In this Section, we are focusing on the countries where these events have taken place. Figure 5 presents a map where countries are colored depending on the number of times it's been hosting the WebConf. At one glance, poles emerge: a North American one and a Western European one. Indeed, the United States and Canada have hosted the conference respectively 7 and 3 times. In parallel, France hosted it 4 times. In terms of proportions, the US. and Canada hosted 10 out of 31 editions and Western Europe 12 editions. Thereby, only one third of the editions happened outside this regions *i.e.* distributed mainly in Asia (6 times).

Overall, we note that the WebConf has never been yet in Africa and only once in South America. This might, in a sense, be a drawback for the adoption of our community's efforts especially considering that some research areas are focusing on "developing regions" (see next Section). We can also quote the IW3C2 about the venues aspect:

> " The location of the conference rotates among America, Europe, and Asia-Pacific. Starting in 2022, the conference will become an ACM/SIGWEB event and the rotation between the three geographical areas will no longer be the rule. *IW3C2*"

We, therefore, hope this change of event-type will increase the mobility of the WebConf.

3.5 Topics & Areas

Among the conference metadata collected, we also listed the sessions names under which research efforts were presented at the conferences. We reviewed all the sessions but the ones for the very first edition and the nineteenth edition. In order to better compare the scientific programs, we had to review the main topics of each sessions as their names might evolve from one edition to another. This led us to group under the same banners/names, the topics that were dealt with at the conferences.

Table 3 presents this task of thematically normalising the research areas explored during the last 30 WebConf editions. In a nutshell, the Table presents in the most-left column the list of broad research topics covered by the conference series across almost 30 years and then each line is composed by information revealing the presence (or not) of the topic for each editions. In addition, color-shading together with number indicates if a topic was the focus of several sessions. For instance the black box on the "Teaching/Education" line indicates that there were 7 distinct sessions focusing on this topics during the second edition of the WebConf.

Visually, the topic list (left column) is sorted by order of chronological appearance in the editions instead of a alphabetical sorting. This allows us to the "curve" of new interests. For instance, one could see the emerging trends in our community following this curve to discover that the "Performance" aspects caught researchers' interest around the 2000s; similarly, more recently "Crowdsourcing" became something around the 21st edition.

With Table 3, it is easy to note the *pillar* topics of the WebConf *i.e.* the ones that have been and are almost present at each edition and often in several dedicated sessions. This is typically the case of the "Security & Pivacy" and "Search" topics.

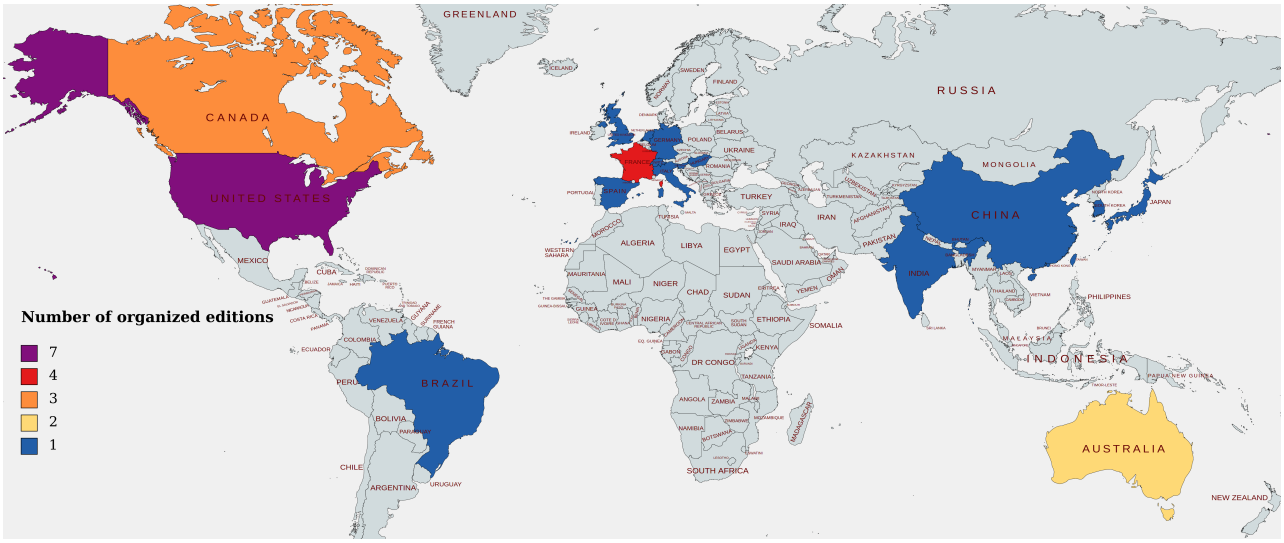

Figure 5: Countries having organised the WebConf 31 editions (from 1994 to 2022).

Using the color coding, we show that some topics had their *glorious days* at some point in the past. This is, for example, the case of the "Semantic Web" in the 2000s where editions could have up to 5 sessions dedicated to the topic. Or, for example, the social network in the 2010s.

On the contrary, other topics were only the focus of a session once (or very sporadically). This is the case of "Programming Language" or "Protocols". We also notice that there are topics that appeared a long time ago but that start to take-off only few editions back, *e.g.* "Question Answering".

Finally, Table 3 allows to find the areas which no longer have dedicated sessions. In the 90s, "Teaching and Education" were important topics. The 2000s witnessed the rise of "XML" but since 2010 it is not a session anymore.

Overall. Through almost three decades, the focuses of the WebConf have evolved while keeping some of them as must-have sessions. The Web community changes with the society see for instance the emerging of the "Health" topic in the recent years.

3.6 Article citations

Citations are collected[8] since www'01 *i.e.* the tenth edition held in Hong Kong in 2001 until the latest 2021 edition. Overall, the 7 368 papers (listed by the ACM) received a total number of 183 274 citations. This corresponds to an average of approximately 25 citations per article, no matter if it is coming from the companion or the main volume of the proceedings.

In Table 4, we present the number of citations received by all the articles of a specific proceedings volume (*i.e.* of a specific edition) since it was published. For instance, articles published in the 2002's edition have been cited 5 275 times as of December 8th 2021. In order to better see the early impact of a volume, we also normalise by the number of past years since the publication. We now have the yearly average citations received by a volume.

[8]According to the ACM <https://dl.acm.org/conference/www> as of Dec.8th 2021.

Since 2001, according to the ACM, the most cited articles published at the WebConf are the following:

#Citation	Info.	Year
4 326	Item-based collaborative filtering recommendation algorithms *By* Sarvar et al.	2001
3 426	What is Twitter, a social network or a news media? *By* Kwak et al.	2010
2 153	Earthquake shakes Twitter users: real-time event detection by social sensors *By* Sakaki et al.	2010
1 879	LINE: Large-scale Information Network Embedding *By* Tang et al.	2015
1 679	Yago: a core of semantic knowledge *By* Suchanek et al.	2007
1 408	Neural Collaborative Filtering *By* He et al.	2017
1 215	The Eigentrust algorithm for reputation management in P2P networks *By* Kamvar et al.	2003
1 131	Mining interesting locations and travel sequences from GPS trajectories *By* Zheng et al.	2009

We can see that these highly cited research efforts and results have been published from 2001 to 2017. This suggests that the WebConf has been and is still a venue where to-be-well-recognized science is discussed and presented.

3.7 Awards

Since 2015, the Seoul Test of Time Award[9] has been given each year to the author, or authors of a paper, presented at a previous World Wide Web Conference, that has, as the name suggests, stood the test of time. The first award was given to S. Brin and L. Page for "The Anatomy of a Large-Scale Hypertextual Web Search Engine" (2008) which led to the status of nowadays search engines. In parallel, since then, the "Search" topic has been one of the WebConf's pillars (see Table 3). Moreover, among the seven awards given, three of them are listed above in the list of most-cited articles

[9]See https://www.iw3c2.org/ToT for a detailed list of the awards.

Table 3: Main conference areas per edition.

Year	Total citations	Avg. per year	Number of papers
2001	10 213	511	78
2002	5 275	278	72
2003	9 178	510	81
2004	9 676	569	78
2005	7 110	444	83
2006	7 628	509	215
2007	16 629	1 188	230
2008	12 518	963	235
2009	11 928	994	198
2010	17 844	1 622	241
2011	7 026	703	90
2012	6 969	774	113
2013	7 053	882	143
2014	3 445	492	94
2015	5 682	947	133
2016	3 579	716	127
2017	5 526	1 382	170
2018	3 848	1 283	199
2019	3 529	1 765	398
2020	1 442	1 442	326
2021	252	*N/A*	363

Table 4: Total citations obtained for the main papers.

from the WebConf: "Item-based collaborative filtering recommendation algorithms", "Yago: a core of semantic knowledge" and "The Eigentrust algorithm for reputation management in P2P networks". More generally, having such awards implies that the community has been gathering for long enough to be able to recognize among its contributors the ones who changed paradigms and behaviors.

3.8 Recurrent sponsors

Finally, we review the sponsor lists still available[10] on the conference editions' websites. These lists are very different from one to another, mainly because among the sponsors a great share is often local. That is why we searched for the most common ones. It appears that the top-4 is as follows: Microsoft, IBM, Google and Yahoo! Moreover, these companies are usually listed among the biggest sponsors *i.e.* gold or platinum. This shows that these companies share the same focuses as the Web community and therefore tend to promote our events. In addition, it also means that their own research agendas are fitting with the topics of the conference. It is therefore no longer a surprise to find 3 of these 4 sponsors among the top-5 of the most represented affiliations on Table 2.

4 SIMILAR INITIATIVES IN OTHER COMMUNITIES

Analysing the history of relevant conference series is common practice, especially for conferences that have spanned through several decades. These analyses are typically useful for conference organisers in order to understand the evolution of a series and the research topics. Relevant examples are available for the WEBIST conferences [6] and the SIGMOD series [4], or even the STM Journals [3]. Others have looked at academic publications only within a particular country (UK) [5] instead of a conference series. And there are also larger worldwide observatories for the Web, and Web Science [2, 7], which have a wider scope and are not focused on Web research publications. However, with this paper we go beyond a straightforward analysis of the numbers of submissions for the different tracks and authors. This contribution aims at providing additional insight into the topics of the Web and its development. By exploring the topics of the papers and the respective research tracks, together with the authors, their affiliations, the citations and their geographical locations, this paper provides a comprehensive analysis of the evolution of Web research for more than two decades. Finally, the collected data will be available online for reproducible research and additional analyses. To the best of our knowledge, this is the first time that such a corpus of aggregated data regarding the WebConf is readily available as a resource for the Web community.

[10]During the first editions, there were no declared sponsors.

5 CONCLUSION

In this article, we looked back at what the WebConf series has achieved and how it has been received by the community. In particular, we conducted an analytical review of the metadata of the conference editions and of the submitted & accepted articles during the WebConf's almost three decades of existence. Using this metadata, we reviewed the various facets of the conference series: from high-level statistics to sponsoring companies, passing by an in-depth analysis of the research areas across the decades.

Our analysis led us to highlight the vividness of the research Web community considering both the number of authors (or affiliations) and their collaborations through co-authored paper publications. While at the same time, we also pinpointed the lack of diversity when it comes to organising countries. We hope this study will help the community self-reflect on its evolution, in order to keep growing and to continue gathering Web-passionate researchers each year.

Acknowledgments. This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT Centre at Trinity College Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by the SFI Research Centres Programme.

REFERENCES

[1] Fabien Gandon and Wendy Hall. 2022. A never-ending project for humanity called "the Web". *The ACM Web Conference* (2022).
[2] James Hendler, Nigel Shadbolt, Wendy Hall, Tim Berners-Lee, and Daniel Weitzner. 2008. Web science: an interdisciplinary approach to understanding the web. *Commun. ACM* 51, 7 (2008), 60–69.
[3] Steve M. Hitchcock, Les A. Carr, and Wendy Hall. 1996. A Survey of STM Online Journals 1990-95: the Calm before the Storm. In *World Wide Web Internet And Web Information Systems*. Vol. 1996. Association of Research Libraries, 1–17. http://eprints.ecs.soton.ac.uk/742/1/survey.html
[4] Mario A. Nascimento, Jörg Sander, and Jeffrey Pound. 2003. Analysis of SIGMOD's co-authorship graph. *ACM SIGMOD Record* 32, 3 (sep 2003), 8–10. https://doi.org/10.1145/945721.945722
[5] Nigel Payne and Mike Thelwall. 2004. A statistical analysis of UK academic web. *Cybermetrics* 8, 1 (2004), 19–31.
[6] Giseli Rabello Lopes, Bernardo Pereira Nunes, Luiz André P. Paes Leme, Terhi Nurmikko-Fuller, and Marco A. Casanova. 2015. Knowing the past to Plan for the Future - An In-depth Analysis of the First 10 Editions of the WEBIST Conference. In *Proceedings of the 11th International Conference on Web Information Systems and Technologies*. SCITEPRESS - Science and and Technology Publications, 431–442. https://doi.org/10.5220/0005447704310442
[7] Thanassis Tiropanis, Wendy Hall, Nigel Shadbolt, David De Roure, Noshir Contractor, and Jim Hendler. 2013. The web science observatory. *IEEE Intelligent Systems* 28, 2 (2013), 100–104. https://doi.org/10.1109/MIS.2013.50