KnE Social Sciences

IC-HEDS 2019
International Conference on Humanities, Education, and Social Sciences
Volume 2020

Knowledge E
enriching | engaging | empowering

Conference Paper

SEforRA: A Bibliometrics-ready Academic Digital Library Search Engine Alternative

Muhaemin Sidiq[1], Ivan Hanafi[2], and Fajar J. Ekaputra[3]

[1]Educational Technology Doctorate Program, Universitas Negeri Jakarta, Indonesia
[2]Vocational Education Program, Universitas Negeri Jakarta, Indonesia
[3]Institute of Systems Engineering, Faculty of Informatics - TU Wien, Austria

Abstract

Naturally, not all researchers can develop their own software to search for academic publications from digital libraries. Nevertheless, at several stages of their research, they will need to search digital libraries for relevant scientific publications and bibliometric information. There are typically two approaches used by researchers to search for scientific publications: (i) using Google Scholar search, or (ii) using publication metadata available from several sources, such as CrossRef and publishers. However, in developing countries like Indonesia, neither option provided users with complete information, since (i) Google Scholar does not provide bibliometric details, and (ii) complete bibliometric information from other sources is often not available due to incomplete data (e.g., CrossRef) or the necessity to pay a subscription fee (e.g., Springer and Elsevier). The development of Search Engine for Research Articles (SEforRA) is a solution to this issue which provides researchers with bibliometric-ready publication metadata. SEforRA extracts and processes data from CrossRef, publishers, and other sources to provide an integrated platform for researchers to search and retrieve publication metadata, which is ready to use further in their research.

Keywords: search engine for research articles, academic search engines, text data mining, bibliometrics

Corresponding Author:
Muhaemin Sidiq
MuhaeminSidiq_9902918009@mhs.unj.ac.id

Published: 11 November 2020

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the IC-HEDS 2019 Conference Committee.

1. Introduction

Science and its application in daily life have developed rapidly and affect the quality of human life [1], this is inseparable from the role of scientific publications as part of the documentation of research activities to disseminate trusted scientific findings after going through a very rigorous and quality peer-review process, which enables the dissemination of information regarding research achievements and further recommended research, scientific publications are also a means to exchange ideas and criticize one another [1–4]. Several parties also used this development for commercial purposes,

primarily scientific publications controlled by Elsevier, Wiley, Taylor & Francis, Springer-Nature, and Sage, resulting in an institutional obligation to pay subscription fees to facilitate individuals accessing the articles they publish. Although there are options available to publish articles that are online open access for readers, most publishers still apply a quite expensive article processing costs to the author, even though most scientific publications come from research at universities, which in turn forces university libraries to make agreements with publishers about the trade-off between subscription fees and the open-access element [3].

Web of Science (WoS) as a platform for searching for scientific literature and analytical information has been widely used in thousands of academic studies over the past 20 years [5]. Even though extracted data from WoS is useful and widely used as a data source for bibliometric methods [6], it has a significant limitation that researchers and academics cannot access WoS individually, only through a subscribed institution [3, 7]. Likewise, scientific publications indexed by Scopus often use reliable literature searches by journal's publication performance rank through ScimagoJR, but to use Scopus to obtain meta-data for bibliometric methods is a paid service [7, 8].

The bibliometric method is a statistical analysis of publications which are widely used by researchers, governments, and organizations to identify patterns of scientific publication as a basis for decision-making [9], direction and novelty of research [10], and even to assess technological maturity of research [9]. VOSViewer software developed by the Center for Science and Technology Studies (CWTS) at Leiden University by Nees Jan van Eck and Ludo Waltman, can be used to carry out publication analysis using the bibliometric method [11].

To facilitate discussion of research carried out in various separate repositories it requires a way to share data and to calculate metrics, Crossref can do this processes [12], Crossref provides missing links in linking various large and small publishers through open and interoperable systems with different other connecting systems [13]. The scope and impact of Crossref have been recognized globally, besides being an agent for registering DOI for scientific content, it also provides tools and open-source the global research community widely uses that, Crossref provides metadata for international academics community [14]. Searches from Crossref, Dimensions, Google Scholar, Microsoft Academic, Scopus, and Web of Science show differences in limitations and search options. Free access to the facilities can also use as a weighted option for academics to literature finding and citation analysis [15]. Therefore, Crossref deserves to be the primary source in the search for scientific literature, but Crossref direct search results cannot be exported and analyzed further using bibliometric tools.

The Crossref REST API exposed the metadata provided by publishers to Crossref when they registered the content, and it was not just bibliographic metadata. This data mining solution simplifies access for researchers who wish to mine and analyze research outputs and content depend on search keywords when searching on a search facility on a publisher's website or an academic search engine [16]. Due to current research developments, researchers increasingly need to access full-text content for data mining and analysis, so researchers need ways to avoid the complexity of negotiations with individuals and publishers [17]. But for the needs of study using the bibliometric method not only requires metadata but also requires metric data from the publication as a condition to keep the metadata analyzed is the publication metadata that has a high impact. The H-Index value is often used as a reference to determine the impact of scientific publications, although some studies find weaknesses of the H-Index [17, 18]. Researchers can get H-index scores yearly from complete scientific journals at ScimagoJR.

Based on the previous presentation, the main problem faced by researchers is there no search engine able to provide scientific publications metadata by filtering based on journal rankings and metric data and equipped with bibliometrics ready data export features. Search Engine for Research Articles (SEforRA) development is a solution to overcome the problems as stated that cannot be solved by such as other scientific publications search engines.

2. Methods and Development Steps

2.1. Methods

Agile software development works mainly for smaller projects [20] in software development, is increasingly being used because it allows for changes during development so that it is following the consumers' needs, far more effective than the traditional software development paradigm [21–23]. Agile development methods are used in this research because they support active end-user involvement, tolerance of change, and evolutionary product delivery [24-26]. This research uses a Successive Approximation Model (SAM) to accommodate user feedback and limited in identifying project needs, much less daunting than other methods, sharing many concepts with Agile [27]. Web applications are attractive because they require no installation or deployment steps on clients and enable large scale collaborative experiences, using the same programming language is not enough because the client and server programming environments

are not the same [28], PHP programming language is used as the main programming discussion with CSS support for display settings, and JavaScript as the backbone of AJAX.

The system development project in this research involved only a small amount of human resources. It was suggested by Allen [29] to use 2 phase version of SAM: a preparatory or backgrounding phase and an interleaved design, build, and review phase, Figure 1 shows the stages, portrayal project development perfectly executed in small steps rather than directly in giant steps.

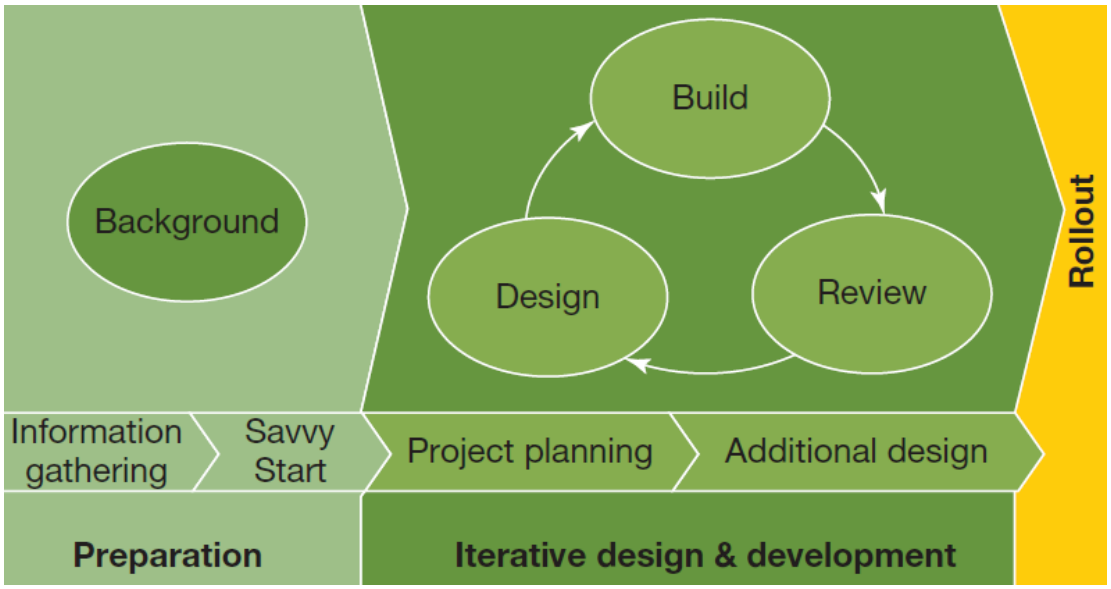

Figure 1: Two-phase successive approximation model (SAM) [29]

2.2. Development Steps

2.2.1. Prototype Development

VOSviewer, as one of the supporting software to carry out analysis using bibliometric methods, at least requires DOI list data, corpus files, reference manager files, and score files [30, 31]. DOI list data is useful for mining scientific publications metadata from the Crossref, EuropePMC, Semantic Scholar, OCC, COCI, and Wikidata databases using the metadata mining features available on VOSviewer. However, mining metadata through this feature does not always provide complete data, often interrupted when mining large amounts of metadata, and when the internet connection is unstable. VOSviewer can extract metadata from the manager's reference file. Still, only RIS files can contain almost all metadata information than other reference files, including EndNote and RefWorks files that can be read by VOSviewer. A corpus file used for VOSviewer is a collection

of structured texts from scientific publications' metadata. For example, the text comes from the title, abstract, or the author's keyword. The score file is the H-index score data and publication year used by VOSviewer for calculation of relevance and visualization.

Some publishers like Elsevier and Springer provide text data mining API facilities to their repositories, but most of them don't. At this time, there are several tools for text data mining. Still, most are available in python or R scripts, which are more widely used for individual interests and require programming knowledge from researchers to use them.

Almost all universities in Indonesia require graduate students to use literature from Scopus indexed journals at least the fourth quartile rank available in ScimagoJR from 1999. It's had their difficulties in finding relevant scientific publications and levels in ScimagoJR, especially when looking for so much scientific publication metadata.

From the user's needs and the availability of scientific publications metadata from the publisher, how the search engines development proposed: users enter search keywords, limitations on the year of publication, and filtering journal rankings that contain articles; the data mining system uses the API text data mining facility from the publisher, and raw data mining from the publisher that does not provide the text data mining facility; clean up the data obtained; store cleaned data in a database for advanced development needs; display metadata of search results through the user's web browser; and prepare RIS files, corpus files, score files, DOI file lists, and HTML files that can be downloaded by users. Figure 2 displays the flowchart of how search engines work at an early stage.

2.2.2. Iterative Design and Development

In the early stage of search engine development, it was done on a local server using PHP as the backbone of the webserver, CSS, and JavaScript for user interface design. Figure 3 and Figure 4 are the results of the initial development stage.

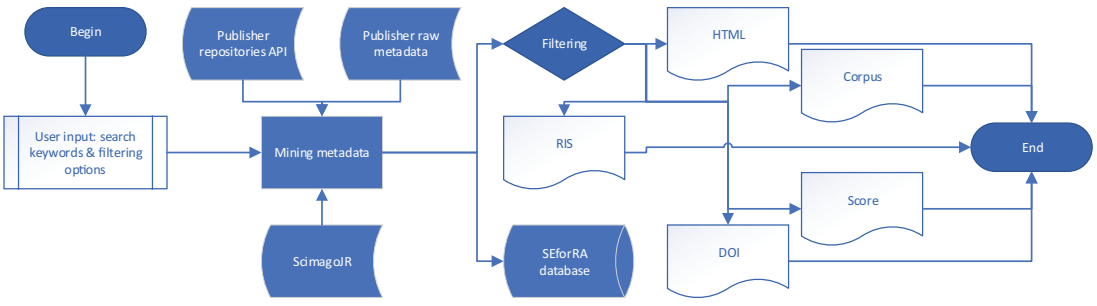

Figure 2: Early-stage flowchart

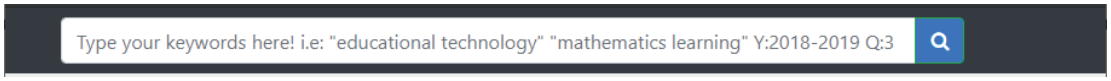

Figure 3: Early-stage user interface

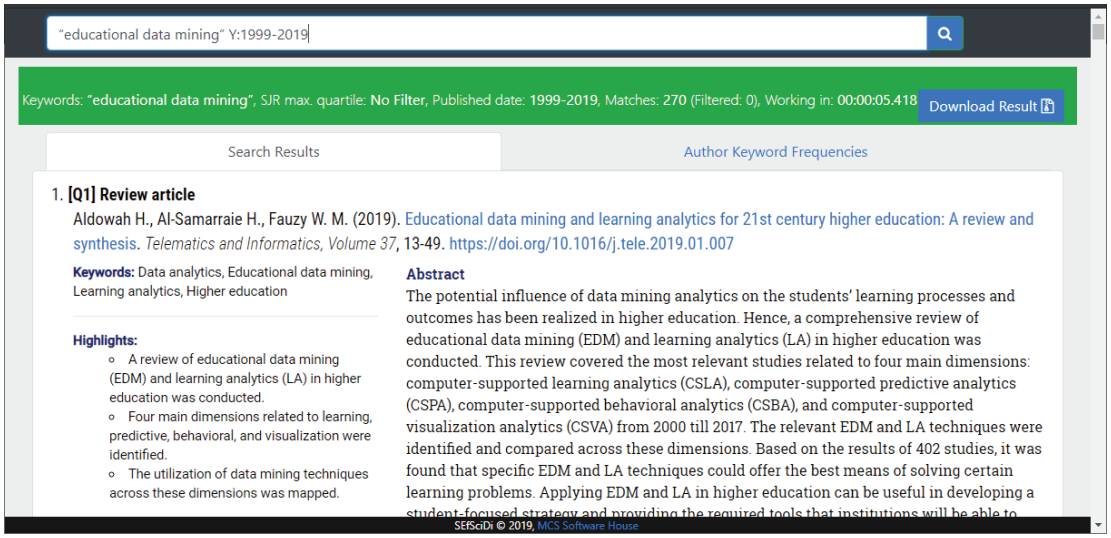

Figure 4: Early-stage search results

Search engines developed at an early stage present the data needed for analysis using the bibliometric method, but there are some fundamental disadvantages: execution time that exceeds 30 seconds as a standard applied by most shared hosting providers because it mines much metadata at one time; vulnerable to JavaScript injection because it doesn't use the HTTPS protocol; only able to handle a few concurrent users; IP Address suspended by the publisher system because it mines many data continuously.

Based on initial development stage results and user feedback, improvement at this stage is: mining metadata from Crossref supplemented by results of mining metadata from the publisher; save data from ScimagoJR in SEforRA's internal database; only retrieve metadata from the publisher when complete metadata is not available in SEforRA's internal database; mining data incrementally using AJAX, so it does not exceed the execution time limit; users can continue searching more without repeating from the start. Figure 5 is the final development flowchart.

3. Results and Evaluation

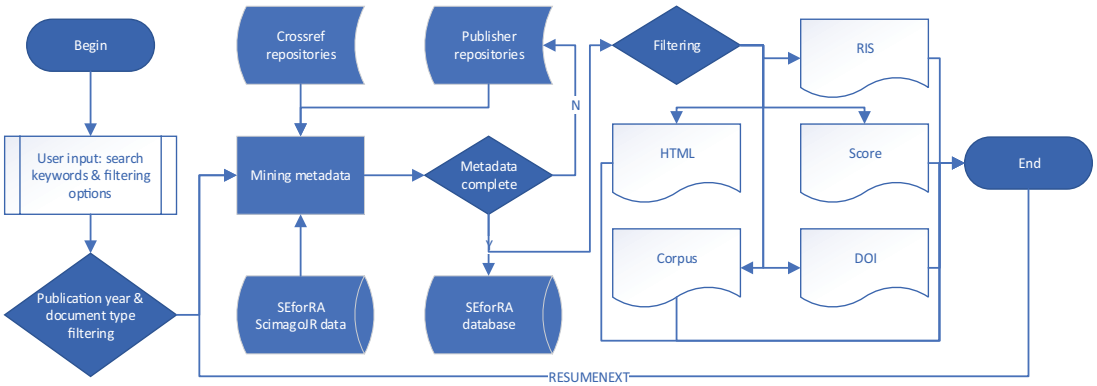

Figure 5: Final development stage flowchart

3.1. Results

The results of the final stage of SEforRA development present features: The scope of all research publication articles and books recorded at Crossref (has DOI); The identity of the publication follows the APA citation style; Abstracts of all research publication articles and books filed have abstracts on Crossref and some articles from journals published by Elsevier and Springer; Keywords from some articles from journals published by Elsevier and Springer; The author country of most articles from journals published by Elsevier and Springer; Publisher of all research publication articles and books recorded at Crossref; Number of references from all research publication articles and books filed at Crossref; Number of references from all research publication articles and books recorded at Crossref; Journal H-Index indexed by Scopus; Journal Ranking Quartiles indexed by Scopus from 1999 to 2018 (Scimagojr); The method of sorting results is based on the SEforRA relevance score, the relevance score of Crossref, and the H-Index; Limitation of publication year based on a range of years or year; Filtering of search results based on Scopus Journal (Scimagojr) Quartile Data; Search results are displayed in a web browser; Export data for offline archives in an HTML file; Export H-Index score data in TXT file; Export of corpus text data includes title, abstract and keywords data in a TXT file; Export data to citation manager software in RIS files; Export DOI data in TXT file; Resume more searches based on the search ID in less than 48 hours from searching with the same search ID; Continue interrupted searches based on search ID.

Starting from 25 August 2019, SEforRA's final development results can now be accessed by academics at https://seforra.com. The user interfaces for mobile and desktop browsers can be seen in Figure 6, search results in Figure 7, and exported data ready for bibliometric in Figure 8. Based on SEforRA usage data obtained from

Google Analytics (Figure 9) shows an increase in the use of SEforRA with mobile users by 25.7% and desktop users by 74.3%.

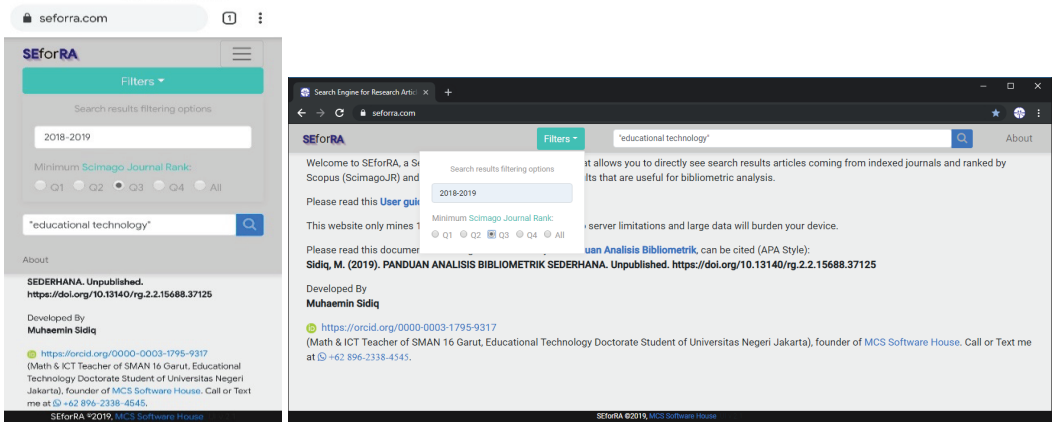

Figure 6: SEforRA user interface

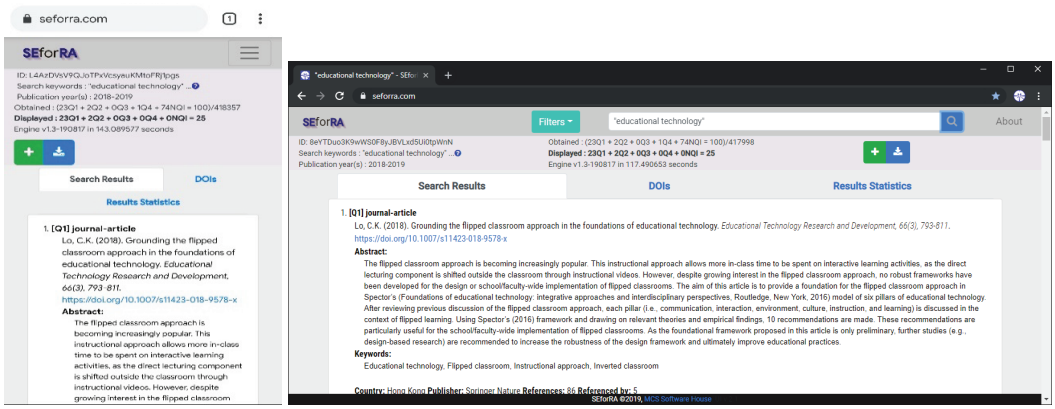

Figure 7: SEforRA search results

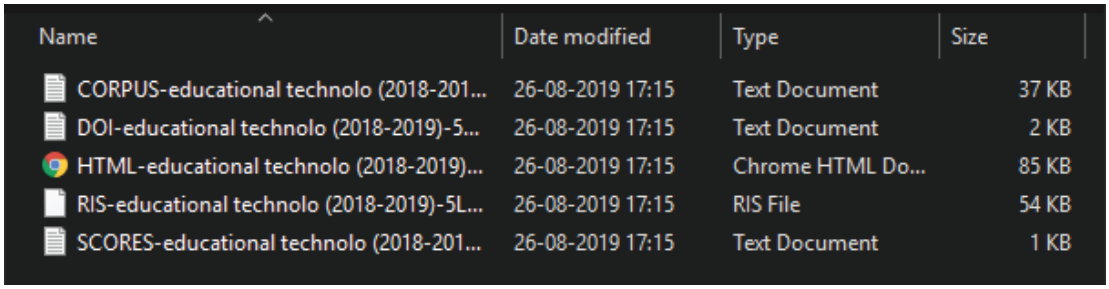

Figure 8: SEforRA exported data ready for bibliometric

3.2. Evaluation

SEforRA can collect a lot of meta-data, but when using shared hosting services, there is a limited time of execution, the use of processors and memory that ultimately limits

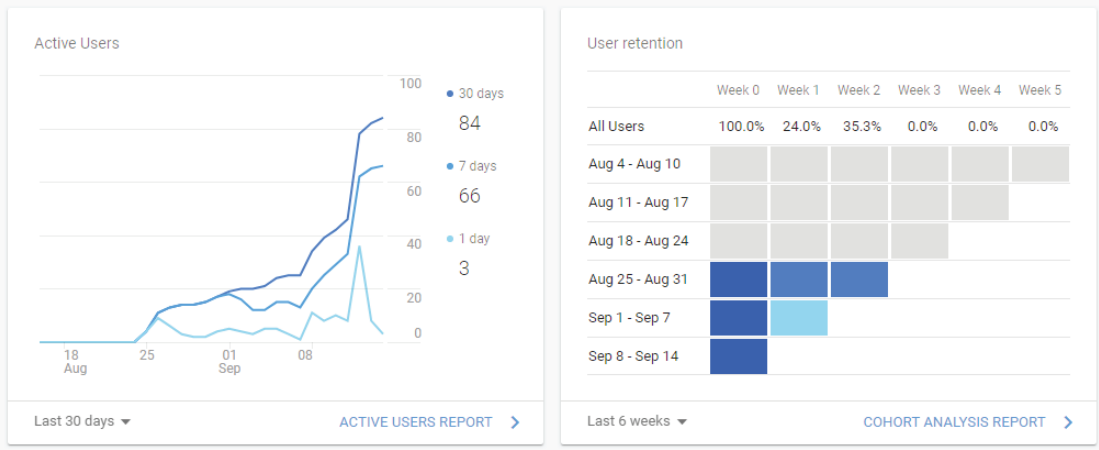

Figure 9: SEforRA user analytics

SEforRA collecting metadata little by little, it would be better and convenient to use if SEforRA uses a specialized server.

To evaluate the system, SEforRA uses Google Analytics and an internal evaluation system ti minimize various weaknesses as soon as possible. So, SEforRA does not stop developing and will continue to evolve as long as the bibliometric method, and the search for academic literature can keep forever. To further introduce SEforRA to the educational environment, SEforRA is indexed in the Google search engine by using Google Search Engine Optimization. Socialization using social media can also be an alternative to socialization through seminars, conferences, and workshops.

4. Discussion

Google Scholar is currently still the most widely used academic search engine for searching scientific publications [32]. However, Google Scholar search results not export-able for analysis using the bibliometric method. SEforRA can be a metadata search engine for scientific publications that is ready to be analyzed using the bibliometric method. Because it can be used freely by academics, SEforRA continues to grow according to user needs. Google, in September 2018, introduced Google Dataset Search, which helped researchers locate online data that are freely available for use. Experts say that it fills a gap and could contribute significantly to the success of the open-data movement, which aims to make data freely available for use and re-use [33]. But the search results are not ready for bibliometric analysis, because they do not provide data storage features for search results such as WoS or SEforRA. Dimension provides complete scientific publication data that is ready for scientometrics [34, 35]. Still, the search results for bibliometric analysis can not be displayed directly like

SEforRA, 2,500 data is prepared and then sent to the user's email. WoS and Scopus provide more sophisticated tools for measuring trends of scholarly publications [8], but not all researchers in Indonesia can access paid services WoS and Scopus.

5. Conclusion

The API metadata repository facility for scientific publications will facilitate the dissemination of research results because the facility enables faster and structured data mining, which will only require a short time when needed by researchers.

The web-based scientific publication search engine is an alternative for a system that requires high interoperability, must also be equipped with a user-friendly user interface for mobile and desktop users.

A two-phase SAM is appropriate for development projects that involve limited human resources because it can adapt to the needs of users in small development stages, does not require expensive costs and development time is relatively shorter than other methods.

The use of SEforRA as a bibliometric-ready academic search engine has proven to be able to collect a lot of metadata in a short time, so it can help analysis using bibliometric methods faster, and researchers can more concentrate on critical review activities based on bibliometric analysis.

For further research, it is advisable to look for more effective and efficient database query methods, develop metadata compression, and review the effectiveness of SEforRA based on the history of its use.

Acknowledgment

The authors would like to thank their colleagues for their contribution and support to the research. They are also thankful to all the reviewers who gave their valuable inputs to the manuscript and helped in completing the paper.

Conflict of Interest

The authors have no conflict of interest to declare.

References

[1] Wong, C. Y. (2019). A Century of Scientific Publication: Towards a Theorization of Growth Behavior and Research-Orientation. *Scientometrics*, vol. 119, issue 1, pp. 357–377.

[2] Nordtveit, B. H. (2019). Scholarly Publication as Scientific Knowledge Production: Vision of the Editors Versus Review by Peers. *Comparative Education Review*, vol. 63, issue 3, pp. 309–314.

[3] Leeder, S. (2019). The IJE and the Volatile World of Academic Publication. *International Journal of Epidemiology*, vol. 48, issue 2, pp. 323–331.

[4] Attyé, A. (2019). Data Sharing Improves Scientific Publication: Example of the "Hydrops Initiative". *European Radiology*, vol. 29, issue 4, pp. 1959–1960.

[5] Li, K., Rollins, J. and Yan, E. (2018). Web of Science Use in Published Research and Review Papers 1997–2017: A Selective, Dynamic, Cross-domain, Content-based Analysis. *Scientometrics*, vol. 115, issue 1, pp. 1–20.

[6] Ellegaard, O. and Wallin, J. A. (2015). The Bibliometric Analysis of Scholarly Production: How Great is the Impact? *Scientometrics*, vol. 105, issue 3, pp. 1809–1831.

[7] Walters, W. H. (2016). Information Sources and Indicators for the Assessment of Journal Reputation and Impact. *Reference Librarian*, vol. 57, issue 1, pp. 13–22.

[8] Cavacini, A. (2015). What is the Best Satabase for Computer Science Journal Articles? *Scientometrics*, vol. 102, issue 3, pp. 2059–2071.

[9] Lezama-Nicolás, R., *et al*. (2018). A Bibliometric Method for Assessing Technological Maturity: The Case of Additive Manufacturing. *Scientometrics*, vol. 117, issue 3, pp. 1425–1452.

[10] Park, I. and Yoon, B. (2018). Identifying Promising Research Frontiers of Pattern Recognition through Bibliometric Analysis. *Sustainability (Switzerland)*, vol. 10, issue 11.

[11] Wong, D. (2018). VOSviewer. *Technical Services Quarterly*, vol. 35, issue 2, pp. 219–220.

[12] Lammey, R. (2019). How Publishers Can Work with Crossref on Data Citation. *Science Editing*, vol. 6, issue 2, pp. 166–170.

[13] Pentz, E. (2019). CrossRef: The Missing Link. *College & Research Libraries News*, vol. 62, issue 2, pp. 206–228.

[14] Fairhurst, V. (2018). The International Reach of Crossref. *Science Editing*, vol. 5, issue 1, pp. 62–65.

[15] Harzing, A. W. (2019). Two New Kids on the Block: How do Crossref and Dimensions Compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science?. *Scientometrics*, vol. 120, issue 1, pp. 341–349.

[16] Aujla, H., *et al*. (2019). The Semantic Librarian: A Search Engine Built from Vector-Space Models of Semantics. Behavior Research Methods.

[17] Meddings, K. (2017). *REST API: Give and Get Metadata that's Open and Useful*. Crossref.

[18] Rochim, A. F., Muis, A. and Sari, R. F. (2018). Improving Fairness of H-index: RA-index. *DESIDOC Journal of Library and Information Technology*, vol. 38, issue 6, pp. 378–386.

[19] Cann, D. (2019). The Other H-index: the Hyperbole Index. *Journal of Materials Science*, vol. 54, issue 18, pp. 11757–11758.

[20] Jorgensen, M. (2019). Relationships between Project Size, Agile Practices, and Successful Software Development: Results and Analysis. *IEEE Software*, vol. 36, issue 2, pp. 39–43.

[21] Rawat, S., Goyal, N. and Ram, M. (2017). Software Reliability Growth Modeling for Agile Software Development. *International Journal of Applied Mathematics and Computer Science*, vol. 27, issue 4, pp. 777–783.

[22] Ebert, C. and Paasivaara, M. (2017). Scaling Agile. *IEEE Software*, vol. 34, issue 6, pp. 98–103.

[23] Meyer, B. (2018). Making Sense of Agile Methods. *IEEE Software*, vol. 35, issue 2, pp. 91–94.

[24] Surendra, N. C. and Nazir, S. (2019). Creating "Informating" Systems using Agile Development Practices: An Action Research Study. *European Journal of Information Systems*, vol. 00(00), pp. 1–17.

[25] Cram, W. A. and Marabelli, M. (2018). Have your Cake and Eat it Too? Simultaneously Pursuing the Knowledge-sharing Benefits of Agile and Traditional Development Approaches. *Information and Management*, vol. 55, issue 3, pp. 322–339.

[26] Dingsoeyr, T., Falessi, D. and Power, K. (2019). Agile Development at Scale: The Next Frontier. *IEEE Software*, vol. 36, issue 2, pp. 30–38.

[27] Allen, M. W. and Merrill, M. D. (2018). SAM and Pebble-in-the-Pond: Two Alternatives to the ADDIE Model. In R. A. Reiser and J. V. Dempsey (Eds.), *Trends and Issues in Instructional Design and Technologies*. Pearson, pp. 31–41.

[28] Richard-Foy, J., Barais, O. and Jézéquel, J. M. (2013). Efficient High-level Abstractions for Web Programming. Presented at the *12th International Conference on Generative Programming: Concepts and Experiences*, pp. 53–60.

[29] Allen, M. W. (2018). The Successive Approximation Model (SAM): A Closer Look. In Robert A. Reiser and J. V. Dempsey (Eds.), *Trends and Issues in Instructional Design and Technologies* (4th ed.). Pearson Education, Inc., pp. 42–51.

[30] van Eck, N. J. and Waltman, L. (2010). Software Survey: VOSviewer, A Computer Program for Bibliometric Mapping. *Scientometrics*, vol. 84, issue 2, pp. 523–538.

[31] van Eck, N. J. and Waltman, L. (2019). *VOSviewer Manual: Manual for VOSviewer version 1.6.12*. Centre for Science and Technology Studies, Leiden University.

[32] Gusenbauer, M. (2019). Google Scholar to Overshadow them all? Comparing the Sizes of 12 Academic Search Engines and Bibliographic Databases. *Scientometrics*, vol. 118, issue 1

[33] Castelvecchi, D. (2018). Google Unveils Search Engine for Open Data. *Nature*, vol. 561, issue 7722, pp. 161–162.

[34] Van Noorden, R. (2018). Science Search Engine Links Papers to Grants and Patents. *Nature*.

[35] Thelwall, M. (2018). Dimensions: A Competitor to Scopus and the Web of Science? *Journal of Informetrics*, vol. 12, issue 2, pp. 430–435.