

Robust Clustering of Open Access Journal Based on Scopus Journal Metrics Database

Rizki Agung Wibowo^{1)*} , Khoirin Nisa²⁾ , Amril Samosir³⁾ 

¹⁾³⁾Department of Management, Faculty of Economics and Management, Malahayati University, Lampung, Indonesia

²⁾Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Lampung, Lampung, Indonesia

^{1)*}rizkiagungw@malahayati.ac.id, ²⁾khoirin.nisa@fmipa.unila.ac.id, ³⁾amril@malahayati.ac.id

Abstract

Background: Open-access is free online access to articles, journal, conferences proceedings, book series and trade journal which provides unrestricted and permit the users to read, download, print, copy and link to the articles. Many articles that discuss the journal metrics using basic statistical methods to describe the journal.

Objective: This research groups journals based on numerical quality measures, identifying quality characteristics for each group. The findings provide a reference for researchers to select suitable journals and for journal owners to improve journal quality.

Methods: There is another method to describe the open-access journal by grouping it into groups with the homogeneous characteristics based on five types of numerical quality measure that are analyzed simultaneously, namely cluster analysis. By using cluster analysis, the article's owner can determine which journals he can choose to publish it in according to the desired journal quality. Based in the result, 5146 open-access journals can be divided into four clusters by using CLARA algorithm. Cluster 1, 2 and 3 have high characteristics in all numerical quality measure and cluster 4 have low characteristics in all numerical quality measure. So that researchers can choose journals in clusters 1, 2, and 3 as a place to publish their research by adjusting the journal's scope.

Results: This study demonstrates that the CLARA algorithm successfully grouped 5146 open-access journals indexed by SCOPUS into four clusters based on quality characteristics. Cluster 1 consists of 39 journals with high values across all quality variables, Cluster 2 includes 50 journals with similarly high values, Cluster 3 contains 430 journals with comparable characteristics, and Cluster 4, comprising 4627 journals, exhibits low values in all quality variables. Furthermore, the majority of journals (89.914%) have numerical quality measures below the average.

Conclusion: This study concludes that journals in Clusters 1, 2, and 3 can be recommended as suitable options for researchers to publish their work, considering the relevance of the journal's scope. Additionally, these findings can serve as a reference for journal owners to improve the quality of their journals to meet higher standards.

Keywords: Open-Access; SCOPUS, Robust, Clustering, CLARA

INTRODUCTION

In this modern era, everyone has a huge opportunity to obtain a lot of information, especially those that can be used in terms of research. We can obtain information through many platforms, one of which is scientific journals. Many journals in this world provide a variety of information, some are open-access and some are paid. Open-access journals are very beneficial

* Corresponding Author

Cite this article: Wibowo, R.A., Nisa, K., & Samosir, A. (2024). Robust Clustering of Open Access Journal Based on Scopus Journal Metrics Database. *Lentera Pustaka: Jurnal Kajian Ilmu Perpustakaan, Informasi dan Kearsipan*, 10(2), 107-116. <http://doi.org/10.14710/lenpust.v10i2.68282>

for readers because they do not need to pay to access the information in the journal, so it has a great chance of being read by many people around the world.

Open-access is free online access to articles, journal, conference proceedings, book series, and trade journal which provides unrestricted and permit the users to read, download, print, copy, and link to the articles without any financial, and lawful barriers (J. et al., 2021), so it has a great chance of being read by many people around the world. Open-access is very helpful for researchers.

The quality of a journal needs to be considered by readers and researchers who will publish in the journal. One of the platforms that assess the quality of a journal is Scopus. Scopus is an abstract and indexing database that is produced by Elsevier Co. (Burnham, 2006; Puspita, 2021). It includes four types of sources: journal, conference proceedings, book series and trade journal. All journal indexed by Scopus are reviewed according to types of quality measure (Puspita, 2021), Through the Scopus database, we can obtain information related to the types of numerical quality measures used by it to assess the quality of a journal such as H-Index (HI), Citation count (CC), CiteScore (CS), SJR (SCImago Journal Rank) and SNIP (Source Normalized Impact per Paper). In this study citation count is also included.

CiteScore measures average citations received per peer-reviewed document published in the serial, Citation received in a range of 4 years for the documents published in same 4 years, SJR measure weighted citations received by the serial, and SNIP measure actual citations received relative to citation expected for the serial's subject field (*Scopus Preview - Scopus - Sources*, n.d.; Vijayan et al., 2021). In 2020 based on dataset provided by Scopus, globally there are 5197 title open-access including 5146 (99.02%) journal, 13 (0.25%) conference proceedings, 32 (0.62%) book series and 6 (0.12%) trade journal that indexed by Scopus, the data visualized in Figure 1 below.

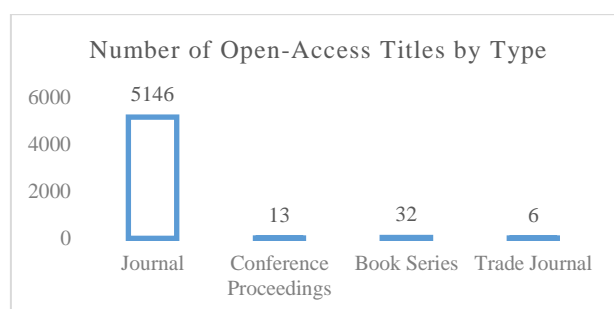


Figure 1. Number of open-access titles by type

Especially in open-access journals, we can do a simple descriptive analysis of the types of numerical quality measures, for the example, Table 1 enumerates the top citation count of ten open-access journal indexed by Scopus

TABLE 1
TOP 10 CITATION COUNT OF OPEN-ACCESS JOURNAL INDEXED BY SCOPUS

No.	Journals	Citation Count
1.	Scientific Reports	591671

Cite this article: Wibowo, R.A., Nisa, K., & Samosir, A. (2024). Robust Clustering of Open Access Journal Based on Scopus Journal Metrics Database. *Lentera Pustaka: Jurnal Kajian Ilmu Perpustakaan, Informasi dan Kearsipan*, 10(2), 107-116. <http://doi.org/10.14710/lenpust.v10i2.68282>

2.	Nature Communications	417503
3.	PLoS ONE	372024
4.	ACS applied materials & interfaces	288151
5.	Science	226134
6.	Proceedings of the National Academy of Sciences of the United States of America	224103
7.	Journal of Cleaner Production	203300
8.	IEEE Access	201619
9.	Physical Review Letters	162453
10.	The Lancet	147190

Table 2 enumerates the top CiteScore of ten open-access journal indexed by Scopus

TABLE 2
TOP 10 CITESCORE OF OPEN-ACCESS JOURNAL INDEXED BY SCOPUS

No.	Journals	CiteScore
1.	The Lancet	91.5
2.	Living Reviews in Relativity	67.4
3.	IEEE Communications Surveys and Tutorials	62.1
4.	Science	46.8
5.	Molecular Cancer	34.3
6.	The Lancet Global Health	32.1
7.	Living Reviews in Solar Physics	30.8
8.	Annals of the Rheumatic Diseases	28.7
9.	The Lancet Public Health	28.7
10.	Studies in Mycology	24.9

Table 3 enumerates the top SNIP of ten open-access journal indexed by Scopus

TABLE 3
TOP 10 SNIP OF OPEN-ACCESS JOURNAL INDEXED BY SCOPUS

No.	Journals	SNIP
1.	The Lancet	23.639
2.	Foundations and Trends in Finance	11.522
3.	IEEE Communications Surveys and Tutorials	11.397
4.	The Lancet Global Health	10.022
5.	Living Reviews in Solar Physics	8.950
6.	Living Reviews in Relativity	8.893
7.	H2Open Journal	8.713
8.	Science	7.789
9.	Journal of Statistical Software	7.237
10.	The Lancet Public Health	7.171

Table 4 enumerates the top SJR of ten open-access journal indexed by Scopus

TABLE 4
TOP 10 SJR OF OPEN-ACCESS JOURNAL INDEXED BY SCOPUS

No.	Journals	SJR
1.	The Lancet	13.103
2.	Science	12.556
3.	Living Reviews in Relativity	11.551
4.	Foundations and Trends in Finance	9.231
5.	Genome Biology	9.027
6.	Nucleic Acids Research	9.008
7.	Molecular Systems Biology	8.523

Cite this article: Wibowo, R.A., Nisa, K., & Samosir, A. (2024). Robust Clustering of Open Access Journal Based on Scopus Journal Metrics Database. *Lentera Pustaka: Jurnal Kajian Ilmu Perpustakaan, Informasi dan Kearsipan*, 10(2), 107-116. <http://doi.org/10.14710/lenpust.v10i2.68282>

8	The Lancet Global Health	7.970
9	Physical Review X	7.940
10	Journal of Statistical Software	7.636

And last table is Table 5 enumerates the top H-Index of ten open-access journal indexed by Scopus

TABLE 5
TOP 10 H-INDEX OF OPEN-ACCESS JOURNAL INDEXED BY SCOPUS

No.	Journals	H-Index
1.	Proceedings of the National Academy of Sciences of the United States of America	805
2.	Physical Review Letters	647
3.	Nucleic Acids Research	569
4.	Journal of Biological Chemistry	528
5.	Nature Communications	410
6.	Biomaterials	397
7.	NeuroImage	381
8.	Neurology	378
9.	PLoS ONE	367
10.	Journal of Chemical Physics	361

The five example above are very simple descriptive analysis of each types of numerical quality measures. There is another technique to describe the open-access journal by grouping it into groups with the homogeneous characteristics based on five types of numerical quality measure that are analyzed simultaneously , namely cluster analysis.

Cluster analysis is a statistical technique for finding groups of objects from multivariate data. The aim of cluster analysis is to construct groups with homogeneous properties out of heterogeneous large samples (Hair et al., 2014). Clustering algorithms are designed to identify an underlying structure of data and use the detected relationships within the structure to group the objects into distinct groups. One of the most commonly used algorithms among the partitioning methods in cluster analysis is the K-means algorithm (Shang et al., 2021; Wu et al., 2021). K-means starts by assigning K initial cluster centroids, either randomly or by an initialization algorithm. All objects are distributed into each cluster based on their distance to the centroids. The solution is refined by first electing a new cluster centroid, based on the mean values of each object in the cluster, and then redistributing the object accordingly.

K-means refines the solution until changes are no longer made or until a maximum limit of iterations has been reached (Suharjo & Utama, 2021). However, K-means is very sensitive to outliers. Even one outlier can affect the result of K-means clustering (Devi & Kaur, 2014; Jin & Han, 2017). Therefore, a robust cluster algorithm is needed when we deal with data containing outliers. One of the efficient robust clustering techniques is the K-medoids, there are several kinds of K-medoids algorithms, namely CLARA (Clustering Large Application) which uses the sampling approach to handle the large dataset (Gupta et al., 2019; Gupta & Panda, 2019). The CLARA algorithm was introduced in 1990 by Kaufman and Rousseeuw (Schubert & Rousseeuw, 2019).

Cite this article: Wibowo, R.A., Nisa, K., & Samosir, A. (2024). Robust Clustering of Open Access Journal Based on Scopus Journal Metrics Database. *Lentera Pustaka: Jurnal Kajian Ilmu Perpustakaan, Informasi dan Kearsipan*, 10(2), 107-116. <http://doi.org/10.14710/lenpust.v10i2.68282>

This research will be conducted to group journals based on types of numerical quality measures into groups of journals that will be given characteristics of quality, so that it can be used as a reference for readers and researchers to choose journals according to the characteristics they want and for journal owners can be used as a reference to improve the quality of their journals.

METHODS

This research was conducted in May 2024. The data used in this research is open-access journals sourced from Scopus database website, with variables used are five types of numerical quality measure, namely H-Index (HI), Citation count (CC), CiteScore (CS), SJR (SCImago Journal Rank) and SNIP (Source Normalized Impact per Paper), there are 5146 open-access journal. Because there are 5146 open-access journal, so the CLARA algorithm is used to construct groups with homogeneous properties. This study used cluster R Package to perform CLARA. The CLARA algorithm can be found at (Gupta et al., 2019; Gupta & Panda, 2019; Schubert & Rousseeuw, 2019).

Processed resulted are analyzed using Rstudio software. The analysis procedure can be described as follow:

- a) Outliers' detection by using the Mahalanobis distance:

$$d_i(\mathbf{x}_i, \bar{\mathbf{x}}) = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})} \quad (1)$$

\mathbf{x}_i is an outlier if $d_i^2(\mathbf{x}_i, \bar{\mathbf{x}}) > \chi_{p,1-\alpha}^2$, where p is the number of variables and α is a significance level with the default cut off commonly used is $\alpha = 0,05$ (Larasati et al., 2021; Venelia et al., 2021)

- b) Clustering the open-access journal using CLARA algorithm based on five variables with number of clusters (k) is 4 and assign journals to each homogeneous cluster
- c) Evaluate the cluster result by using silhouette width. The silhouette width value ranges from -1 to 1, if it is close to the value of 1, it can be said that the members of each cluster are well grouped, if it is close to the value of -1, it can be said that the members of each cluster are poorly grouped. it was published in (Brock et al., 2008) .

RESULTS AND DISCUSSION

The first step is conducted outlier detection using robust squared of Mahalanobis distance and the result is shown in Figure 2.

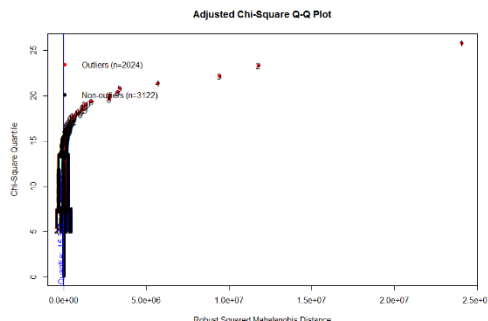


Figure 2. Outlier's detections

The result shows that the number of outliers in the data is 2024 journals, so a robust clustering algorithm is required using the CLARA algorithm. Using R, the CLARA algorithm's result for four clusters yields a silhouette width value of as much as 0.82 because it is close to 1, which means the members of each cluster are well grouped. The resulting clusters are shown graphically in Figure 3.

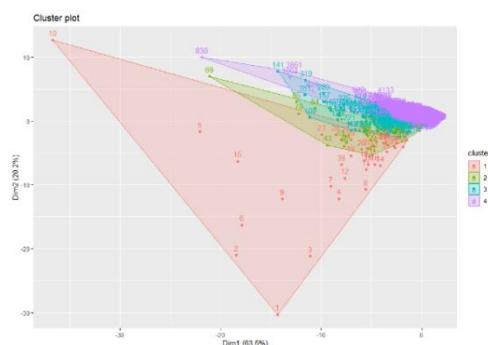


Figure 3. Cluster plot result

To determine cluster characteristics based on each variables and paid more attention to in each cluster, it is necessary to carry out a descriptive analysis (Larasati et al., 2021). From all open-access journal in the data, the average value of each variable is $\bar{X}_{HI} = 29.677$, $\bar{X}_{CC} = 2304.604$, $\bar{X}_{CS} = 2.621$, $\bar{X}_{SJR} = 0.585$ and $\bar{X}_{SNIP} = 0.886$. These average values are compared to the cluster centers ($\bar{X}_{c,j}$), if $\bar{X}_{c,j} \leq \bar{X}_j$ (where j is the variable: HI, CC, CS, SJR, SNIP; c is cluster: 1, 2, 3, 4) it is interpreted as "low", if $\bar{X}_{c,j} > \bar{X}_j$, it is interpreted sequentially as "high". Details of the result are described in Table 6 and Table 7.

TABLE 6
 CLUSTER CENTRE ($\bar{X}_{c,j}$)

Cluster	HI	CC	CS	SJR	SNIP
1	220.256	134135.718	12.454	2.774	2.420
2	165.840	30647.280	10.260	2.235	1.841
3	87.165	7200.742	6.684	1.626	1.568
4	21.257	432.140	2.078	0.452	0.800

Cite this article: Wibowo, R.A., Nisa, K., & Samosir, A. (2024). Robust Clustering of Open Access Journal Based on Scopus Journal Metrics Database. *Lentera Pustaka: Jurnal Kajian Ilmu Perpustakaan, Informasi dan Kearsipan*, 10(2), 107-116. <http://doi.org/10.14710/lenpust.v10i2.68282>

TABLE 7
 CLUSTER CHARACTERISTICS

Cluster	HI	CC	CS	SJR	SNIP	Cluster Size
1	High	High	High	High	High	39 (0.758%)
2	High	High	High	High	High	50 (0.972%)
3	High	High	High	High	High	430 (8.356%)
4	Low	Low	Low	Low	Low	4627 (89.914%)

TABLE 8
 10 OPEN-ACCESS JOURNAL MEMBERS PER CLUSTER

Cluster	Journals
1	Scientific Reports, Nature Communications, PLoS ONE, ACS applied materials & interfaces, Science, Proceedings of the National Academy of Sciences of the United States of America, Journal of Cleaner Production, IEEE Access, Physical Review Letters, The Lancet
2	BMJ Open, Journal of Chemical Physics, Materials and Design, Biomaterials, Optics Letters, Medicine (United States), NeuroImage, Cell Death and Disease, Frontiers in Psychology, Frontiers in Pharmacology
3	Advances in Experimental Medicine and Biology, PLoS Genetics, Frontiers in Oncology, European Radiology, Journal of Medical Internet Research, Frontiers in Neurology, PLoS Computational Biology, Arabian Journal of Chemistry, JCI insight, OncoTargets and Therapy
4	npj Quantum Information, Journal of Ophthalmology, The Lancet Planetary Health, Computational and Structural Biotechnology Journal, Ciencia e Saude Coletiva, Egyptian Journal of Petroleum, Health Expectations, Climate of the Past, Vaccines, Journal of Healthcare Engineering

Table 7 shows that the members in Cluster 1, 2, and 3 have high characteristics in all variables, while Cluster 4 has low characteristics in all variables. Based on Table 7, most open-access journal members in cluster 4 have quality measure-related conditions below the measured average, there are 4627 open-access journals, or 89.914% of all journals. Table 8 shows 10 open-access journal members per cluster and the full table can be found in following here: https://s.id/members_of_clustering_journal. So that researchers can choose journals in clusters 1, 2, and 3 as a place to publish their research by adjusting the journal's scope, and journal owners or journal editors included in cluster 4 must improve the quality of their journals.

CONCLUSIONS

In this paper, we applied the CLARA algorithm to group 5146 open-access journals indexed by SCOPUS. Based on the result and discussion it can be concluded that 5146 open-access journals can be divided into four clusters by using the CLARA algorithm. Most open-access journal members in cluster 4 have quality measure-related conditions below the measure average. Cluster 1 has high characteristics in all variables with 39 members, Cluster 2 has high characteristics in all variables with 50 members, Cluster 3 has high characteristics in all variables with 430 members and the last cluster has low characteristics in all variables with 4627 members. So that researchers can choose journals in clusters 1, 2, and 3 as a place to publish their research by adjusting the journal's scope. In addition, most open-access journals indexed by SCOPUS have a low of five types of numerical quality measure, there are 89.914%

Cite this article: Wibowo, R.A., Nisa, K., & Samosir, A. (2024). Robust Clustering of Open Access Journal Based on Scopus Journal Metrics Database. *Lentera Pustaka: Jurnal Kajian Ilmu Perpustakaan, Informasi dan Kearsipan*, 10(2), 107-116. <http://doi.org/10.14710/lenpust.v10i2.68282>

of all journals. This research can serve as a reference for journal owners to improve the quality of their journals.

AUTHOR CONTRIBUTIONS

[Rizki Agung Wibowo]: Conceptualization, methodology, writing the original draft, review and editing, supervision. [Khoirin Nisa]: Software development, investigation, data analysis, writing the original draft. [Amril Samosir]: Investigation, data curation.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no specific grant from any funding agency.

ACKNOWLEDGMENTS

We would like to acknowledge and thank all those who have given valuable contributions to this study.

REFERENCES

- Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clValid: An R Package for Cluster Validation. *Journal of Statistical Software*, 25(4 SE-Articles), 1–22. <https://doi.org/10.18637/jss.v025.i04>
- Burnham, J. F. (2006). Scopus database: A review. *Biomedical Digital Libraries*, 3(1), 1–8. <https://doi.org/10.1186/1742-5581-3-1/TABLES/2>
- Devi, P., & Kaur, K. (2014). A Robust Cluster Head Selection Method Based on K-Medoids Algorithm to Maximize Network Life Time and Energy Efficiency for Large WSNs. *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT)*, 3(5), 1430–1432.
- Gupta, T., Gupta, T., & Panda, S. P. (2019). A Comparison of K-Means Clustering Algorithm and CLARA Clustering Algorithm on Iris Dataset. *International Journal of Engineering & Technology*, 7(4), 4766–4768. <https://doi.org/10.14419/ijet.v7i4.21472>
- Gupta, T., & Panda, S. P. (2019). Clustering Validation of CLARA and K-Means Using Silhouette DUNN Measures on Iris Dataset. *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Perspectives and Prospects, COMITCon 2019*, 10–13. <https://doi.org/10.1109/COMITCON.2019.8862199>

Cite this article: Wibowo, R.A., Nisa, K., & Samosir, A. (2024). Robust Clustering of Open Access Journal Based on Scopus Journal Metrics Database. *Lentera Pustaka: Jurnal Kajian Ilmu Perpustakaan, Informasi dan Kearsipan*, 10(2), 107-116. <http://doi.org/10.14710/lenpust.v10i2.68282>

- Hair, J. F., Black, W. C., & Anderson, R. E. (2014). *Multivariate Data Analysis: Pearson New International Edition* (7th ed.). Pearson Education Limited.
- J., A., Prakash, M., & Balasubramani, R. (2021). A Study on Contribution of Open Access Journals on Robotics in Directory of Open Access Journals (DOAJ) Platform. *Department of Library and Information Science*.
- Jin, X., & Han, J. (2017). K-Means Clustering. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 695–697). Springer US. https://doi.org/10.1007/978-1-4899-7687-1_431
- Larasati, S. D. A., Nisa, K., & Herawati, N. (2021). Robust Principal Component Trimmed Clustering of Indonesian Provinces Based on Human Development Index Indicators. *Journal of Physics: Conference Series*, 1751(1), 0–8. <https://doi.org/10.1088/1742-6596/1751/1/012021>
- Puspita, A. T. (2021). Comparing between Scopus, Web of Science and Dimensions Indexation: Case of 100 Most Cited Articles on Waqf. *Journal of Islamic Economic Literatures*, 2(2).
- Schubert, E., & Rousseeuw, P. J. (2019). Faster -Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11807 LNCS, 171–187. https://doi.org/10.1007/978-3-030-32047-8_16
- Scopus preview - Scopus - Sources*. (n.d.). Retrieved April 9, 2022, from https://www.scopus.com/sources?zone=TopNavBar&origin=NO_ORIGIN_DEFINED
- Shang, R., Ara, B., Zada, I., Nazir, S., Ullah, Z., & Khan, S. U. (2021). Analysis of Simple K-Mean and Parallel K- Mean Clustering for Software Products and Organizational Performance Using Education Sector Dataset. *Scientific Programming*, 2021. <https://doi.org/10.1155/2021/9988318>
- Suharjo, B., & Utama, M. S. U. (2021). K-Means Cluster Analysis of Sex, Age, and Comorbidities in the Mortalities of Covid-19 Patients of Indonesian Navy Personnel. *JISA(Jurnal Informatika Dan Sains)*, 4(1), 17–21. <https://doi.org/10.31326/JISA.V4I1.869>
- Venelia, H., Nisa, K., Wibowo, R. A., & Muda, M. A. (2021). Robust Biplot Analysis of Natural Disasters in Indonesia from 2019 to 2021. *Jurnal Aplikasi Statistika & Komputasi Statistik*, 13(2), 61–68. <https://doi.org/10.34123/jurnalasks.v13i2.349>
- Vijayan, D. S., R, R. V, & R, A. V. (2021). Web of Science (WoS) Indexed Library and Information Science (LIS) Journals in Scopus: An Analysis. *Library Philosophy and Practice (e-Journal)*. <https://digitalcommons.unl.edu/libphilprac/6348>

Cite this article: Wibowo, R.A., Nisa, K., & Samosir, A. (2024). Robust Clustering of Open Access Journal Based on Scopus Journal Metrics Database. *Lentera Pustaka: Jurnal Kajian Ilmu Perpustakaan, Informasi dan Kearsipan*, 10(2), 107-116. <http://doi.org/10.14710/lenpust.v10i2.68282>

Wu, C., Yan, B., Yu, R., Yu, B., Zhou, X., Yu, Y., & Chen, N. (2021). K -Means Clustering Algorithm and Its Simulation Based on Distributed Computing Platform. *Complexity*, 2021. <https://doi.org/10.1155/2021/9446653>

Cite this article: Wibowo, R.A., Nisa, K., & Samosir, A. (2024). Robust Clustering of Open Access Journal Based on Scopus Journal Metrics Database. *Lentera Pustaka: Jurnal Kajian Ilmu Perpustakaan, Informasi dan Kearsipan*, 10(2), 107-116. <http://doi.org/10.14710/lenpust.v10i2.68282>