

PRACTICE PAPER

Use of Available Data To Inform The COVID-19 Outbreak in South Africa: A Case Study

Vukosi Marivate^{1,2} and Herkulaas MvE Combrink^{1,3}

¹ University of Pretoria, ZA

² Council for Scientific and Industrial Research, ZA

³ University of the Free State, ZA

Corresponding author: Vukosi Marivate (vukosi.marivate@cs.up.ac.za)

The coronavirus disease (COVID-19), caused by the SARS-CoV-2 virus, was declared a pandemic by the World Health Organization (WHO) in February 2020. Currently, there are no vaccines or treatments that have been approved after clinical trials. Social distancing measures, including travel bans, school closure, and quarantine applied to countries or regions are being used to limit the spread of the disease, and the demand on the healthcare infrastructure. The seclusion of groups and individuals has led to limited access to accurate information. To update the public, especially in South Africa, announcements are made by the minister of health daily. These announcements narrate the confirmed COVID-19 cases and include the age, gender, and travel history of people who have tested positive for the disease. Additionally, the South African National Institute for Communicable Diseases updates a daily infographic summarising the number of tests performed, confirmed cases, mortality rate, and the regions affected. However, the age of the patient and other nuanced data regarding the transmission is only shared in the daily announcements and not on the updated infographic. To disseminate this information, the Data Science for Social Impact research group at the University of Pretoria, South Africa, has worked on curating and applying publicly available data in a way that is computer readable so that information can be shared to the public – using both a data repository and a dashboard. Through collaborative practices, a variety of challenges related to publicly available data in South Africa came to the fore. These include shortcomings in the accessibility, integrity, and data management practices between governmental departments and the South African public. In this paper, solutions to these problems will be shared by using a publicly available data repository and dashboard as a case study.

Dashboard: <https://dsfsi.github.io/covid19za-dash/>, **Data repository:** <https://github.com/dsfsi/covid19za>.

Keywords: COVID-19; Data repository; Dashboard; Data management

1 Introduction

Accurate data are at the centre of mitigating risk, preventing widespread panic and sensationalism during a natural disaster. Evidence-based information obtained from accurate data is an asset, and one of the only strategic resources the public have during a crisis. The practice of sharing information to the public about the current state of things is dependent on specific data that has to be captured and shared to the public in a way that is useful, usable and desirable. During a crisis, the public needs to minimize exposure to the situation or act accordingly to provide support where needed. The contribution of this paper is a framework that substantiates the type of data employed to capture and modify shared information with the public during a crisis of a biological nature, such as the COVID-19 pandemic. This paper summarises findings (e.g., demography and neighbourhood) based on the public data repository (Marivate et al., 2020) and dashboard,¹ to support general understanding and lessons learned from the COVID-19 epidemic.

¹ <https://dsfsi.github.io/covid19za-dash/>.

The disease, COVID-19, is a severe acute respiratory syndrome (SARS) caused by the SARS-CoV-2 virus (Bai et al., 2020). The first reported case of COVID-19 was in December 2019 in Wuhan, China (Mo et al., 2020). The problem with SARS-CoV-2 is the pressure it puts on the healthcare system because of its high infection rate (Zhao et al., 2020). Globally, sudden spikes in the confirmed COVID-19 cases are severe because there are limited resources to effectively manage and treat patients in any of the current healthcare systems (Ferguson et al., 2020). As of February 2020, the World Health Organization (WHO) declared SARS-CoV-2 a pandemic- one of the first biological threats to modern society in the 21st century (Remuzzi & Remuzzi 2020). In order for the South African government to reduce the widespread of SARS-CoV-2, travelling restrictions were placed on high-risk regions and border-crossings were closed from one country to another. South Africa experienced the first COVID-19 case on the 28th of February 2020, and as of the 25th of March 2020, the number of confirmed COVID-19 cases increased to 702, despite efforts to contain the virus by putting a ban on international travel (World Health Organization 2020). The increase in confirmed cases may cause widespread panic and anxiety, which is why the public relies on good, reliable information and data, now more than ever.

Currently, information regarding the COVID-19 outbreak in South Africa is shared with the public across various platforms, of which two are most popular/most widely used. Firstly, the National Institute for Communicable Diseases (NICD) publishes an infographic that contains limited information, providing a bird's-eye view of the outbreak. This information is limited to South Africa and only reports the number of tests performed, number of confirmed cases, which regions are affected, and COVID-19 related mortality (NICD 2020). Secondly, the minister of the Department of Health (DoH) in South Africa, Zwelini Mkhize, updates the public regarding the cases as they arise on a daily basis. These updates are published sentences on the DoH website, containing some demographic information about the confirmed cases, including age, gender, travel history and mode of contraction of SARS-CoV-2 (DoH 2020). Although these sources of information are valuable, they are potentially ineffective ways of sharing information to the public regarding their usability for a variety of reasons. Amongst these is the number of different platforms a person has to navigate through to gain access to accurate data. Additionally, the format in which the data are presented is not in a computer readable format and has to undergo processing in order to be used and stored. This further complicates legibility, simplicity and accessibility of the information that is shared, a concern about South African government data that was highlighted in prior work (Marivate & Moorosi 2018). The impact of not having useful, usable and desirable information has a direct effect on management strategies and responses from the public in relation to the disease (Gonulal 2019).

To counter the aforementioned problem, the Data Science for Social Impact (DSFSI) research group at the University of Pretoria, South Africa, has developed an open repository for the data integrity of South African COVID-19 cases. DSFSI Lab members and willing volunteers are responsible for the mining, validation and storage of data related to the COVID-19 patients in South Africa. To work collectively on a project related to public data in a way that can be scaled, the DSFSI manages and consolidates the available data related to the COVID-19 cases in South Africa. Once consolidated, the data are shared in an open, publicly available repository on GitHub.com, and then linked to the dashboard (Marivate & Data Science For Social Impact Research Group 2020). On the repository, any member or user has the freedom to critique and propose new features or data to be added to the repository or dashboard. This includes data integrity issues as well as information to be added to the repository that is otherwise outdated or virtually inaccessible to the public. The workflow of this process is illustrated in **Figure 1**.

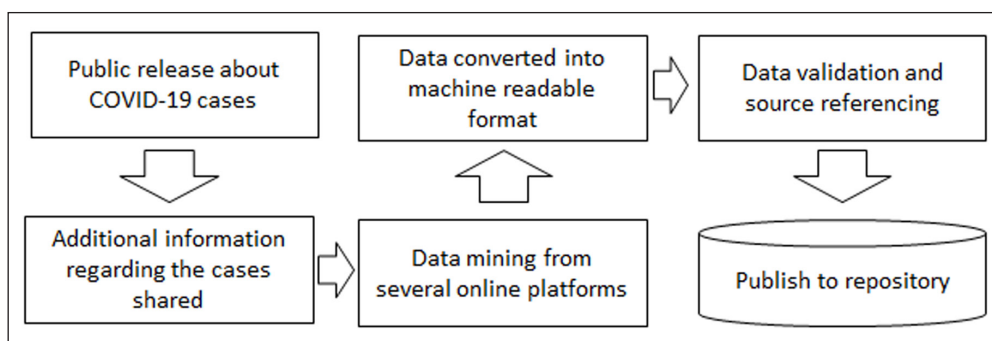


Figure 1: Data publishing cycle of COVID-19 data.

In order to link information to the COVID-19 records, data needs to be accurate and restructured. Unfortunately, when data are not updated and information changes about the state of the evidence, then it becomes difficult to track when the changes occurred, especially if the changes relate to the unique ID or name of the item. One such example relates to the publicly available South African hospital data. Some items in the hospital data were last updated more than two years prior to 2020. This includes hospital data such as coordinates, accurate contact information, and facilities available within the hospital, as well as the population size of a/the district.

Most of the aforementioned information is not consolidated in one place – a great risk factor in a time where the public needs to know where to seek care, and which hospitals are equipped to test and manage COVID-19 patients. Another example came about in a sudden spike of confirmed cases between the 23rd and 25th of March 2020. This increase of 435 cases came unaccompanied by any of the aforementioned demographic information, transmission type, nor their travel history, and these results are still pending. In subsequent days, there has been the same challenge, although the latest available data on deaths has demographic information and some travel history.

2 Aims and Methodology

The primary objectives of this study were to determine what data should be included in a public repository amidst the COVID-19 outbreak and how this data should be disseminated within a public dashboard. The public repository of data followed a Creative Commons licence for data, and MIT License for Code, with copyright for the Data Science for Social Impact research group at the University of Pretoria, South Africa (JSOUP 2020). All of the data were gathered and consolidated on the public repository which is hosted on GitHub² and uploaded on Zenodo (Marivate et al., 2020).

To determine if the dashboard (Figure 2) and data repository were used, data and analytics were performed on the basis of descriptive statistics related to the number of views, clicks, comments and recommendations on the public repository and dashboard. The use of the repository goes further than just the views, but that other researchers actually use it for analysis.

To measure whether or not the repository and dashboard were useful and desirable, the public repository allowed for the posting of issues, comments and recommendations. These items were categorised according to their submission on the repository. An item could be sorted into more than one category depending on the nature of the problem. The DSFSI and public have the opportunity to choose which problems they would like to work on, and solutions are approved by the DSFSI group.

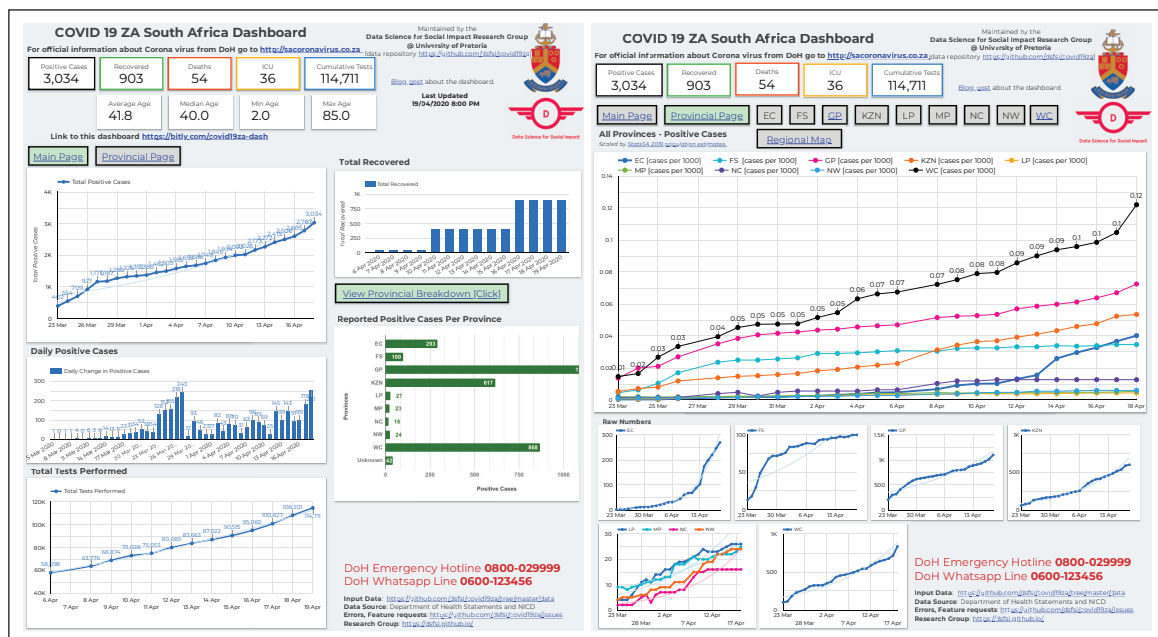


Figure 2: Consolidated dashboard using data from the repository. *Left:* Front page with aggregated national statistics. *Right:* Aggregated statistics by province.

² <https://www.github.com/dsfsi/covid19za>.

3 Results

In total, 58,169 users accessed the dashboard in the time period of 17th March to 18th April 2020 (**Figure 3**).

As seen in **Figure 3**, at least 20.4 percent of users return to the dashboard, with an average session duration of 1.35 minutes per session. In addition to this, the repository from which the data are drawn have had over 2,000 clones of the dataset from at least 200 different people. The majority of the users were from the GitHub community, but a few unique users (meaning they had not visited before) accessed the repository from other platforms (**Table 1**).

At least 404 unique visitors viewed the issues that were posted on the repository. A further 150 unique visitors viewed the pull requests. These were requests for data that members of the DSFSI and the public contributed to the actual datasets. To manage the issues that were posted, ten different labels were created to categorise the issues. Currently, only six of the ten labels were used for the issues that were posted. These were: bugs, data, enhancements, good first issue, help wanted and questions (**Table 2**).

Bugs referred to any error in the data or issue related to a feature of either the dashboard or function within the repository. The single unresolved issue in bugs related to a single incorrect data entry, but finding the source confirming the correct data proved challenging. Data were any inquiry about the data including differences between data sets, missing data, or additions to datasets. To resolve these issues, data needs to be updated from the source. Enhancement meant improvements to current implementations of either the data in the repository or dashboard. Pending information for most of these enhancements. These include

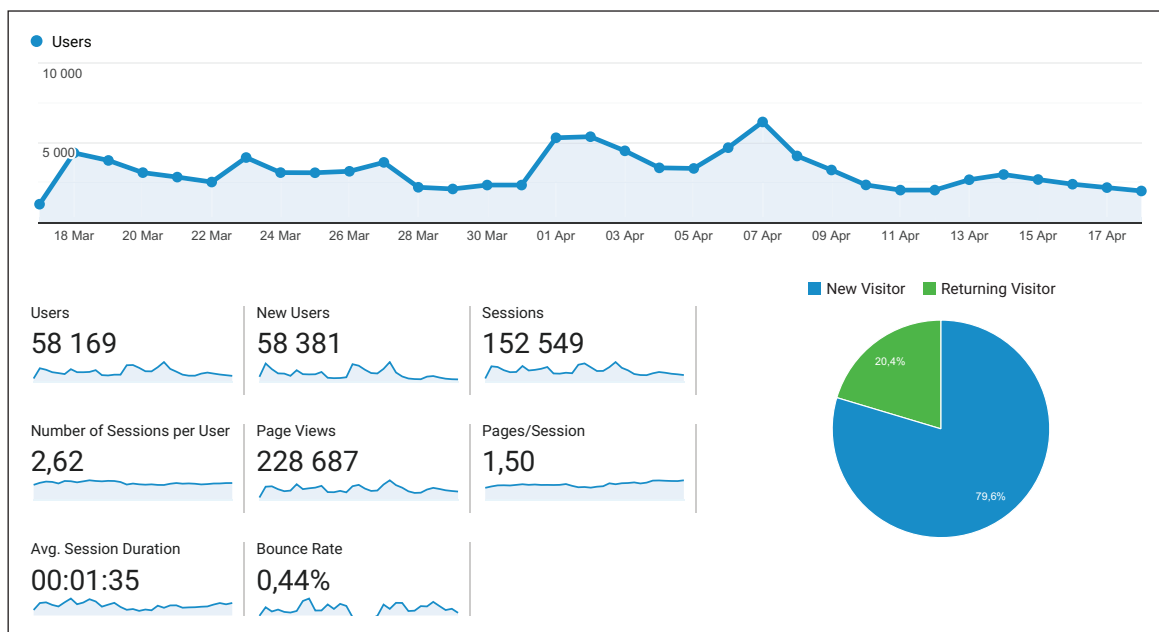


Figure 3: Data usage information from the website.

Table 1: Data usage information from the website.

Non-English or Site	Views	Unique visitors
github.com	2,229	797
twitter.com	675	431
Google.com	647	353
Linkedin.com	212	125
zendo.org	58	26
Bing.com	57	27
facebook.com	43	27
m.facebook.com	39	26
vima.co.za	20	9
Ink.in	17	8

Table 2: Categories of lodged issues.

Label	Resolved issues	Unresolved issues
Bugs	7	2
Data	19	10
Enhancement	26	17
Good first issue	2	2
Help wanted	2	1
Question	1	1

additional fields that were not provided in the publicly available data. Good first issues were entry level problems that could be completed by people from any background. These were labelled as such for newcomers to the project that did not require either a lot of time or expertise to work on. Help wanted translated to problems that require additional attention. The unresolved issues require data that are not currently available to the public in order to solve the issues. Questions were presented as general request for clarity or required more information on a particular issue in the repository that another person posted. To resolve the one unresolved question, a decision has to be made internally about the data to resolve the matter. There were more than 10,000 additions to the repository data, and 1,430 deletions, all reviewed by different members of DSFSI and accepted if they were noteworthy contributions to the repository. In addition to this, there were 26 different contributors who pushed 345 commits to all branches within the repository.

The majority of requests and changes to the repository and dashboard were associated with the data or enhancements to the data. In total, there are fifteen datasets, with seven of them related to information about hospitals in SA. Once created, the subsequent data were displayed in a dashboard (Marivate et al., 2020). Included in the dashboard were information related to COVID-19, a South African helpline, sources of the information, when last the information was updated, a blog post containing the purpose of the dashboard, links to the open public repository, and general information about the research group. Some analysis that can be accomplished with the data is shown in **Figure 4**.

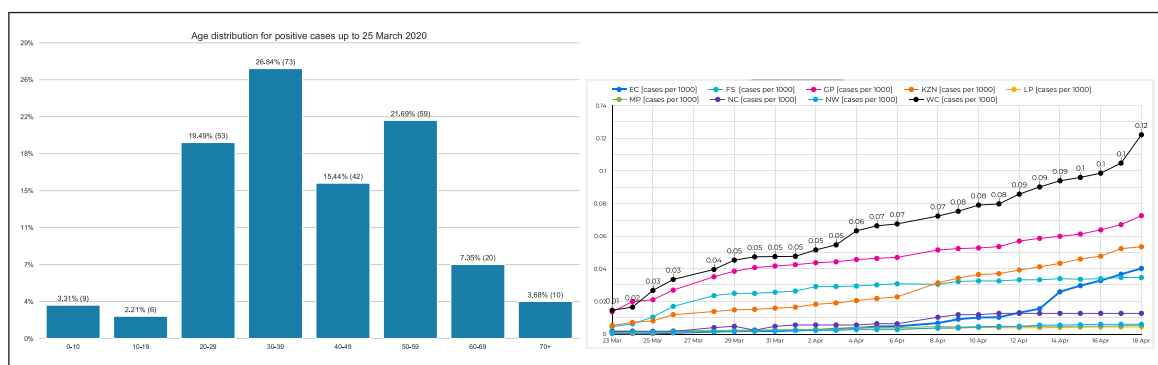


Figure 4: Examples of analysis made possible by the data repository. *Left* Age distribution of positive cases. *Right:* Provincial growth.

4 Conclusion

Data are one of the most important assets during a crisis. Unfortunately, not prioritising this commodity had complications during the early days of the COVID-19 pandemic from a South African perspective. To prevent this from happening, the DSFSI research group have started collaborating and expanding this type of methodology to create a line list for the rest of the African continent³. The data from this project led to discussion between DSFSI and the NICD and DoH, in an attempt to assist the situation. The COVID-19 pandemic showed the world the value of data, and that people and systems should effectively prepare for a time of crisis. Enabling the public to engage with data in a way that is open and collaborative is an overlooked service that can aid during a calamity. Furthermore, having readily available data is useful when needed

³ <https://www.github.com/dsfsi/covid19africa>.

during an emergency, but can seem redundant during peaceful times. Therefore, prioritising data management practices, getting input from the people who use the information, and collaborating between different organisations to gain the same result should be a proactive approach and standard, not one that is only implemented during a catastrophe.

Acknowledgements

We would like to acknowledge every person from the public that dedicated their time, effort and energy to assist during this pandemic. At the time of this publication, in alphabetical order, Alta de Waal, Jay Welsh, Nompumelelo Mtsweni, Ofentswe Lebogo, Shiven Moodley, Vutlhari Rikhotso, S'busiso Mkhondwane. As this is an open contribution project, the updated list of contributors is available on the github repo⁴. We would like to thank the DSFSI research group at the University of Pretoria for all their expertise, patience and hard work during this time. We also would like to thank all the employees of the NICD, DoH and WHO who assisted with data during this time. We would like to acknowledge ABSA for sponsoring the industry chair and it's related activities to the project.

Competing Interests

The authors have no competing interests to declare.

References

- Bai, Y, Yao, L and Wei, T.** March 2020. Presumed asymptomatic carrier transmission of COVID-19. *Jama*, E1. DOI: <https://doi.org/10.1001/jama.2020.2565>
- DoH.** 2020. *South African Department of Health*. Retrieved March 26, 2020 from <http://www.health.gov.za/>.
- Ferguson, NM, Laydon, D, Nedjati-Gilani, G, Imai, N, Ainslie, K, Baguelin, M, Bhatia, S, Boonyasiri, A, Cucunubá, Z, Cuomo-Dannenburg, G, Dighe, A, Dorigatti, I, Fu, H, Gaythorpe, K, Green, W, Hamlet, A, Hinsley, W, Okell, LC, van Elsland, S, Thompson, H, Verity, R, Volz, E, Wang, H, Wang, Y, Walker, PGT, Winskill, P, Whittaker, C, Donnelly, CA, Riley, S and Ghani, AC.** March 2020. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. London: Imperial College COVID-19 Response Team, 1–20. DOI: <https://doi.org/10.25561/77482>
- Gonulal, T.** Feb. 2019. Missing data management practices in L2 research: the good, the bad and the ugly. *Erzincan University Education Faculty Journal*, 21(1): 56–73. DOI: <https://doi.org/10.17556/erziefd.448559>
- JSOUP.** 2020. *KSOUP Licence*. Retrieved February 29, 2020 from <https://jsoup.org/license>.
- Marivate, V, de Waal, A, Combrink, H, Lebogo, O, Moodley, S, Mtsweni, N, Rikhotso, V, Welsh, J and Mkhondwane, S.** 2020. *Coronavirus disease (COVID-19) case data – South Africa*. DOI: <https://doi.org/10.5281/zenodo.3732419>
- Marivate, V and Data Science For Social Impact Research Group.** 2020. *COVID 19 ZA South Africa Dashboard*. Retrieved March 21, 2020 from <https://datastudio.google.com/u/0/reporting/1b60bdc7-bec7-44c9-ba29-be0e043d8534/page/hrUIB>.
- Marivate, V and Moorosi, N.** 2018. Exploring data science for public good in South Africa: evaluating factors that lead to success. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, 1–6. DOI: <https://doi.org/10.1145/3209281.3209366>
- Mo, P, Xing, Y, Xiao, Y, Deng, L, Zhao, Q, Wang, H, Xiong, Y, Cheng, Z, Gao, S, Liang, K, Luo, M, Chen, T, Song, S, Ma, Z, Chen, X, Zheng, R, Cao, Q, Wang, F and Zhang, Y.** March 2020. Clinical characteristics of refractory COVID-19 pneumonia in Wuhan, China. *Clinical Infectious Diseases*, 119, 103670: ciaa270. DOI: <https://doi.org/10.1093/cid/ciaa270>
- NICD.** 2020. *National Institute For Communicable Diseases (NICD)*. Retrieved March 26, 2020 from <http://www.nicd.ac.za/>.
- Remuzzi, A and Remuzzi, G.** March 2020. COVID-19 and Italy: what next? *The Lancet*, 395, 10229: 1–4. DOI: [https://doi.org/10.1016/S0140-6736\(20\)30627-9](https://doi.org/10.1016/S0140-6736(20)30627-9)
- World Health Organization.** 2020. *Coronavirus disease 2019 (COVID-19): situation report*. Retrieved March 16, 2020 from <https://apps.who.int/iris/bitstream/handle/10665/331475/nCoVsitrep11Mar2020-eng.pdf>.
- Zhao, W, Yu, S, Zha, X, Wang, N, Pang, Q, Li, T and Li, A.** March 2020. Clinical characteristics and durations of hospitalized patients with COVID-19 in Beijing: a retrospective cohort study. *MedRxiv*, 119, 103670: 1–6. DOI: <https://doi.org/10.1101/2020.03.13.20035436>

⁴ <https://github.com/dsfsi/covid19za/graphs/contributors>.


How to cite this article: Marivate, V and Combrink, HM. 2020. Use of Available Data To Inform The COVID-19 Outbreak in South Africa: A Case Study. *Data Science Journal*, 19: 19, pp. 1–7. DOI: <https://doi.org/10.5334/dsj-2020-019>

Submitted: 31 March 2020

Accepted: 21 April 2020

Published: 06 May 2020

Copyright: © 2020 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 