REALM: A Framework to Explore Research Impacts by Social Network Analysis, Bibliometrics, and Altmetrics

Luís Fernando Monsores Passos Maia
luisfmpm@ufrj.br
Graduate Program in Informatics, UFRJ
Rio de Janeiro, Brazil

Jonice Oliveira
jonice@dcc.ufrj.br
Graduate Program in Informatics, UFRJ
Rio de Janeiro, Brazil

ABSTRACT

Currently, there is a big concern of governments and research institutes on evaluating the population awareness about scientific innovations, such as new food-production technologies and the development of drugs. Unmet demand is to find new methods to measure the impact of scientific research and its social outreach. This work presents an Altmetrics-based framework to map the research impacts using alternative metrics based on the exchange of scientific knowledge on social media and online environments. This master thesis contributed to the ZIKAlliance consortium, enabling an online platform to monitor the scientific evolution and its social perception on the Zika epidemic.

KEYWORDS

Altmetrics, Bibliometrics, Social Network Analysis, Zika, Chikungunya, Dengue, COVID-19

1 INTRODUCTION

This article presents a summary of results obtained in the thesis defended in the Graduate Program in Informatics on 05/31/2019. The work was developed by Luís Fernando Monsores Passos Maia, in a period of 18 months, under the supervision of Jonice Oliveira.

The measurement of technological advances and scientific impact is an issue that follows researchers since Science was institutionalized. Usually, the scientific impact is measured by metrics such as the number of citations and h-index. These metrics estimate researchers' reputation and productivity based on the impact of their publications [4]. However, these metrics have been criticized in several aspects. Some people claim that those ignore the more subtle and informal elements of academic influence, such as the engagement with the scientific community, leading of research groups, and result dissemination beyond the academic milieu [27]. Moreover, the demand for faster results dissemination and knowledge exchange has led researchers to use social media to publish their achievements [10, 23, 25].

Consequently, alternative metrics - called as Altmetrics- based on social media have been used to get a bigger influence and contextualize scientific works [28, 29]. The alternative metrics tries to map the correlation between researchers and society. This relationship has been strengthened by the exchange of experiences, opinions, ratings, and content publishing in social media and online sources (e.g., online news, blogs, discussion forums, and Online Social Networks (OSN)) [5].

Due to this scenario, we created a computational framework - called REALM (Researcher Evaluation ALternative Metrics) - to identify the scientific and social reputation of researchers and their work, using alternative metrics. The main goal of this framework is to answer: "who are the most influential scientists on the ACADEMIC perspective?" and "who are the most influential scientists on the SOCIAL perspective?" Based on the REALM, we developed a web system, which was used in ZIKAlliance (international Zika research consortium) [35] and Oswaldo Cruz Foundation (Fiocruz) to analyze the impact of research related to the Zika epidemic.

2 METHODOLOGY

This thesis followed the activities: 1. Systematic Literature Review; 2. Creation of a framework to data collection and integration; 3. Creation of a method to measure a researcher's academic reputation, based on metrics of social network analysis; 4. Creation of a mechanism to measure a researcher's social reputation, based on alternative metrics; 5. Creation of a ranking mechanism for the social relevance; 6. Development of the system; 7. Experiments in the Zika scenario; and 8. Publications.

3 FRAMEWORK DESCRIPTION

The framework is divided in four modules, as shown in Figure 1. It was implemented in PHP (native) with the library EasyRDF[1] version 0.9.0; Javascript (native) with the libraries Cytoscape.js[2] version 3.2.9, jQuery version 2.1.4, and Google Charts[3], in the GeoChart and BubbleChart types; HTML and framework CSS Bootstrap.

3.1 Academic data extraction and processing

This module is responsible for retrieving data from publications in indexing databases (e.g., PubMed, Web of Science, Scopus, etc.) to create Scientific Co-authorship Networks (SCN) [12, 13, 24, 26] on a domain (e.g., Zika or Chikungunya). The module operates extracting pieces of information from the publications such as title, authors' name, affiliations, date of publication, article id, and others. Using this information, we identify the co-authorship networking. This module executes the: (i) Association of two nodes (authors), based on the title of a publication, characterizing an edge; (ii) Removal of edges without associated nodes; (iii) Representation of the social network described in item (i) using a matrix; (iv) Removal of duplicate items; (v) Identification of edges weight, based on the co-authorship frequency; (vi) Assignment of identifiers at each node

In: II Concurso de Teses e Dissertações (CTD 2020), São Luís, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2020.

ISSN 2596-1683

[1] http://www.easyrdf.org/
[2] http://js.cytoscape.org/
[3] https://developers.google.com/chart

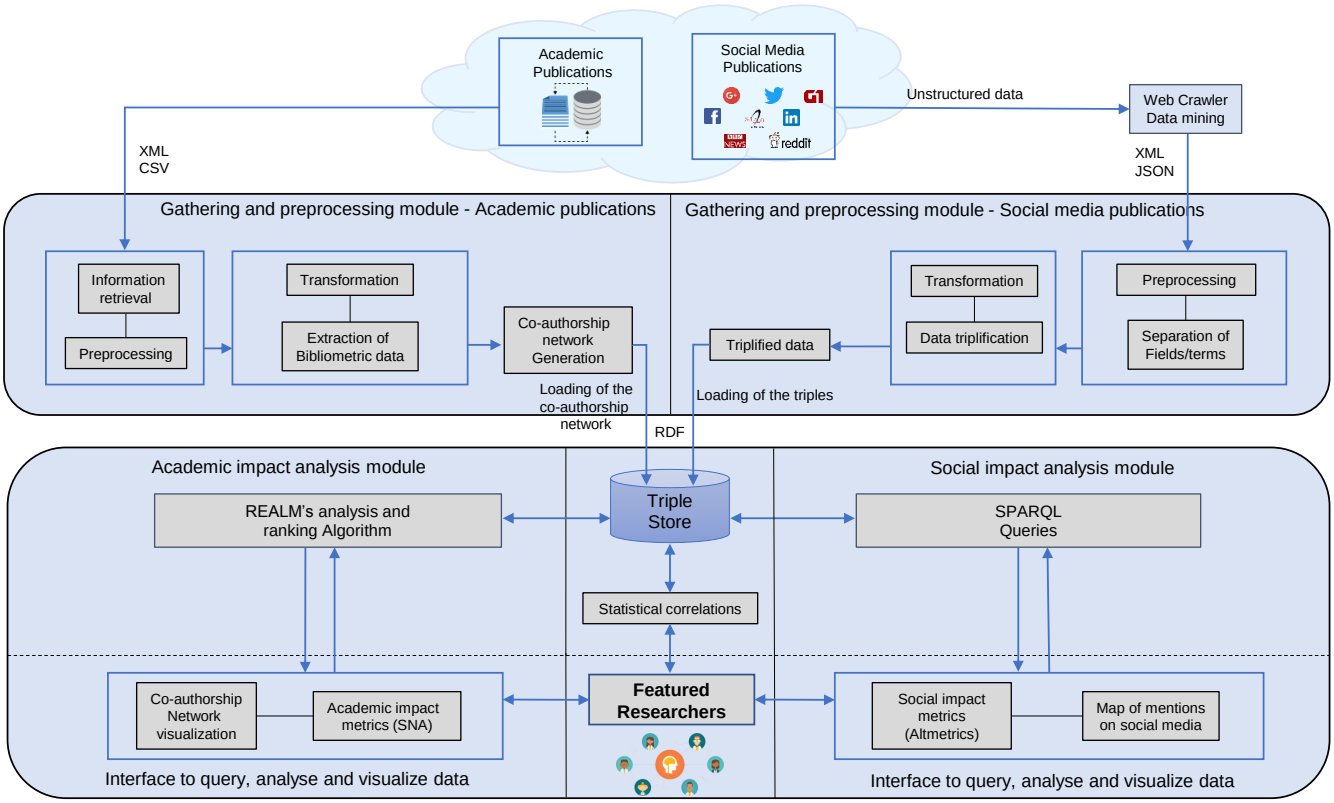

Figure 1: The framework REALM

and edge, enabling visualization of the co-authorship graph and extraction of academic impact metrics.

3.2 Social data extraction and treatment

This module is responsible for the collection, preprocessing and triplification of data from social media, such as online news, blogs, discussion forums and OSN (e.g. Facebook, LinkedIn, Google+, etc.). The data extraction is based on Webhose API[4], which allows the monitoring of social media in real-time. Data is unstructured. Then, it is converted to a semi-structured format (JSON/XML), and the metadata (such as URI, title, text, author, country, domain, date, language, and shares on OSN) are extracted. The next step is the data description using RDF triples [11, 22, 32], based on the RDF data model available at <https://realm0.github.io/1>. Finally, the triples are stored in the Apache Jena Fuseki triplestore[5]. The procedures described in 2.1 and 2.2 were executed using Knime tool[6], which optimized the preprocessing of the large volume of the text contained in the XML, JSON, and CSV files [3].

3.3 Academic impact analysis

This module detects relevant clusters and individuals, using productivity and academic impact metrics of the SCN (section 3.1) in three levels: (i) Global - using the global graph, which allows comparing publications and collaboration among researchers from different areas; (ii) Local - analyzing the subnets, which helps to identify clusters and the relevance of researchers in a group; (iii) Individual - maps the most influential researchers using the number of publications of a researcher and his network centrality. For the academic ranking, we use the centrality metrics Degree, Closeness, and Betweenness [1, 9, 12, 33] for each researcher. Additionally, the PageRank values of each researcher are calculated as indicative of other important co-authorships [6, 7, 12].

3.4 Social impact analysis

This module is responsible for extracting social impact metrics, based on SPARQL queries performed on the triples database (section 3.2), which allows the detection of important researchers/research according to its visibility/reach. This module measures: (Query 1) the reach of the researches in primary (online news) and secondary (e.g., scientific blogs and forums) communication vehicles; (Query 2) its acceptance by the population, by its dissemination on OSN such as Facebook and Google+; and (Query 3) its visibility at the global level, identifying the country of origin of the publication. The queries use as parameters the mentions to a researcher in publications, the mentions/shares on OSN, and mentions by country.

Afterward, an altmetric ranking is created, based on the mentions and shares returned by the first two queries. Also, this module generates a map that shows the geographical distribution of the mentions based on the results of the third query.

[4]https://webhose.io/
[5]https://jena.apache.org/documentation/fuseki2/index.html
[6]https://www.knime.com/

The results are saved in the triplestore, enabling the researchers' categorization. Four impact categories are possible: (i) high academic impact and high social impact - outstanding researchers in the scenario. They have a significant influence in their domain, belonging to networks of scientific collaboration with strong geopolitical/institutional references and strong online presence. (ii) High academic impact – researchers that connect and participate in well-defined research nuclei, but with little online presence. (iii) High social impact - researchers that do not have a well-defined collaboration network, but often prefer other ways to share results, such as fast-tracks, OSN (e.g. Facebook) and scientific blogs, making their dissemination more practical and faster. Also, their works are very mentioned by the public. (iv) Low academic impact and low social impact - researchers of minor importance in the scenario and irrelevant online presence. The researchers and their categories are plotted in a graph, where the dimensions are "academic index" (X-axis, normalization of the academic score) and "social index" (Y-axis, normalization of the social score).

3.5 Data storage

The academic and social impact information are triplified, based on the RDF data model available at <https://realm0.github.io/2>, and stored in the Apache Jena Fuseki triplestore. All information is registered by sessions, enabling the analyses of temporal progress in a domain or the comparison among different areas (e.g., how scientists collaborated to drive the significant advances in Zika research and how the population reacted to the findings).

4 EXPERIMENTS: NEGLECTED DISEASES

In the Zika scenario, the evaluation was three-fold. First, we conducted a proof of concept involving researchers from Fiocruz and the international consortium ZIKAlliance, as shown in [14, 19]. The researchers wholeheartedly agreed with the presented rankings (academic and social impact). The second evaluation consisted of analyzing the usefulness and correctness of a system created from REALM, using Technology Acceptance Model (TAM) [8]. The participants indicated the success of the approach. In other words, they found the system useful, easy to use and would rather prefer to use it in future than other methods to find specialists [16]. The results of the TAM evaluation can be found in <https://realm0.github.io/3>. The third evaluation was a quasi-experiment, which examined the quality and precision of the results (academic and social rankings) generated by REALM using a Goal, Question, Metric (GQM) approach [2]. The results demonstrated a good performance of the framework, which had values above 70% in the metrics of Similarity, Accuracy, and Utility. According to the specialists, the results are consistent and reflect very evident realities, among them: (i) The mapping of the outbreak evolution in its most critical period. (ii) The mapping of the interactions between researchers, population and media. (iii) The impacts of the scientific output dissemination on social media, allowing us to better understand how the population sees and interprets the findings made by the scientists.

Nowadays, this approach has been used to analyze the triple arbovirosis outbreak caused by the Aedes aegypti mosquito (Zika-Dengue-Chikungunya) [14, 15, 17] and COVID-19, in Brazil and worldwide. The web system is available at <http://www.realm.net.br>. We provided a brief description on how the system interface works. The goal is to guarantee our studies replicability to other researchers. For this purpose, we made the Zika and Chikungunya datasets available at <https://goo.gl/vsBeK3>. We also made a vídeo tutorial explaining how the system interface works, using a smaller dataset as example (Zika - crawled period: Oct - Dec 2016), which is available at <https://goo.gl/pVFJmP>. The video tutorial is available at <https://youtu.be/NfcdG8o0PyE>.

5 CONCLUSION: RESEARCH APPLICABILITY AND CONTRIBUTIONS

In this work, we presented the framework REALM, which provides a set of metrics to assess the reputation of researchers as well as mechanisms for measuring and visualizing the impacts of science on specific research scenarios, in this case, on Zika domain. We can highlight:

5.1 Scientific Contribution

The use of 3 perspectives (productivity, academic influence, and social impact) in a single approach represents a step forward in measuring the impacts of science. This master thesis brings contributions to the areas: Altmetric, Scientometric, and Social Network Analysis. This thesis is related to the challenge "Computational modeling of complex systems: artificial, natural, socio-cultural, and human-nature interactions", defined by Brazilian Computer Society [31]. This thesis also supports the priority areas defined by the Ministry of Science, Technology, Innovations and Communications (MCTIC): "Production Technologies" and "Technologies for Sustainable Development". This thesis directly contributed to the ZIKAlliance consortium and Fiocruz, enabling the wide analysis of Zika epidemic, helping in the definition of communication strategies to society and supporting the national recognition in the international scenario. We have important publications (see section 5.3), a travel award in ACM International Conference on Web Intelligence - WI '19 (because the student was the only Latin author of full paper), and this thesis was financed by an international consortium (ZIKAlliance), consolidating the Brazilian participation on it.

5.2 Multidisciplinarity and social contribution

During the thesis, the author was involved in different areas from Computer Science (computer-human interface, information visualization, database systems, software experimentation, information systems, and graphs), Public Health and Neglected Diseases. Through this thesis, Maia et al. [17] identified how scientific interactions on the Zika epidemic occurred (studying in-depth the Zika SCN). This study was pioneer in this domain and an important resource for understanding the evolution of Zika research. The approach has been used to analyze other scenarios (Dengue, Chikungunya, and COVID-19). It is important to emphasize that this solution can be applied in any area or domain. REALM can aid researchers, universities, and funding agencies to better visualize and study different domains and compare the national reality with abroad. All the data is open, using LOD principles, easily reused and can support other researchers.

5.3 Knowledge dissemination

The knowledge dissemination was made through publication in journal [17]; important international venues such as the Web Conference (ex-WWW) [16], the IEEE/WIC/ACM International Conference on Web Intelligence - WI [15], and the International Symposium on Zika Virus Research [20, 30]; and national events such as WebMedia [14, 18], SBBD [19], and SBSI [21].

5.4 Future works

Future studies should include the implementation of new functionalities on REALM and its application on new research domains/scenarios such as Computer Science topics (e.g. Crowdsourcing, Altmetrics and Social Network Analysis) other neglected diseases (e.g., Dengue and Mayaro) and COVID-19. This research is currently being applied in the study of the COVID-19 pandemic, in partnership with Fiocruz and the Wellcome Trust Foundation [34].

ACKNOWLEDGMENTS

We wish to thank the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES), Brazilian National Council of Scientific and Technological Development (CNPq), Rio de Janeiro State Research Foundation (FAPERJ), the Zika Social Sciences Network (Fiocruz), and ZIKAlliance. The study was partially financed by the European Union's Horizon 2020 Research and Innovation Programme, ZIKAlliance Grant Agreement no. 734548, and CNPq.

REFERENCES

[1] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. 2008. *Dynamical Processes on Complex Networks.* Cambridge University Press.

[2] Victor R Basili, Gianluigi Caldiera, and H Dieter Rombach. 1994. The goal question metric approach. *Encyclopedia of software engineering* (1994), 528–532.

[3] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Kilian Thiel, and Bernd Wiswedel. 2009. KNIME-the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD explorations Newsletter* 11, 1 (2009), 26–31. http://dl.acm.org/citation.cfm?id=1656280

[4] Johan Bollen, Herbert V Sompel, A Hagberg, and R Chute. 2009. A principal component analysis of 39 scientific impact measures. *PloS one* 4, 6 (2009).

[5] Lutz Bornmann. 2014. Validity of altmetrics data for measuring societal impact: A study using data from Altmetric and F1000Prime. *Journal of Informetrics* 8, 4 (2014), 935–950.

[6] Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems* 30, 1 (1998), 107–117. http://www.sciencedirect.com/science/article/pii/S016975529800110X

[7] Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the Seventh International Conference on World Wide Web 7* (Brisbane, Australia) *(WWW7)*. Elsevier Science Publishers B. V., NLD, 107–117. https://dl.acm.org/doi/10.5555/297805.297827

[8] Fred D Davis. 1985. *A technology acceptance model for empirically testing new end-user information systems: Theory and results.* Ph.D. Dissertation. Massachusetts Institute of Technology.

[9] Linton C. Freeman. 1978. Centrality in social networks conceptual clarification. *Social networks* 1, 3 (1978), 215–239.

[10] Amy Harmon. 2016. Handful of Biologists Went Rogue and Published Directly to Internet. *The New York Times* (March 2016). https://www.nytimes.com/2016/03/16/science/asap-bio-biologists-published-to-the-internet.html

[11] Olaf Hartig, Christian Bizer, and Johann-Christoph Freytag. 2009. Executing SPARQL Queries over the Web of Linked Data. In *The Semantic Web - ISWC 2009*. Vol. 5823. Springer Berlin Heidelberg, Berlin, Heidelberg, 293–309.

[12] Xiangjie Kong, Yajie Shi, Shuo Yu, Jiaying Liu, and Feng Xia. 2019. Academic social networks: Modeling, analysis, mining and applications. *Journal of Network and Computer Applications* 132 (April 2019), 86–103.

[13] Barabási László. 2016. *Network science.* Cambridge university press.

[14] Luís Fernando Monsores Passos Maia, Marcia Lenzi, and Jonice Oliveira. 2019. REALM: An altmetrics-based web system to map science impacts on society: a case study on chikungunya research. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web - WebMedia '19*. ACM Press, Rio de Janeiro, Brazil, 481–488. https://doi.org/10.1145/3323503.3361697

[15] Luis Fernando Monsores Passos Maia, Marcia Lenzi, Elaine Rabello, and Jonice Oliveira. 2019. REALM: An Altmetrics-based Framework to Map Science Impacts on Society. A Case Study on Zika Research. In *IEEE/WIC/ACM International Conference on Web Intelligence on - WI '19*. ACM Press, Thessaloniki, Greece, 233–241. https://doi.org/10.1145/3350546.3352523

[16] Luís Fernando Monsores Passos Maia, Marcia Lenzi, Elaine Rabello, and Jonice Oliveira. 2020. A Web Tool to Map Research Impacts Via Altmetrics. In *Companion Proceedings of the Web Conference 2020 (WWW '20)*. Association for Computing Machinery, Taipei, Taiwan, 235–239. https://doi.org/10.1145/3366424.3383549

[17] Luis Fernando Monsores Passos Maia, Marcia Lenzi, Elaine Teixeira Rabello, and Jonice Oliveira. 2019. Scientific collaboration in Zika: identification of the leading research groups and researchers via social network analysis. *Cadernos de Saúde Pública* 35, 3 (2019), 1–21. https://doi.org/10.1590/0102-311x00220217

[18] Luís Fernando Monsores Passos Maia and Jonice Oliveira. 2017. Investigation of Research impacts on the Zika Virus: An Approach Focusing on Social network Analysis and Altmetrics. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web - WebMedia '17*. ACM Press, Gramado, RS, Brazil, 413–416. https://doi.org/10.1145/3126858.3131593

[19] Luis Fernando Monsores Passos Maia and Jonice Oliveira. 2018. REALM: Um Framework Computacional para Investigar os Impactos de Pesquisas Através de Métricas Alternativas. In *Proceedings of 33rd Brazilian Symposium on Databases (SBBD 2018)*, Vol. 01. Sociedade Brasileira de Computação (SBC), Rio de Janeiro, RJ, Brazil, p. 37. http://sbbd.org.br/2018/wp-content/uploads/sites/5/2018/08/037-sbbd_2018-fp.pdf

[20] Luis Fernando Monsores Passos Maia, Jonice Oliveira, Elaine Teixeira Rabello, Marcia Lenzi, and Kenneth Rochel de Camargo Jr. 2018. Scientific collaborations in Zika: identifying the main research groups through Social Scientific Network analysis. In *Book of Abstracts - International Symposium on Zika Virus Research*. Paris: European Union's Horizon 2020 Research & Innovation Programme, Marseille, France, 101. https://zikalliance.tghn.org/site_media/media/medialibrary/2018/05/Book_of_Abstracts_Zika_Symposium_-_Marseille_2018.pdf

[21] Luís Fernando Monsores Passos Maia and Marcela Mayumi Mauricio Yagui. 2017. Data Triplification of Zika News. In *Anais do Simpósio Brasileiro de Sistemas de Informação (SBSI)*. Sociedade Brasileira de Computação, Lavras, 40–47. https://doi.org/10.5753/sbsi.2017.6024

[22] Marcela Mayumi Mauricio Yagui, Luís Fernando Monsores Passos Maia, Jonice Oliveira, and Adriana Vivacqua. 2019. A Crowdsourcing Platform for Curating Cultural and Empirical Knowledge. A Study Applied to Botanical Collections. In *IEEE/WIC/ACM International Conference on Web Intelligence on - WI '19 Companion*. ACM Press, Thessaloniki, Greece, 322–326. https://doi.org/10.1145/3358695.3360920

[23] Donald G. McNeil Jr. 2016. Zika Data From the Lab, and Right to the Web. *The New York Times* (July 2016). https://www.nytimes.com/2016/07/19/health/zika-data-monkey-studies.html

[24] Guilherme Vale Menezes, Nivio Ziviani, Alberto H.F. Laender, and Virgílio Almeida. 2009. A geographical analysis of knowledge production in computer science. In *Proceedings of the 18th international conference on World wide web - WWW '09*. ACM Press, Madrid, Spain, 1041. https://doi.org/10.1145/1526709.1526849

[25] Lilian Nassi-Calò. 2016. Saiu no NY Times: Biólogos se rebelam e publicam diretamente na Internet. *SciELO em Perspectiva* (April 2016). http://blog.scielo.org/blog/2016/04/07/saiu-no-ny-times-biologos-se-rebelam-e-publicam-diretamente-na-internet/

[26] Mark E.J. Newman. 2004. Who Is the Best Connected Scientist?A Study of Scientific Coauthorship Networks. In *Complex Networks*, Eli Ben-Naim, Hans Frauenfelder, and Zoltan Toroczkai (Eds.). Vol. 650. Springer Berlin Heidelberg, Berlin, Heidelberg, 337–370. https://doi.org/10.1007/978-3-540-44485-5_16

[27] PLoS Medicine Editors. 2006. The impact factor game. *PLoS Med* 3, 6 (2006).

[28] Jason Priem. 2013. Scholarship: Beyond the paper. *Nature* 495, 7442 (2013), 437–440. http://www.nature.com/nature/journal/v495/n7442/full/495437a.html

[29] Jason Priem, Paul Groth, and Dario Taraborelli. 2012. The Altmetrics Collection. *PLoS ONE* 7, 11 (2012). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3486795/

[30] Elaine Teixeira Rabello, Marcia Lenzi, Luis Fernando Monsores Passos Maia, Jonice Oliveira, Marcelo Pereira Garcia, Janine Cardoso, Gustavo Correa Matta, and Kenneth Rochel de Camargo Jr. 2018. Production and circulation of knowledge about Zika: from scientists to social media users. In *Book of Abstracts - International Symposium on Zika Virus Research*. Paris: European Union's Horizon 2020 Research & Innovation Programme, Marseille, France, 106. https://zikalliance.tghn.org/site_media/media/medialibrary/2018/05/Book_of_Abstracts_Zika_Symposium_-_Marseille_2018.pdf

[31] SBC. 2006. Grandes Desafios da Pesquisa em Computação no Brasil – 2006 – 2016, Relatório sobre o Seminário dos Grandes Desafios da Computação. https://www.sbc.org.br/documentos-da-sbc/category/141-grandes-desafios

[32] W3C. 2014. RDF 1.1 Primer. https://www.w3.org/TR/rdf11-primer

[33] Stanley Wasserman and Katherine Faust. 1994. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press.

[34] WellcomeTrust. 2020. About us | Wellcome. https://wellcome.org/about-us/

[35] ZIKAlliance. 2020. ZIKAlliance - About Us. https://zikalliance.tghn.org/about/