How to Extend your Data Lifetime: Research Data Management in Indonesia's Context

Dasapta Erwin Irawan[1]
Applied Geology Research Group,
Faculty of Earth Sciences and Technology
Institut Teknologi Bandung
Bandung, Indonesia
dasaptaerwin@gmail.com

Cahyo Darujati[2]
Faculty of Computer Science
Universitas Narotama
Surabaya, Indonesia
cahyo.darujati@narotama.ac.id

Santirianingrum Soebandhi[3]
Faculty of Economics and Business
Universitas Narotama
Surabaya, Indonesia
santirianingrum@narotama.ac.id

Fierly Hayati[4]
Dr Soetomo General Hospital
Surabaya, Indonesia
h.fierly@gmail.com

Deffy Ayu Puspito Sari[5]
Environmental Engineering Department
Faculty of Engineering and Computer Science
Universitas Bakrie
Jakarta, Indonesia
deffi.sari@bakrie.ac.id

***Abstract*— Data is the basis of research. On the other side, the world has a problem of replication. The first problem is we don't really know how to manage our own data to able to reanalyze it at some point after the research has been finished. The lifetime of data is very short, in only one or two fiscal years. In this article we will describe on how to write a research data management in order to extend the lifetime of data. There are seven basic components to remember before writing a proper research data management: (1) Data storage and software, (2) Metadata, (3) Structure, (4) Persistent link, (5) Licensing, (6) Data maintainer, (7) Indexing. In several fields, including medicine, an anomyzation strategy will be needed. We also need to put into account the Intellectual Property Rights and data ownership in to the equation, as Indonesian scientists are not properly exposed to those subjects.**

Keywords*— *Research data management, data sharing, open data

I. INTRODUCTION

More than 70% scientists have problems to repeat their own experiment [1]. This article echoed some of similar articles exposing the problem in data management, data sharing, and keeping a track of data [2]. In Indonesia's context, the number of problems increases with the low knowledge of data literacy. All instruments only measure the output of the research in form of peer-reviewed article or conference article [3]. The obvious impact had been the very short data lifetime. Data could only be found in the duration of research and at the most one year after that [4].

The objectives of this article are (1) to explore the barrier of data sharing and to add some knowledge and skill to properly manage research data using free tools and infrastructure; (2) to describe the barriers of data sharing; (3) and the benefits of data sharing.

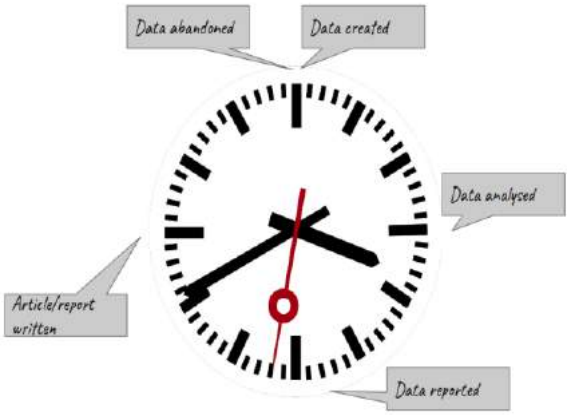

FIG. 1 THE LIFETIME OF DATA [4]

II. MATERIALS AND METHOD

We used several materials to extract the information about research data management and extending data usage

and lifetime. SHARE-research, Google Scholar, Dimensions, databases were used to keep track the development on the mentioned subjects. They are both free to use, with some notes. SHARE is free and open source, they even offer an API for interesting parties to freely access their database. Google Scholar and Dimensions are both freemium application, which means they offer a functional free account and paid service for full features. All searches were conducted in May 29th, 2018. While the size of Google Scholar database remains unclear, several papers have discussed how to estimate its size [5].

TABLE 1. DATA SOURCES AND JUSTIFICATION

	SHARE-research	Google Scholar	Dimensions
Document language	UN language	Potentially all languages	Potentially all languages
Total document volume	99,617 (articles) 2,188,221 (preprints) 169 (sources)	>160 million docs (May 2014, estimation) [5]	> 128 million docs [6]
Accounts	free and fully functional	free and functional	Free with limited function
Analytic tools	No	No	Yes

III. RESULTS AND DISCUSSIONS

A. Search Results and Size of Database

The results of our searches are listed in Table II. We find a large discrepancy between databases in terms of total documents. This is likely due to the age of the database. However, as they grow, each database is developed to harvest older data as well. Therefore, likely, we would see a different result each day. In this case, Dimensions is the youngest database, while GS is the oldest.

TABLE I. DATA SOURCES AND JUSTIFICATION.

	SHARE-research	Google Scholar	Dimensions
Keywords	Potentially all languages	Potentially all languages	Potentially all languages
Total documents (keyword: "RDM")	813	37,100	2,331
Time span	2014-2018	2014-2018	1969-2018

The search results from Dimensions shows the following top 10 fields of research containing: information systems, public health, artificial intelligence, sociology, policy and administration, education, computer software, clinical sciences, psychology, and business-management. More field of researches that likely handle large size of data, eg: earth sciences, atmospheric science, are not in the top list. Moreover, SHARE database shows that Datacite, an organization that creates DOI for data, is the first biggest data source.

B. Current Landscape

The following images show the current setting of data sharing and research data management. From the OpenKnowledgeMap, in which it relates with SHARE database, we could see the following components that are largely involved in the subject: (1) skills (related to human resources), (2) technology including software and hardware, (3) easy to follow guidelines, including IPR and legal related documents.

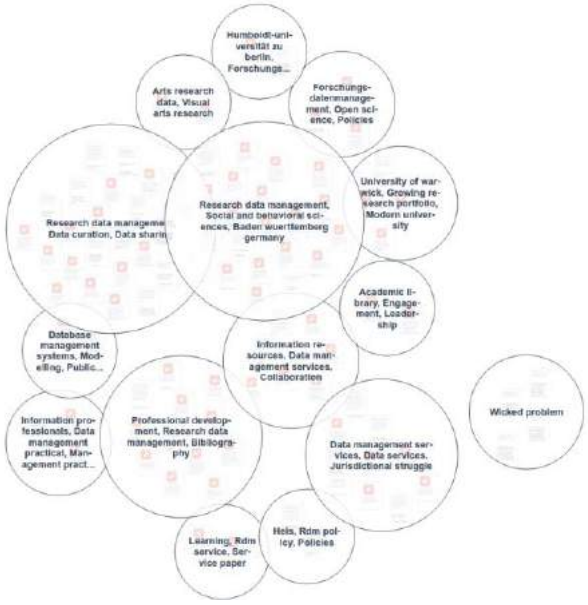

FIG. 2 CURRENT LANDSCAPE OF RESEARCH DATA MANAGEMENT [7]

Based on our search we could generate a qualitative mind map and the connections between components. This mind map should be re-tested to some extent in more quantitative manner.

The following are the barriers of data sharing [8]: (1) ethical/legal, e.g.: patient data, customer data; (2) cultural, e.g.: data abuse, cut-throat competition; (3) financial, e.g.: high cost to maintain a data server; (4) technical, e.g.: lacking of knowledge in data management. For Indonesian context, the technical bits are the major barrier. They do not have enough training to conduct proper data management strategies. Therefore a series of trainings are very much needed for Indonesian scientists.

Based on our search we could generate a qualitative infographic about the barrier of data sharing, specifically within Indonesia's context. There are three common barriers in data sharing as the basis of open science: fear, competition, and power, which all three create an inertia [9]. In Indonesia's point of view, we believe national law is the root of the barriers. However, such law has been

released as Indonesia's response to the large number of data exporting to abroad by overseas scientists [10].

The next barriers are: national and strategic discoveries, commercial discoveries, closed-research culture, peer-pressure from PI, and high cost of data infrastructure. We believe the price of infrastructure is the least problem we have since there are many open and free data hosting services. That leave the nurturing of research culture would be the next target to be taken into account, whereas a wrongful way of thinking about data sharing from senior staffs (principal investigator/PI) could lead a wrong message to be delivered to young researchers. Having said that, the focal message to be spread out is that data sharing won't erase our role as data creator. The property rights are still strongly hold by data creator, and we give the users more access to use our data for their own work. We could set a moderate license to do so. CC-BY (Creative Commons Attribution) is highly recommended, as data sharing leads to more citations [11].

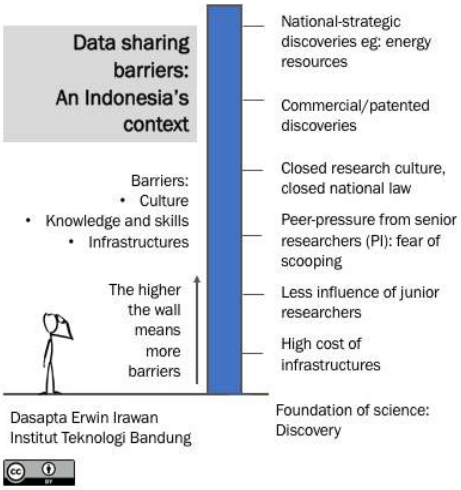

FIG. 3 THE BARRIERS OF DATA SHARING FROM INDONESIA'S POINT OF VIEW [12]

Sensitive data or sensitive information is also a big problem, in terms of lacking the knowledge. The definition of sensitive information depends on local norms and culture [13]. In social science, the following criteria can be applied to assess sensitivity: (1) something that is considered personal, full of pressure, or being valued as sacred; (2) fear attributed issues; (3) political issues that potentially generate social conflict [14]. One of the examples is conversation about transgender that is still considered taboo in some countries. In order to implement RDM, the university needs to classify the data to ensure what data can be shared widely. For instance, Duke University classifies data based on the risk level: sensitive (high), restricted (medium), and public (low) [15], while Western University uses the following group: confidential, sensitive, and public [16, 17]. These classifications are based on the existing law and regulation.

C. The Structure of an RDM

An RDM is built upon perspectives from data creator and data user. Both want a functional system that they can use conveniently. Irawan and Rachmi [4] had explained a SAFE perspectives from data creator's point of view: **S**table and searchable, **A**ccessible and interoperable, **F**lexible, **E**asy to use and re-useable.

They also pointed out the importance to fit to user's needs in setting up a data respository. Users need a CUTE repository: **C**ompact but systematic, **U**sable, **T**imely, and **E**asy to follow: (1) Data storage and software: we have options to choose between static repositories: Eprints-based platform (eg: ITB-Eprints, Undip Eprints, UGM-ETD Eprints dan UNP-Eprints), Dataverse-based platform, or DSpace-based platform. Dynamic repository (with version control), eg: OSF, Figshare, and Zenodo; (2) Metadata: this is another major issue to be raised in RDM for Indonesian scientists. They frequently speak about the importance of searchability to increase impact, but they do not have the need the knowledge and skill to determine minimum metadata that works with indexing services; (3) Structure: Folder and data structure is important, along with, writing a small Readme file to make a documentation is essential; (4) Persistent link: DOI has been famous among Indonesian academics, so this should not be major issue; (5) Licensing: Not many Indonesian academics are exposed to licensing policy. We promote the use of Creative Commons-based license; (6) Data maintainer: officials to manage data do not exist in Indonesia's default organization, where as their role is very important to set up a proper data structure, data format, and data preservation at university or project level; (7) Indexing: data indexing is important. Where did you place the data is no longer the focal point as long as it is searchable by available indexing service. Google Scholar is a good place to start as it has a mature crawling system across many platforms. However, currently there many initiatives to index data repositories, such as: Share-research, Datacite, and Base.

IV. CONCLUSIONS

Indonesia has two major problems in data sharing: (1) the lifetime of data is very short, and (2) the lack of initiative to manage data itself. Thos problems must be thoroughly solved because of the big effort to get the data itself. We identify several root problems. The first one would be national law and the last one would be the limited distributed budget to ministries and institutions to setup a data sharing system. Data sharing is essential to increase impact. This would lead to change of research culture to a more open and transparent manner. Such effort could also extend the lifetime of a dataset as more parties could reuse it for their own purposes.

We propose three solutions: (1) we need to endorse the establishment of RDM document at institutional level and/or at research project level. This document is essential as a sustainable data sharing guideline. A good RDM, at university or project level, should be made prior to the

research period, to ensure its applicability; (2) publishing research data in data journals as part of the research output; (3) publishing data as online map or online databased using open source platform such as RShiny or QGIS cloud, or in a searchable traditional blog page.

ACKNOWLEDGMENTS

Authors wish to acknowledge the INArxiv community and Prof. Roos Akbar from ITB for their feedback during the writing phase and our home institution for fund our participation in i-Click 2018.

REFERENCES

[1] M. Baker. (2016). 1,500 Scientists Lift the Lid on Reproducibility: Survey Sheds Light on the 'Crisis' Rocking Research. Available: https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970. [Accessed May 29th, 2018].

[2] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, et al., "Promoting an Open Research Culture," Science vol. 348, pp. 1422-1425, 2015.

[3] Peraturan Menteri Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia Nomor 20 Tahun 2017 tentang Pemberian Tunjangan Profesi Dosen dan Tunjangan Kehormatan Profesor, 2017.

[4] D. E. Irawan and C. N. Rachmi, "Promoting Data Sharing Among Indonesian Scientists: A Proposal of Generic University- Level RDMP," INA-Rxiv, 2018.

[5] E. Orduna-Malea, J. M. Ayllón, A. Martín-Martín, and E. D. López-Cózar, "Methods for Estimating the Size of Google Scholar," Scientometrics vol. 104, pp. 931–949, 2015.

[6] Digital-Science blog. (2018). Reimagining Discovery and Access to Research: Grants, Publications, Citations, Clinical Trials and Patents in One Place. Available: https://www.dimensions.ai/. [Accessed May 29th, 2018].

[7] Open Knowledge Maps. (2018). Overview of Research on "Research Data Management". Available: https://openknowledgemaps.org/map/9feca8d40bef7804bc1b29757ec64430. [Accessed May 29th, 2018].

[8] A. S. Figueiredo, "Data Sharing: Convert Challenges into Opportunities," Frontiers in Public Health, vol. 5, p. 327, 2017, doi: 10.3389/fpubh.2017.00327.

[9] J. Tennant. (2017). Barriers to Open Science for Junior Researchers. Available: https://figshare.com/articles/Barriers_to_Open_Science_for_junior_researchers/5383711. [Accessed June 5th, 2018].

[10] D. Rochmyaningsih. (2018). Indonesian Plan to Clamp Down on Foreign Scientists Draws Protest. Available: https://www.nature.com/articles/d41586-018-05001-7. [Accessed June 5th, 2018].

[11] H. A. Piwowar and T. J. Vision, "Data Reuse and the Open Data Citation Advantage," PeerJ, vol. 1, 2013.

[12] D. E. Irawan. (2018). Data Sharing Barriers: An Indonesia's Context. Available at: https://figshare.com/articles/Data_sharing_barriers_An_Indonesia_s_context/6470708/1. [Accessed June 10th, 2018].

[13] H. McCosker, A. Barnard, and R. Gerber, "Undertaking Sensitive Research: Issues and Strategies for Meeting the Safety Needs of All Participants," Forum: Qualitative Social Research, vol. 2, p. 22, 2001, Available at: https://eprints.qut.edu.au/23625/. [Accessed June 5th, 2018].

[14] R. M. Lee, Doing Research on Sensitive Topics. London: Sage, 1983, Available at: https://books.google.co.id/books/about/Doing_Research_on_Sensitive_Topics.html?id=AVW_MGH5ZsIC&redir_esc=y. [Accessed June 5th, 2018].

[15] University IT Security Office (ITSO). (2014). Data Classification Standard. Available: https://security.duke.edu/policies/data-classification-standard. [Accessed May 29th, 2018].

[16] Western University. (2018). What are the Data Classifications. Available: https://security.uwo.ca/information_governance/standards/data_classification/data_classifications.html. [Accessed May 29th, 2018].

[17] D. Patel, "Research Data Management: A Conceptual Framework", Library Review, 16(4/5), 226-241, doi: 10.1108/LR-01-2016-0001 [Accessed June 5th, 2018].