# HOW TO LINKING SCIENTIFIC PUBLICATION TO RESEARCH DATA

**[1]SLAMET RIYANTO, [2]HENDRO SUBAGYO, [3]EKAWATI MARLINA, [4]YANIASIH, [5]HERMIN TRIASIH**

[1,2,3,4,5]Indonesian Institute of Sciences, Jl. Jend. Gatot Subroto Kav. 10 Jakarta 12710
E-mail: [1]slamet.riyanto@lipi.go.id, [2]hendro.subagyo@lipi.go.id,
[3]ekawati.marlina@gmail.com, [4]yani.asih@gmail.com, [5]hetha23@gmail.com

**Abstract-** The policy that ensures research data availability in public archive is getting more frequently implemented among government, funding institutions and journalistic scope. Research data availability is vastly affected by the articles' age when there is no supporting policy. In Indonesia, as well as world-wide, OJS is a very popular software for managing and publishing scientific journal articles. With its Plugin, OJS can be functioned as a system to automatically deposit research data through SWORD protocol. Authors are able to determine whether the research data will be directly deposited or deposited upon publication. The scientific journal administrator also holds a very significant role in formulating policies on supporting data in each submitted manuscript. Data management implementation could begin with scientific journal administrator as the main port in collecting research data. This essay discusses about the importance of research data accessibility for others. The method used in this implementation is survey for scientific journal administrator and literature study on similar researches that have been conducted before. The analysis output can be utilized as reference for related parties that are actively involved in manifesting research data lifecycle in Indonesia.

**Index terms-** data repository, research data management, research data lifecycle, scientific data repositories, research data publishing.

## I. INTRODUCTION

Serving data "Big sciences", is a new instrument that mobilizes science [1]. In data management, the practice and objective are different. Scientific Data Repositories holds key role in science, which in general is to implement a systematic data usage practice to sustain scientific data collection, curation, preservation, long-term accessibility, dissemination as well as adequate access. As a result, scientific data repositories is often proposed as an instrument to support data publishing since it provides facilities to different actors involved in the whole process. Research data publishing, in the meantime, is an approach for research data sharing, where the objective is to release research data to be re-used by others[2].

Despite the nearly-universal agreement on the "share and reuse" merit as a tool to accelerate science performance, the systematic and effective manifestation of the objective is hindered by several obstacles. The obstacles are methodological, legal, and technical, that they are often associated with the lack of incentive for researchers who share their data. Research data availability is vastly affected by articles' age when there is no supporting policies, therefore, data sharing practice and policy needs to be initiated in order to sustain data availability. Apart from that, a specific mechanism needs to be identified to help scientific communities recognize available data collection, to facilitate their understanding as well as to encourage effective data reuse[3].
Recently, the policy that ensures research data availability in public archive is frequently implemented among government, funding institutions and journalistic scope. These policies depart from the idea that authors are their own data keeper, mainly in the long run. There are indeed numbers of researches that have proven that authors are often unable or not willing to share their data [4].

One of the measures to overcome the obstacles in research data management is through the policy of attaching supporting data in journal publishing, starting from manuscript acceptance up to articles publishing phase. In journal publishing process, there are phases where authors upload appendix data as supporting argumentation delivered in the article. Research data could be integrated into a repository system as data deposit.

By integrating two applications with different objectives, we can complement in providing articles and data simultaneously on separate systems. One of the electronic journal publishing is Open Journal System (OJS). OJS has a feature that supports Simple Web-service Offering Repository Deposit (SWORD) protocol-based interoperability. Through the protocol the data are automatically stored in the repository system in accordance with the applied policy by the journal publishing.

This article will discuss about the importance of journal publishing's role in simultaneously publishing and depositing the data to the repository system. The objective is to ensure that the articles are verified by the provided data. Others cannot easily find the articles and data, no matter where they are located, regardless the data format and the software type they

are using. To strengthen the argument in a research, an author would conduct a survey for periodical scientific journal publishing.

## II. METHODOLOGY

To support this research, authors use surveys to randomly-selected journal administrators. Respondent data came from the Indonesian Scientific Journal Database (ISJD) on the website http://isjd.pdii.lipi.go.id. The survey is created using Google Form and sent online to respondents. As the next step, the survey output will be analyzed using Microsoft Excel 2010. Authors use several literature studies related to prior researches as well as surveys to randomly-selected journal administrators The literature study analysis result will then be compared with the current condition in Indonesia, and the probability to implement the output of this research.

A data sponsored by a journal or researcher has to contain a citation and enable easier data extraction for future research needs. A scholar may read an article or book and read data visualization to access the exact same data used in the article. The data has to be available, regardless how long it has been existing since the original publication, regardless the data format or the physical location where data collections are stored.

This article also explains that data can be publicly distributed through Dataverse Network, even when users require specific authorization for confidentiality or copyright reason. Journal manuscript editors and researchers can validate the existence of certain data collections that are not changeable, even when users are not authorized to access the data collections, or when they are authorized but incapable of extracting or analyzing the data. This, enables data replication to affect science where researchers seem to take cover in the privacy issue instead of adopting scientific norms on data sharing that have been widely spread and evolving

## III. BASIC STRUCTURE OF BEC LOGIC

Two open source-based applications that function as scientific data repository and scientific data publishing are Dataverse Project and Open Journal System (OJS). OJS is developed by the Public Knowledge Project (PKP) in partnership with Stanford University and Simon Fraser University. OJS is an open source software platform utilized in management and peer-review journal publishing. Dataverse Project is developed in Harvard University's IQSS, an open source web application built for sharing, citating, analyzing and preserving research data. This application works as a facilitator of the provided data for others through persistent data citation using Digital Object Identifier (DOI). This also enables researchers to easily adopt and reuse others' work [5].

An article written by Altman, at al (2015) mentioned that Open Journal System (OJS) is one of the open source web applications used as journal publishing that functions as a tool to deposit data to repository system. By utilizing SWORD API plugin, OJS can be integrated with any application. The repository system that is integrated with the OJS is Dataverse Project. Funded by the the Sloan Foundation for two years (2012-2014), the Public Knowledge Project (PKP) and the Dataverse Project worked together to link journal articles with its underlying research data, this was done by integrating two stable open source systems. This project yielded two major outputs: a plugin for the OJS that sustains data delivery, citation, review and publication work flow, as well as extention to the Dataverse that provides standard API deposit.

OJS enables attachment file publication along with the article. Previously OJS did not have certain provisions on review, publication or research data format presentation in general. This integration feature facilitates data sharing and archiving, which enables authors to store their research data at the same time with the articles. It simplifies the data collection process as well as providing permanent identification which is indexed for the published data. In line with the OJS' focus on the solid scientific work flow, the integration to the Dataverse enables a consistent editorial policy to be implemented in the attachment data submission and exposure, in a certain case, for an instance, it may be required to publish and to grant identification to the data collection, even when the related article is declined. This also enables an appropriate content embargo. Users can also configure whether they need to make the article data available immediately upon receiving articles from the journal editor, or to wait until the article itself has been published. It enables a faster scientific work flow.

For researches in the 21st century, scientific articles' extensive availability and accessibility, as well as data set and digital resources become a norm. The increase in scientific resources repository numbers enable researchers to discover, comprehend and develop the preceding work in a bigger scale than ever. Ensuring articles, data, software and other resources' availability requires a lot of effort, as well as time and money investment[6].

The essay written by June, M. and colleagues (2016) discussed about the significant disparity in scientific research infrastructure i.e. lack of structure in linking related resources in the entire organization, repository and system. There are series of reciprocal linking between research article, data, software and other

resources that are being utilized in scientific invention. However, the repository for this resources normally only supports certain types of resources, such as software. This condition has caused crossed-information in majority of repository. It will be advantage for resources producers and users if a repository, with specific specialization and varying user communities could operate on a technical level as well as process to provide a vastly-related scientific resources network. This article recommends; firstly, to sustain the ongoing progress carried by several existing groups. The coalition that links the publication and data needs to be articulated and exchanged . DataCite, meanwhile has also formed a broaden community, along with a set of tools, metadata structure and partnership. Secondly, what this article is recommending, is to promote cross-linking in the new scientific communication initiative as well as the existing initiative, if possible. Data repository, publishers, and institutional repositories have to ensure that programmed resources are accessible in order to facilitate automatic citation detection towards data and other scientific resources.

The basic element of reproductivity and usability is an open and constantly available research data presentation. However there are plenty of general approaches for primer data publication being used do not cover adequate long term sustainability, openess, accessibility or consistency. They also disallow comprehensive exploitation done by modern web technology. As a result, several authoritative researches recommend standardized direct citation from data that are archived in the persistent repositories. Data have to be considered as first-class scientific object and be treated equally in multiple manners to cite and archive scientific literatures. From this point, we could briefly review a set of the state-of-the-art principle-based recommendations that have been widely approved to be used as scientific data citation, the Joint Declaration of Data Citation Principles (JDDCP). The article then presented a work frame for JDDCP operationalization; as well as a set of initial recommendations on identifier scheme, resolution identifier behavior, required metadata element, and the best practice to manifest programmatic action ability from the cited data. The main audience target for the general implementation guidelines in this article consists of publishers, scientific organizations, persistent data storage, including technical staffs in these organizations. In the meantime, researchers could also benefit from this recommendation. The manual provided here is meant to help achieving a broaden access, equal human beings as well as access to the machine from the stored data, in order to sustain proof, validation, reproductivity and reusability of the significantly improved stored data.

The image of the future, the old ways in scientific research conduct is replaced with computers that divulge knowledge out of data simply by pressing a button [3]. Scientific research activities conducted by institutions such as LIPI, BATAN, BPPT or other institutions are deemed not optimized before managing research data. Publication of research outputs are easy to find for they are available in the form of journal articles, proceedings, monographs or books while research data are the other way round. The reason why research data are very difficult to find is because by the end of each research, it is only the budget that is audited, and not the data research itself. Thus, we can conclude that research data is not managed. This issue can't be separated from the actors involved in producing the data, namely; the researchers, journal publishers, donor institution and government. Actor who holds key role in providing data is government since they have the authority to formulate policy on data management throughout the research activity, while the other actors merely comply with the implemented policies. As a technology, data sharing, data management practice and policies from funding institutions as well as scientific organizations have evolved in the last few years[7].

Scientific publication is a high-value critical entry point in making data available, open, accessible and usable. Majority of publishers convey statements related to data inclusion or data release as part of the publication, admitting that a complete data inclusion can increase the value and is a part of research integrity[8]. Publication and data have to be easily accessed despite the difference between its location and management. One of the objectives is to verify between the argument conveyed in the publication with the data used.

In 2016, the Center for Scientific Documentation and Information – Indonesian Institute of Sciences (PDII LIPI) conducted a research to develop a big data repository system that works as a storage, preservation and analysis tool. From the research, a repository system based on Dataverse Project framework was generated. Before determining the software that suits the majority Indonesian researchers' character and behavior, PDII team had conducted a comparison analysis on the repository system.

One of the key points available in the Dataverse is the access right settings on the data and data curation mechanism to ensure the quality of published data. Data curation has become a new responsibility to researchers, librarians and information professionals within the digital library environment[9]. As producer, researchers want to believe that they generate high-quality data; while as consumer, they want to extract data with the highest quality[10].

Beside considering the quality of published research data, researchers also consider the data exchange mechanism, with the interoperability mechanism. The objective is to facilitate researchers in managing publication and data simultaneously. At times, researchers need the data to be the fundamental of a publication so that it attracts other researchers, therefore the data needs to be openly provided, to enable verification and validation process by others. If a desired data is unavailable or even lost, the credibility of a research will be questioned. Others may think the researcher has used a dummy data, or a manipulated data. To achieve progress in science, we have to be transparent and willing to share data[11]. Data sharing will only be successful if it makes benefit for a researcher in ensuring that his/her data is openly accessible[12].

Digital repository has the capability to be integrated with research life. They can collect, manage, preserve and provide the output of a research that has been conducted. Research output covers different type of object, including journal articles, books, chapters, data collections, rich media and many others. One of the success keys in digital repository management is when one is able to integrate a research institution effectively and efficiently in capturing research output as close as possible with its invention sources[13]. Meanwhile, Ghergina and Katsanidou (2013) stated that "each step towards the data depends on the journal cooperation". However, this may not be achieved by solely relying on journal. Researchers that have done excellent work in archiving their data to the public repository may generate a more valuable essay for the scientific communities or for other publications.

Journal publishers use OJS which functions as the main pillar in research data management. When an author submits his/her manuscript, there is a facility to attach supporting data so that the data can be deposited in the repository system that works on SWORD API protocol. The attached data can be procedurally adjusted with the journal management settings. If the manuscript is approved, then the data can be deposited in the repository system, or the data will be deposited once the manuscript has been published.

In the publisher to repository scenario, SWORD can be used by the publisher to push the content to the repository. Publishers also have the option to be involved in a more complete storage life-cycle. For an instance, a recently published content can immediately be archived in the repository with embargo set so that the content will not published until the copyright policy allows to do so. This scenario has been implemented in scientific journal publishing that uses OJS framework. SWORD API plugin on the OJS is highly ideal for simultaneous data collection and publication (Figure 1).

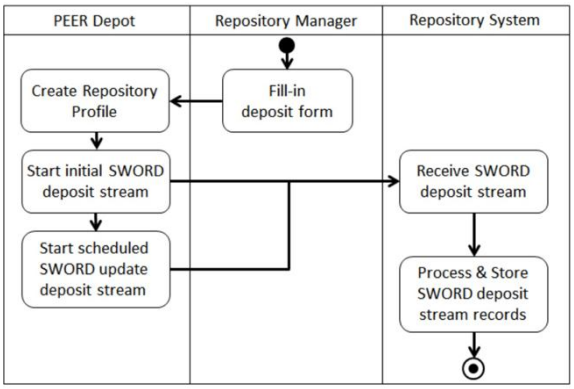



**Figure 1. SWORD Interaction between Depositor and Repository (Lewis, S., et al: 2012)**

A few of publishers and OJS journal editor are suggested to give feedback on each case of relevant data publication use, which is consistent with their journal publication work flow. Below are illustrations of two different data publication work flow. Work flow A (Figure2) illustrates an author who stores his/her research data while at the same time submitting the article to journal publisher using the same platform for data and article. If the article is rejected for publication, then author has the option to store the research data. Meanwhile, work flow B (Figure2) illustrates an author who submits his/her manuscript to the journal and deposits the research data in the repository[13].

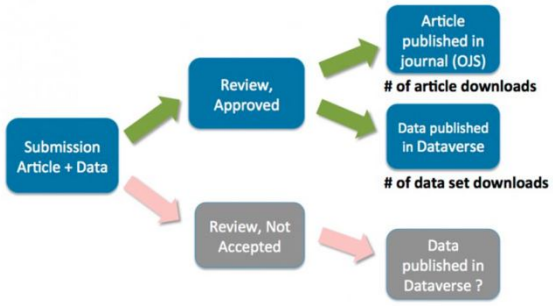



**Figure 2. Work flow A, Data publishing between OJS – Dataverse**(source: https://projects.iq.harvard.edu)

The key factor in research data management is to ensure that the system is perfectly secure; the data are encrypted yet remain indexed by the search engine. Indonesia has the policy that all periodical scientific publishing process have to be done on electronic journal-based application, mainly using the OJS framework as journal publishing. The published journals will then be accredited by the Ministry of Research, Technology and Higher Education of the Republic of Indonesia. The objective of Journal accreditation is to ensure the quality of publication in accordance with the procedure, to eventually generate a quality content. The elements that are assessed in the accreditation process is editing, management, article substances, writing style, layout, regularity as well as dissemination[14].

Integration will contribute in replication increase and research output reuse by improving infrastructure for

practices and incentives related with publication and data citation[5]. Motivation to cite dataset came from the acknowledgment that data generated during research process are equally valuable with an ongoing academic dissertation, such as essays and monographs. Scientific journals traditionally sustain researches by distributing knowledge in details; firstly, fellow scientists can measure a conclusion's intensity based on the quality of location and methods used in the research, and secondly, a further exploration can be based on the above factors. However, in many disciplines, the essay itself is no longer adequate to cover these objectives, also, the underlaying data need to be shared [15].

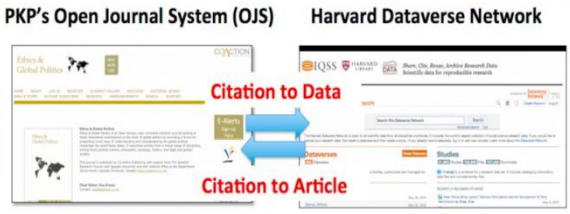



**Figure 3. Diagram on integration level between OJS journal and Dataverse** (source https://projects.iq.harvard.edu)

Journal publishers hold critical role in sustaining the research data availability for others. Therefore in this study, we conducted a survey to learn the opinion from publishers regarding research data availability. Major publishers who participated in this survey came from college background, amounting to 55 publishers (local government 1,1%; private 3,4%; ministry 8,11%; non-ministerial 10% and academic 74%). The group of knowledge from 74 journals shows varying (Figure 4). There is one journal publisher claiming that a journal has four groups of knowledge, namely linguistics, religious science and philosophy, economics and education.

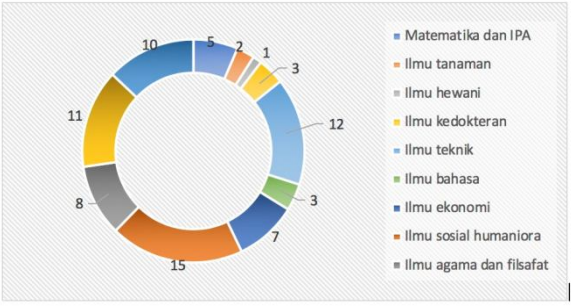



**Figure 4. Group of Knowledge**

The opinion from journal publishers on the questions asked in the survey.

a. **Your journal is integrated with repository system**. 48 publishers strongly agreed and the other 22 agreed if the journal they manage integrates with repository system. The integration with repository simplifies research data storage process in the repository. The integration feature is highly beneficial to the researchers and institutions who need to deposit their research data. According to Lewis, S., de Castro, P., & Jones, R. (2012), one of the keys to success in digital repository management is by capturing research output as close as possible to the inventor source.
b. **The managed journal becomes a vital part in the research data lifecycle**. 43 publishers strongly agreed and the other 30 publisher agreed if the journal they manage becomes the most vital part in the research data lifecycle. Research data is the main component in a scientific article, processed data that have been published in an article.
c. **Every journal publisher is required to implement a policy where every author has to attach their data**. 36 publishers strongly agreed and the other 35 publishers agreed if every journal publisher is required to implement a policy where authors attach their data. Attaching data increases citation number (Piwowar et all, 2007). Apart from that, by attaching data that are used in the article enables research data to be reused by others.
d. **If data are being cited, it will add up the point for functional credit submission**. 38 publishers strongly agreed and the other 32 publishers agreed if data is being cited, researchers/data generator will get a reward to submit functional credit point.
e. **Your journal will become publisher (data publishing).** 28 publishers strongly agreed and the other 40 publishers agreed if journal becomes data publishing.

Based on the conducted survey, we can hypotetically conclude that journal publishers support research data availability for others.

## CONCLUSION

Each research institution is demanded to manage their research data and store it to a system where others can openly access the data. The generated data can increase the value for researchers, intitutions and journal publishers. One of the most effective approaches is to manage research data by increasing scientific journal publishers' role. Journal publishers' functions which previously limited to management and peer-review journal publication could be extended to data publishing. Research data lifecycle becomes priority in a research activity in each institution, covering: creating data, processing data, analysing data, preserving data, giving access data, and re-using data. Research data availability does not solely depend on researchers as data producer. Institutions and journal publishers can also function as research data manager for the society. Another hindering factor is availability and readiness of the data repository manager to manage, archive, curate and up to data publishing has to be highly qualified.

The integration between journal administrator and data repository will not be realized if done one-sidedly. Each party bears different objective but complements each other to ensure that the society can find publication and research data without difficulty.

## RECOMENDATIONS

Based on the discussion in this article, authors recommend several points to scientific journal administratrors, research institutions, funding institutions and stakeholders.

1. Journal administrators: to formulate a policy that requires each author to include/attach research data that are used in the article.
2. Research Institutions: to provide scientific data repository system that is integrated with journal administrators.
3. Funding institutions: to formulate a policy on auditing research data before the research activity ends.
4. Stakeholders: to bestow compensation or incentive for data owner whose data have been cited by others

## REFERENCES


[1] Murray-Rust, "Big Science and Long-tail Science.' petermr's blog - A Scientist and the Web. Unilever Centre for Molecular Informatics," 2018. [Online]. Available: https://blogs.ch.cam.ac.uk/pmr/2008/01/18/932/. [Accessed: 08-Aug-2018].

[2] M. Assante, L. Candela, D. Castelli, and A. Tani, "Are Scientific Data Repositories Coping with Research Data Publishing?," Data Sci. J., vol. 15, pp. 1–24, 2016.

[3] F. Mazzocchi, "Could Big Data be the end of theory in science?: A few remarks on the epistemology of data-driven science," EMBO Rep., vol. 16, no. 10, pp. 1250–1255, 2015.

[4] T. H. Vines, A. Y. K. Albert, R. L. Andrew, F. Débarre, D. G. Bock, M. T. Franklin, K. J. Gilbert, J. S. Moore, S. Renaut, and D. J. Rennison, "The availability of research data declines rapidly with article age," Curr. Biol., vol. 24, no. 1, pp. 94–97, 2014.

[5] E. Castro and A. Garnett, "Building a Bridge Between Journal Articles and Research Data: The PKP-Dataverse Integration Project," Int. J. Digit. Curation, vol. 9, no. 1, pp. 176–184, 2014.

[6] M. June, M. S. Mayernik, J. Phillips, and E. Nienhouse, "Linking Publications and Data: Challenges, Trends, and Opportunities," D-Lib Mag., vol. 22, no. 5/6, pp. 1–12, 2016.

[7] Angela Wipperman, "Research in progress blog," 2013. [Online]. Available: http://blogs.biomedcentral.com/bmcblog/2013/06/17/g8-science-ministers-recommendations-on-access-to-research/. [Accessed: 25-Oct-2018].

[8] M. Thelwall, L. Vaughan, and L. Bjorneborn, "Webometrics," Annu. Reveiw Inf. Sci. Technol., pp. 81–135, 2005.

[9] A. M. Tammaro, K. Matusiak, F. A. Sposito, V. Casarosa, and A. Pervan, "Understanding roles and responsibilities of data curators: an international perspective," Libellarium J. Res. writing, books, Cult. Herit. institutions, vol. 9, no. 2, pp. 39–48, 2017.

[10] K. Ashley, "Data Quality and Quration," Data Sci. J., vol. 12, no. 10, pp. 65–68, 2013.

[11] N. Kroes, "Opening Science Through e□Infrastructures," 2012. [Online]. Available: http://europa.eu/rapid/press-release_SPEECH-12-258_en.htm. [Accessed: 25-Oct-2018].

[12] H. Pampel and S. Dallmeier-Tiessen, "Opening science," in Open Research Data: From Vision to Practice, 2014, pp. 213–224.

[13] S. Lewis, P. de Castro, and R. Jones, "SWORD: Facilitating Deposit Scenarios Stuart," vol. 18, no. 1, pp. 1–13, 2012.

[14] Ditlitabmas, Pedoman Akreditasi Terbitan Berkala Ilmiah. Kementerian Pendidikan dan Kebudayaan, 2014.

[15] C. Mayo, T. J. Vision, and E. A. Hull, "The location of the citation: changing practices in how publications cite original data in the Dryad Digital Repository," Int. J. Digit. Curation, vol. 11, no. 1, p. 150, 2016.


★★★