Repository of Biodiversities' Research Data

Ekawati Marlina, Slamet Riyanto and Hendro Subagyo

Center for Scientific Documentation and Information Indonesian Institute of Sciences, Indonesia, {ekawati.marlina@lipi.go.id, slamet.riyanto@lipi.go.id, hendro.subagyo@gmail.com}

ABSTRACT

Indonesian researchers from various institution have done many scientific research activities in the field of biodiversity. Many of them documented in each institution. Online availability of Indonesian biodiversity data still becomes a challenge. To support access and preservation, we develop a framework that could integrate biodiversity data and information among the institution. We develop repositories as a place to share and preserve research result. Besides that, the repository is a tool to disseminate information for the stakeholder. This paper will discuss our challenges to develop a repository and provide a brief explanation of information dissemination related to repositories.

Keywords: biological diversity repositories, research data management, knowledge management, information dissemination

I INTRODUCTION

Biodiversity is natural resources that have an economic value. It is essential for the economic development and human well-being. Biodiversity data and information become crucial for decision making and researchers. When environmental changes happen, they require data and information to protect the environment.

Beginning with the establishment of Lends Plantetuin (1817) and the establishment of reference collection in Bogor, documentation and conservation of Indonesian biodiversity has long started in Indonesia (Darajati et al., 2016). Data collection of biodiversity expedition in Indonesia from 1891 until now has documented at Indonesian Institute of sciences in the form of live collection, microbial culture, and specimen.

As one of megabiodiversities region, Indonesia has many species of flora, fauna, and microbe. Many research institute in Indonesia has conducted many scientific research activities to explore that kind of species. From the exploration, researchers gain much valuable information about Indonesian biodiversity. The availability of that information is still a problematic issue. The availability of the biodiversity data is important for the conservation and preservation of the endemic species from Indonesia.

Biodiversity data and information spread in many places, each research institute manages their research result. This cause information problematics, redundancy and ambiguous research data and information. The lack of data and information on biodiversity that can be known and accessed quickly and accurately causes the potential of natural resources, geostrategic position, and local wisdom is not maximally utilized. One repository that contains complete information about Indonesian biodiversity is needed. We would develop a framework that could integrate biodiversity data and information among the institution. The repository is making a catalogue of the Indonesian species data. It will inform a distribution of species in Indonesian province and information about endangered species. The repository should have an impact on protection, conservation, and utilization of biodiversity. This paper will discuss our challenges to develop a repository and provide a brief explanation of information dissemination related to repositories.

Biodiversity repositories collect information from much biological diversity database into centrally manage and retrieve. Digital repositories are essential to aid in coastal wetland biodiversity conservation (Krishnan et al., 2017). Repository facilitates knowledge discovery over a massive amount of biodiversity literature (Batista-Navarro, Zerva, Nguyen, & Ananiadou, 2017). Brazilian Biodiversity Information Facility Repository (SiBBr) is developed because of availability online biodiversity data are essential for conservation planning (Dias et al., 2017).

II METHODOLOGY

Requirement analysis for the development of biodiversity repositories determined through Focus Group Discussion (FGD). It is a qualitative research method and a data collection technique in which a selected group of people discusses a given topic or issue in-depth.

The participant of the FGD is three institution that has been managing biodiversity database and two institutions that has a role as an infrastructure provider. All of that institution involved in the providing data and information for biodiversity repositories. Agreement among institution to work together to actualize the biodiversity are essential agenda for the FGD.

Main topics for the discussion is the existing database in each institution, the challenges in developing the centralized repository, and the facilities that should be provided by repositories.

III DISCUSSION

The participant of the FGD is consist of 15 people. In the FGD, two institutions explain about their existing repository. In the discussion, we also discuss using the data from the repository to make an information dissemination.

All participants agreed that a centralized repository that has complete content is required. The most information needed by stakeholder from the biodiversity research is about a distribution of species in Indonesian province and information about endangered species. The content of the repository comes from researchers who directly incorporate into the system or the integration of an existing repository in each institution.

The more complete of the content, information from the repository will be more valuable. Beside provide an access, the repository will be used to disseminate information through information packaging. We make an analysis of the biodiversity research data into information packaging to fulfil the need of the community and policymakers.

A. Manage researchers knowledge through the repository

The repository is a tool to preserve and publish the research data. The repository provides a catalogue for discovery and access. Good practice of research data management is needed to support the repository. Research data management concern with the process of handling and organizing the research data throughout the entirety of the research data lifecycle (Figure 1).

Source: http://www.open.ac.uk/blogs/the_orb/?p=52

Figure 1. Research Data Lifecycle

Research data lifecycle (Figure 1) starting with creating data until re-using data. Organizational knowledge lifecycle consists of five dimensions, i.e. create, capture, secure, describe, share, and re-use (Figure 2). Comparing with research data management, the dimensions of the organizational lifecycle is almost the same.

In the context of biodiversity repositories, data created by research activity are capture to store in the repositories. All of the data managed securely. Every data has control access and sharing regulation. To easily understood and access, every data is described through metadata. The repository is a tool to share biodiversity data for the stakeholder. Re-use of data for other research activity is one of the benefits of repositories.

Source:https://fireoakstrategies.com/knowledge-management/

Figure 2. Organizational Knowledge Lifecycle

Knowledge management (KM) attempt to capture, retain, and disseminate knowledge (Hakopov, 2016). Indonesian biodiversity is the valuable asset. Biodiversity has significant role and contribution for national development. Indonesian researchers have done the expedition to explore the Indonesian biodiversity. The specific problem from research data management at research institute spread, vulnerable lost, and difficult to access (Marlina, Riyanto, & Yaniasih, 2016). Common practice, researchers store their data on the personal computer without well back up. Collected research result in a repository is one way of hindering lost of researchers knowledge.

For the institution, people or researchers knowledge is institution asset. Repository, manage by the institution, is a tool to preserve knowledge resulted from researchers activity. And this support to achieve the big goal from our repository, preserve Indonesian biodiversity data.

B. Challenges to developing Biodiversity repositories

There are some challenges to developing centralized biodiversity repository. In developing the repository, the institution should provide reliable hardware and software infrastructure. In developing repository

infrastructure, we will use an open source and Apache license.

In the FGD, two institutions reported their existing database. One institution has plant collection database. The other has a database of zoology, microbiology, and botany. The database is stand-alone, not connected with each other. Each institution has biodiversity database with different platform and format data. Access policy for each database also different. To summarize, the challenges to developing national biodiversity repositories are metadata standard, data integration, and access.

The function of metadata is to build a shame interpretation of data. The metadata is crucial for finding and sharing the data in repository system. Metadata standard implemented in biodiversity informatics is Darwin core standard (Schindel, Miller, Trizna, Graham, & Crane, 2016). The Darwin Core is based on the standards developed by the Dublin Core Metadata Initiative [DCMI] and can be viewed as an extension of the Dublin Core for biodiversity information. The Darwin Core is primarily based on taxa, their occurrence in nature as documented by observations, specimens, samples, and related information (http://rs.tdwg.org/dwc/). The existing database actually also use Darwin core as their metadata standard, but inconsistency in fulfilling the field of metadata. Missing or incorrect geographic coordinates (geo-coordinates) and inconsistencies in plant names are the barriers to data reuse and synthesis (Wiser, 2016). The metadata heterogeneity of the repositories poses a challenge for their effective integration (Goldfarb & Le Franc, 2017).

Beside easily to find, the repository should provide qualified data, regarding completeness and currency. It is not an easy job. Validate the data is labor extensive. One of the strategies is the call of community curation to validate the data that are published in repositories. For our biodiversity repositories, biodiversity scientist act as subject specialist. Community curation is also done by The Global Registry of Biodiversity Repositories (Schindel et al., 2016).

Another challenge is user access. Some of the biodiversity data is a sensitive data. Each database has different access policy. Different stakeholders have different access level. Not all information is open or available online. Data classification can be one solution. Data are classified into three groups: limited data, personal data, and general data. The classification of this data is the first step that must be done in building a national repository (Patel, 2016). Another proposal, to reduce the complexity of data sharing rules Sweeney, Crosas, & Bar-sinai (2015) introduced the data tags repository. Data tags will help scientists in sorting out which data can be accessed by the public.

The last challenge is data integration. Based on the experience from Taiwan (Shao et al., 2013), the most critical element to integrating biodiversity data is to build a species checklist as the scientific name. Via the scientific name of a species, its specimen information, DNA barcode (Barcode of Life, BOL), phylogeny (Tree of Life, TOL), ecological distributional data (in EML format), and other information (Encyclopedia of Life, EOL) can all be accessed.

Three of the challenges is the bottleneck for providing comprehensive biodiversity data. The biodiversity database in Indonesia is scattered in the various institution. It needs a synergy among institution to break down the barriers to developing biodiversity repositories.

C. **Information dissemination**

There is still a lack of data and information on biodiversity that can be known and accessed quickly and accurately. This cause contribution of scientific research results in biodiversity has not been able to contribute maximally to community development.

We choose a dataverse as a framework to provide access for biodiversity data. Dataverse is an open source web application to share, preserve, cite, explore, and analyze research data. It facilitates making data available to others and allows you to replicate others' work more efficiently. Researchers, data authors, publishers, data distributors, and affiliated institutions all receive academic credit and web visibility. (http://dataverse.org)

In biodiversity repository system, management of the data includes describing data using metadata standard and include security settings. In giving convenience and security data, repository system uses Hypertext Transfer Protocol Secure (HTTPS) protocol. Every data in the repository encrypt using Message-Digest algorithm 5 (MD5). To protect tabular data, Dataverse uses Universal Numerical Fingerprint (UNF) encryption.

Providing access to biodiversity data through the internet is one of a method to disseminate biodiversity data. Figure 3 is the information dissemination from biodiversity data that can be accessed online. Brazil is another country that provides online access to their biodiversity data, Brazilian Biodiversity Information Facility (SiBBr) (Dias et al., 2017).

Figure 3. Metadata in repository system.

Repository gives services for the stakeholder. Researchers and decision maker is two main stakeholder for the biodiversity repositories. To provide easily read information, we develop a dashboard. Biodiversity dashboard is a visualization of biodiversity indicators designed to enable tracking of biodiversity and conservation performance data in a clear, user-friendly format (Liu, Wang, Liu, & Tan, 2009). This monitoring resource services would be useful for the government as a policymaker.

We provide dashboard map and graphics. Information indicator in the dashboard is about the growth of species, distribution of species, type of species, and geographic location of species. Trend analysis of the development of biodiversity science also presented.

Many other indicators should be considered to be present in the dashboard. The visualization of the data at SiBBr gives information about Brazilian occurrence records in term of geographic coverage, taxonomic coverage, and temporal coverage. Liu et al., (2009) used four indicators on their dashboard. They are indicators to measure pressure on biodiversity (deforestation rate), state of species (Red List Index), conservation response (protection of key biodiversity areas), and benefits to human populations (freshwater provision). For Africa, the types of biodiversity data currently needed by the decision maker include species populations, distributions, offtake, trade and threat status; habitat cover or distribution; protected area coverage and management effectiveness (Liu et al., 2009).

Besides as information dissemination and knowledge management, a biodiversity repository can function as an analysis tool. Repositories complete with a tool that researchers can use to analyze their data. It can directly explore the Tabular data. Then, Researchers can choice a data visualization model to depict the result. Some tabular data format supported on biodiversity repository are SPSS, Stats, R, Excel, and CSV. For depict analysis in the image format, It can export to PNG, JPG, and PDF format. These tools help researchers to disseminate their data in the paper or presentation.

Biodiversity repository integrated with TwoRavens software. It is a system of interlocking statistical tools for data exploration, review, and meta-analysis. The geospatial data can be analyzed using WorldMap which connect to the system. WorldMap is an online, open-source mapping platform developed to lower barriers for scholars who wish to explore, visualize, edit, and publish geospatial information. The system attempts to address the gap between desktop GIS which is light on collaboration, and web-based mapping systems which often don't support the inclusion of large datasets.

IV CONCLUSION

The extensive database serves as a primary tool for biodiversity research. Collaboration among institution to involve in the developing repository is needed. With repositories, biological diversity literature more accessible and searchable. Repositories provide benefit and solutions towards the national development problems. Biological diversity data and information are an important asset. It should be preserved, accessible, and disseminate. Standard data is the big issue in integrating many databases with a different platform. The other challenges are data integration and access. Metadata of heterogeneity inhibits data integration. The main services that repositories should be provided are information accessed quickly and accurately. Presenting information in a single screen and showing a graphical presentation of the current condition and historical trend of Indonesian biodiversity is our services that could be accessed through the dashboard. The information in the dashboard is about a distribution of species in Indonesian province and information about endangered species. It would be useful for stakeholders, especially for the policymaker.

REFERENCES

Batista-Navarro, R., Zerva, C., Nguyen, N. T. H., & Ananiadou, S. (2017). A Text Mining-Based Framework for Constructing an RDF-Compliant Biodiversity Knowledge Repository. In *In: Lossio-Ventura J., Alatrista-Salas H. (eds) Information Management and Big Data. SIMBig 2015, SIMBig 2016. Communications in Computer and Information Science, vol 656*.

Darajati, W., Pratiwi, S., Herwinda, E., Radiansyah, A. D., Nalang, V. S., Nooryanto, B., … Hakim, F. (2016). *Indonesian Biodiversity "Strategy and Action Plan 2015-2020."* Kementrian Perencanaan Pembangunan Nasional/BAPPENAS.

Dias, D., Baringo Fonseca, C., Correa, L., Soto, N., Portela, A., Juarez, K., … Junior, J. (2017). Repatriation Data: More than two million species occurrence records added to the Brazilian Biodiversity Information Facility Repository (SiBBr). *Biodiversity Data Journal*, *5*, e12012. https://doi.org/10.3897/BDJ.5.e12012.

Goldfarb, D., & Le Franc, Y. (2017). Enhancing the discoverability and interoperability of multi-disciplinary semantic repositories. *CEUR Workshop Proceedings*, *1933*.

Hakopov, Z. N. (2016). Digital Repository as Instrument for Knowledge Management. Retrieved from http://eprints.rclis.org/29046/

Krishnan, P., Deepak Samuel, V., Sreeraj, C. R., Abhilash, K. R., Patro, S., Sankar, R., … Ramesh, R. (2017). Digital repositories for coastal wetland biodiversity i South Asia: A conceptual framework from India. *Wetland Science: Perspectives From South Asia*, (2017), 51–65. https://doi.org/10.1007/978-81-322-3715-0_3

Liu, Y., Wang, J., Liu, Y., & Tan, Z. (2009). Research on data integration of bioinformatics database based on web services. In *2009 First International Conference on Networked Digital Technologies* (pp. 292–296). https://doi.org/10.1109/NDT.2009.5272808

Marlina, E., Riyanto, S., & Yaniasih. (2016). Peran Pusat Dokumentasi dan Informasi dalam Pengelolaan Data Penelitian. In *International Conference on Science Mapping and the Development of Science* (pp. 281–290).

Patel, D. (2016). Research data management : a conceptual framework. *Library Review*, *65*(4/5), 226–241. https://doi.org/10.1108/LR-01-2016-0001

Schindel, D., Miller, S., Trizna, M., Graham, E., & Crane, A. (2016). The Global Registry of Biodiversity Repositories: A Call for Community Curation. *Biodiversity Data Journal*, *4*, e10293. https://doi.org/10.3897/BDJ.4.e10293

Shao, K., Lai, K., Lin, Y., Ko, C., Lee, H., Hung, L., … Chen, L. (2013). Experience and strategy of biodiversity data integration in Taiwan. *Data Science Journal*, *12*. https://doi.org/10.2481/dsj.WDS-008

Sweeney, L., Crosas, M., & Bar-sinai, M. (2015). Sharing Sensitive Data with Confidence : The Datatags System. *Technology Science*, *October 16*, 1–34. Retrieved from http://techscience.org/a/2015101601

Wiser, S. K. (2016). Achievements and challenges in the integration, reuse and synthesis of vegetation plot data. *Journal of Vegetation Science*, *27*(5), 868–879. https://doi.org/10.1111/jvs.12419