Editorial

Healthc Inform Res. 2021 October;27(4):265-266.
https://doi.org/10.4258/hir.2021.27.4.265
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Sharing Biomedical Data Obtained Through Government-Funded Research and Development Projects in Korea

Seungwoo Hwang[1], Hyoun-Joong Kong[2,3,4,5]

[1]Korea Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon, Korea
[2]Medical Big Data Research Center, Seoul National University College of Medicine, Seoul, Korea
[3]Transdisciplinary Department of Medicine and Advanced Technology, Seoul National University Hospital, Seoul, Korea
[4]Department of Biomedical Engineering, Seoul National University College of Medicine, Seoul, Korea
[5]AI Institute, Seoul National University, Seoul, Korea

Scientific research produces ever increasing variety and amount of data. Making the data available in public repositories and sharing with other researchers facilitate scientific advances. Data availability enables researchers to falsify or confirm the reported findings, which is useful not only for checking the scientific integrity but also for devising future research directions. Availability of real-world data is also useful for developing analytical methods and computer software, and for training new researchers. In addition, data availability can maximize economic utility of funding resources by avoiding duplicated reproduction of already produced data. Most importantly, existing data can be reused to conduct new studies that ask new questions beyond the scope of the original study, examples being meta-analysis that combines multiple individual studies addressing the same question, and integrative analysis that combines multiple heterogeneous data.

In recognition of the aforementioned benefits of data sharing, journal publishers and funding agencies, which are the two major driving parties towards data sharing, have been requiring researchers to deposit their data in public repositories as a necessary condition of publication acceptance and grant approval. Parallel efforts of the two parties are required because a single approach alone is not enough. In regard to the publisher-driven data sharing, a report in 2017 reviewed data sharing requirements of biomedical journals and showed that only a minority of journals strictly required data sharing [1]. For example, a total of 11.9% of journals explicitly required data sharing as a requirement of publications, whereas a total of 31.8% of journals did not mention data sharing at all. In regard to the funding agency-driven data sharing, National Institute of Health (NIH), as an example, currently requires sharing of data that were produced by NIH-funded projects, only if the grant amount is US$500,000 or more per year [2]. Therefore, neither journal publishers nor funding agencies alone are enough to ensure data sharing, and parallel efforts of the two parties are needed to maximize data sharing rate. In addition, data sharing requirements made by journal publishers and funding agencies have different audiences because the former usually reach all researchers around the world whereas the latter are generally limited to local researchers of the corresponding nation.

The aforementioned NIH data sharing policy, which was issued in 2003 and is still in effect, will however be replaced in 2023 by an augmented NIH Policy on data management and sharing [3], which requires all NIH-funded projects to be subject to data sharing, regardless of the grant amount. Thus, biomedical data sharing in the United States is ex-

pected to be strengthened in the near future. Similarly, Korea initiated a strengthened nationwide drive towards biomedical data sharing through an announcement of the "Strategy on Biological Research Resource Big Data" in 2021, which is expected in the long run to be applied to all government-funded research projects in any area of bio-health and related fields, including genomics, medicines, biomedical sciences, biomedical engineering, agriculture, fishery, environmental biology, and many others. As a public repository that accepts data submission from research in all these areas, Korea BioData Station (KBDS) was opened with joint efforts between Korea Bioinformation Center at Korea Institute of Bioscience and Biotechnology and Korea Institute of Science and Technology Information (https://kbds.re.kr).

A key to the data management system in the KBDS are the web-based submission forms that were developed for each type of biological data. Depending on the data type, different strategies were adopted to develop the data submission form that can make the inputted information sufficient and useful. For some data types, there is either an official standard file format (for example, Digital Imaging and Communications in Medicine [DICOM] for medical imaging data) or a de facto standard file format (for example, FASTQ for high-throughput genomic sequencing data). Such data types also usually have minimum information guidelines on how to prepare their metadata as well as data archives specific to them, an example for functional genomics data being the Minimum Information About a Microarray Experiment (MIAME) guideline [4] and the Gene Expression Omnibus (GEO) archive [5]. In such cases, data submission forms in the KBDS were kept largely similar to those in the existing standards and archives for international interoperability. Some other data types do not have a clear standard format. In such cases, experts from the corresponding research field reviewed and assembled existing international and domestic data specifications into a submission form. Still other data types are quite unstructured and ad hoc, whose format depends on how individual researchers organize various information. Such data are usually organized into a tabular spreadsheet form, similar to supplementary information data that accompany journal publications. For such unstructured data, submission forms of existing generalist repositories such as Dryad [6], BioStudies [7], FigShare [8], and Zenodo [9] were reviewed and assembled into a generalist data submission form. In developing the submission forms for all the three aforementioned cases of data types, experts from diverse biomedical areas participated to maximize information content of the submission forms as well as to scrutinize them.

In summary, research data sharing policies are being promoted in many countries, since they have a great potential for scientific advances and related business. Korea government recently initiated such a strengthened drive towards sharing of data from all government-funded research and development projects in any area of biomedical fields. A data repository called the KBDS was opened in October 2021 and starts its pilot operation. Subsequently, it is planned to continuously expand its features such as data analysis functionality.

ORCID

Seungwoo Hwang (https://orcid.org/0000-0002-6415-9913)
Hyoun-Joong Kong (https://orcid.org/0000-0001-5456-4862)

References

1. Vasilevsky NA, Minnier J, Haendel MA, Champieux RE. Reproducible and reusable research: are journal data sharing policies meeting the mark? PeerJ 2017;5:e3208.
2. National Institutes of Health. Final NIH statement on sharing research data [Internet]. Bethesda (MD): National Institutes of Health; 2003 [cited at 2021 Nov 4]. Available from: https://grants.nih.gov/grants/guide/notice-files/not-od-03-032.html.
3. National Institutes of Health. Final NIH policy for data management and sharing [Internet]. Bethesda (MD): National Institutes of Health; 2020 [cited at 2021 Nov 4]. Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html.
4. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29(4):365-71.
5. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30(1):207-10.
6. Dryad [Internet]. London, UK: Dryad; c2021 [cited at 2021 Nov 4]. Available from: https://datadryad.org.
7. McEntyre J, Sarkans U, Brazma A. The BioStudies database. Mol Syst Biol 2015;11(12):847.
8. Singh J. FigShare. J Pharmacol Pharmacother 2011;2(2):138-9.
9. Zenodo [Internet]. Geneva, Switzerland: Zenodo; c2021 [cited at 2021 Nov 4]. Available from: https://zenodo.org.

 https://doi.org/10.4258/hir.2021.27.4.265