Guidelines for Management, Disclosure, Sharing, and Use of Research Data at the Center for Advanced Intelligence Project (AIP), RIKEN (AIP Research Data Guidelines)

2021-03-05

The Center for Advanced Intelligence Project, RIKEN

1. Purpose

The Center for Advanced Intelligence Project (hereinafter "AIP") of RIKEN, a National Research and Development Institute, (hereinafter "the Institute"), aims to contribute to the advancement of scientific research and the solution of real-world problems through the development and application of innovative artificial-intelligence technologies, and also conducts researches on ethical, legal, and social issues that arise with the spread of artificial-intelligence technology. These researches and developments utilize not only data generated internally but also data obtained from collaborators including various companies, universities, research institutes, and projects.

The AIP shall contribute to maximizing the Institute's research activities through information technology and maximizing the results of such research, in compliance with the Basic Policy on Management, Disclosure, Sharing, and Use of Research Data (hereinafter "the Basic Policy") set forth by the Institute. The purpose of AIP Research Data Guidelines is to define matters necessary for handling research data at the AIP so that the researchers and other members of the AIP can manage research data in an appropriate and reliable manner.

2. Definitions of Research Data and Useful Data

[Research data]

In the Basic Policy, "research data" means all kinds of data acquired or generated in the course of research undertaken at the Institute. All research data at the AIP shall be managed by electromagnetic means. Data in print and other media that cannot be managed electromagnetically shall be converted appropriately so that they can be managed by electromagnetic means. Research data in principle does not include research notes, research papers, or programs executed on a computer, such as database systems or software components. Research data shall be classified as below, depending on how it is used, how it is collected, or its source.

[Useful data]

According to the Basic Policy, "Useful data" means the research data, including metadata, which shall be disclosed along with publication of research papers (hereinafter "Data of Research Papers") and other academic documents or research data that the AIP has determined on behalf of the Institute will contribute to the development of science and technology.

[Local research data]

“Local research data” means research data that is used exclusively by researchers or research projects at the AIP and is placed in a data management area that cannot be accessed by any third party. Data management media include fixed storage devices connected to computers exclusively used by researchers at the AIP, storage media provided with various discs or non-volatile memory, and storage areas that can be accessed through a network and accessed exclusively by users.

[Local useful data]

“Local useful data” means local research data that is categorized as useful data.

[Disclosure flag]

“Disclosure flag” means the information that indicates whom to disclose useful data to. In accordance with the Basic Policy, the types of disclosure flags are “undisclosed,” “restricted internal disclosure,” “restricted internal/external disclosure,” and “public.”

[Data generator]

“Data generator” means a person who has generated research data.

[Data manager]

“Data manager” means a person who manages research data, using information systems infrastructure or information media. The data manager may be either the same as or different from the data generator.

[Data user]

“Data user” means a person who receives and uses useful data.

3. Selection and Use of Research Data Repository

The research data repository developed by the Institute (hereinafter “the Institute Repository”) shall, in principle, be used for the management of research data. If the main body of research data is stored in another storage, it is desirable to register a link to that data in the Institute Repository.

Requests to the Institute for the enhancement of the Institute Repository functions must accompany objective and quantitative evidence. The AIP shall endeavor to precisely identify the quality and quantity of research data generated at the AIP in order to ensure the effective operation of the Institute Repository.

4. Handling of Personal Information

The research data containing personal information shall be handled in compliance with the relevant laws, regulations, and guidelines. In particular, research data containing special care-required personal information shall be handled strictly when conducting medical research involving people, and so on.

Furthermore, data received from outside the Institute will be treated as research data, after being processed in accordance with the guidelines for information management including the regulations described above.

5. Registration of Research Data in the Institute Repository

The research data acquired or generated in the course of research and managed by the AIP typically has a life cycle from the acquisition or generation of data, to the registration in the Institute Repository for sharing and storage, the release, and the deletion. Specifically, the life cycle of research data can be described as follows. In this case, the data generator and the data manager may be either the same or different.

a. Acquisition or generation of data

The acquired or generated data is managed as local research data by a data manager.

b. Selection of useful data

The data manager selects data to store, share, and disclose from the local research data to define local useful data.

c. Registration in the Institute Repository

The data manager uploads local useful data to the Institute Repository for registration by using the upload functionality provided by the Institute Repository. At the time of registration, it is desirable to set the disclosure flag as “undisclosed” to prevent accidental disclosure.

d. Setting of disclosure flag

The disclosure flag for data is set according to the procedure described below.

e. Deletion of data

Data is deleted according to the procedure described below.

When selecting useful data from research data, the data generator or data manager comprehensively evaluates and determines the need for data storage or data sharing with others, from various perspectives—including the status of research, the management of intellectual property in research, and a basis for the research paper to be written—and according to his/her conscience as a researcher.

In concrete terms, useful data includes:

・Data recorded in a laboratory notebook and necessary to store;

・Data to share in collaborative research;

・Data needed in the next phase of research; and

・Data needed for publication of and as a basis for a future research paper.

Consideration shall also be given to the fact that the operation and maintenance of the Institute Repository necessary for the storage of research data involves a significant cost. Efforts shall be

made to reduce costs and improve the effectiveness, such as omitting some data to an extent that does not hinder the circulation or use of data, compressing data, and avoiding storage of data that can be generated by a program.

The data manager shall be responsible for managing local research data and local useful data. In a case where such data is managed by using information systems infrastructure or media other than the Institute Repository, necessary measures shall be taken to prevent the theft or loss of, or unauthorized access to the information media, and the leakage of data at the time of disposal.

6. Changes of Disclosure Flag Status

For useful data registered in the Institute Repository with disclosure flag "undisclosed data," the data generator or data manager shall change the disclosure flag by analyzing research trends and data users and by obtaining approval for change of disclosure flag from the head of the division. The following is the disclosure flags listed from smallest to largest in terms of the scope of eligible users.

[Undisclosed] This flag is given to data accessible only by the data generator or data manager.

[Restricted internal access] This flag is given to data which data users in the Institute can read. It is possible to target all those in the Institute or only some of them by restricting access, for example by password or IP address.

[Restricted internal/external access] This flag is given to data targeted at a single data user or multiple data users, regardless of whether inside or outside the Institute.

[Public] This flag is given to data that can be read by anyone, without limiting the scope of eligible data users.

Data provision method should be individualized according to data characteristics among others. There is no need to constantly keep data readable through networks. Instead data may be provided in an information medium upon a written request for access, for instance.

The data manager shall endeavor to ensure the setting of appropriate disclosure flag, comprehensively considering research trends, the value of useful data, the progress of collaborative research, and the effectiveness of data sharing, and shall change the disclosure flag as needed with the approval of the head of the division.

7. Period of Grace for Publishing Useful Data

Useful data shall, in principle, be made public within two years since the data is generated or within one year since it is classified as useful data, whichever comes earlier. However, it may be handled case-by-case, if there is any contract or agreement at the time of carrying out research associated with the useful data. In a case where the period of grace for publishing Data of Research Papers is stipulated in the conditions for publication of the research paper, such data shall be published in accordance with such provisions.

8. Licensing Use of Useful Data

Efforts shall be made to appropriately present to the recipient of the data the license to use useful data, along with the disclosure flag (the scope of data disclosure), in order to protect the copyright of the data. In a case where a collaborative research agreement or a Material Transfer Agreement (MTA) has been concluded, a license to use useful data shall be sought accordingly.

To grasp the status of data use without identifying recipients, it is necessary to ask the recipients to provide their information (such as the recipients' names and affiliations), together with IP addresses and other records before downloading the data, in recognition of its being personal information. When a result of utilizing the data is presented in a research paper, the license to use the data shall require the user to provide information on the data providing source in the body or acknowledgement of the paper.

In preparing a data-use license, consideration should be given to allow the readers of a research paper to refer to the source data, such as by including a provision in the license to request that research papers contain identifiers to reference the source data or the data generator. For providing data to more recipients by allowing free access to the data, including data use by a program, a data-use license suitable for open data, such as a Creative Commons license, should be used.

9. Storage Period of Useful Data and How to Manage in the Institute Repository

As set forth in the Basic Policy, the useful data to be registered in the Institute Repository shall be kept for 10 years or longer. Due to various circumstances, such as the advance of science and technology and changes in trends, the utility value of the useful data may change during the storage period of the data. The data manager shall therefore endeavor to check the useful data registered in the Institute Repository on a regular basis and modify the disclosure flag and the data-use license so that they be most suitable for the data.

10. Use of Public Repositories

In the handling of publicly disclosed data, it is sometimes recommended to use public repositories depending on the research field or the type of the data. Publicly disclosed data should be registered in a public repository when it is deemed suitable to use public repositories, after comprehensive consideration of the conditions for publishing data set by the public repository, such as the period for data publication or data-use licensing. At that time, metadata shall be registered and disclosed in the Institute Repository.

11. Providing Metadata

Research institutes of all sizes in Japan and abroad have started to publish metadata on the Web, to form a global network of research data linked by metadata. In line with this trend, the AIP shall endeavor to provide high-quality metadata from very valuable research data in order to promote open science and the use of research data. For describing metadata, it is desirable to use standardized ontology and data items to ensure global interoperability. Researchers who have generated research data shall seek to create metadata of the world's highest quality by using the ontology and tools provided or recommended by and technological support from the division in

charge at the Head Office for Information Systems and Cybersecurity (hereinafter “ISC”).

In view of the costs and convenience of data management, it is most desirable to generate metadata when target data is generated. However, metadata is often provided at the time of selecting useful data or sharing/disclosing data, which entails more efforts to generate metadata. To address this issue, efforts shall be made to work on the development of a plan for handling research data, considering, at the stage of developing a research plan, the level of detail of metadata and the timing of providing it.

When metadata describes details on scientific findings such as experimental conditions and observed phenomena, the metadata itself can be highly valuable research data. In such a case, efforts shall be made to provide an appropriate disclosure flag or a license to use the metadata, similarly to other useful data. However, considering the spread of the standard technology used for queries, which enables users to automatically search for specific metadata from programs, etc., the metadata created as a data catalogue shall be published by providing a Creative Commons license with the approval of the head of the division to lower barriers hindering access.

Although research data and metadata annotating it are closely related, they may differ in the disclosure settings. It is possible to disclose metadata while keeping research data undisclosed, for example when the research content is proactively published by using metadata while preserving the intellectual property rights of the data. In addition, when data should be disclosed urgently but it takes some time to provide detailed metadata, it is acceptable to disclose data provided with a minimum level of metadata, such as the data generator and location.

12. Response to the Transfer of the Data Manager

The useful data registered in the Institute Repository shall, in principle, be managed responsibly by each manager or the head of the division to which the manager belongs to. When a data manager is transferred from the AIP to another center, etc., the head of the division shall take on the responsibility for the data management. The head of the division then may request another person in charge to manage the useful data registered in the Institute Repository. If there is any useful data for which nobody can take on management responsibility, the head of the division shall consult the ISC’s division in charge on how to manage the data.

13. Deletion of Data

As stipulated in the Basic Policy, useful data, including metadata, to be registered in the Institute Repository shall be kept for 10 years or longer. After expiration of the storage period, the data shall be deleted. When it is deemed desirable to extend the storage period of data owing to its scientific importance, however, the storage period may be extended on a temporary basis after approval by the head of the division.

Useful data that is deemed appropriate to delete prior to the expiration of the storage period may be deleted from the Institute Repository after approval by the head of the division, which requires justification for deletion of the data. Once the data is deleted, the reason for the deletion shall be registered in the Institute Repository, together with the same disclosure flag as before the deletion.

There is no need to delete both research data and its metadata at the same time. An appropriate

plan should be developed to delete each.

14. Evaluation of Researchers, etc. and Reward for Their Accomplishments

A person who generates, acquires, and actively shares research data with others may be entitled to receive an appropriate evaluation and reward for his/her accomplishments in this regard.