BERNOULLI INSTITUTE FOR MATHEMATICS, COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE

RESEARCH DATA MANAGEMENT POLICY

November 2018

Contents

1. General framework and principles

1.1 Purpose

The Bernoulli Institute (BI) produces fundamental and empirical knowledge that enables applications in numerous scientific and societal domains. This knowledge entails research data that needs to be archived for reasons of scientific integrity and reproducibility: e.g. the verification of scientific results, the protection of valuable datasets and analysis methods that lead to published experimental results.

The BI Research Data Management policy translates how these principles and the Faculty of Science and Engineering (FSE) and University of Groningen (UG) policy[1] must be applied by every BI scientist.

1.2 Background for the Bernoulli Institute data management

In the context of this document, data is defined as any collection of information based on which scientific conclusions are drawn and possibly disseminated. Research at the BI has a multidisciplinary nature and can produce data through observation, experiment, simulation and data processing. Therefore its volume, privacy restrictions, ownership, storage location can vary greatly.

Each researcher will decide what type of data will be collected and stored and what file format will be the most appropriate for storage. Following the UG Research Data Policy, which mirrors the international FAIR (Findable, Accessible, Interoperable and Re-Usable) principles,[2] data must be:

- accurate, complete, reliable, authentic and provided with metadata in the form of a text file describing the data sources in relation to the performed study;
- safely stored with minimum risk of loss and for at least 10 years;
- registered in a Current Research Information System (CRIS) (e.g. PURE);
- be available for review and study after completion of the research/researcher's departure;
- satisfying legal requirements, criteria for ethically sound research, partnership agreements and conditions laid down by research funding agencies;
- traceable, accessible and citable.

1.3. Ruling principles

To ensure that data have these characteristics, the following principles of verifiability, responsibility and "open, unless…" are essential. These principles translate the Netherlands Code of Conduct for Research Integrity that requests scientists to act honestly, scrupulously, transparently, independently and responsibly in their research practices, including the handling of data.[3] The implementation of the Research Data Management Policy entails that all researchers will describe how they adhere to the requirements and will provide rational considerations for the choices that are made.

1.3.1 Verifiability

Published results clearly have to show: upon what data the conclusions are based; how they are derived; where and how they can be verified. This means that all those involved in data collection and management at the BI will need to meet the standards of good data management and will have to act according to the procedures described in this document.

---

[1] based on the Netherlands Code of Conduct for Research Integrity (2018) and the VSNU regulations in this respect https://www.rug.nl/about-us/organization/rules-and-regulations/algemeen/gedragscodes-nederlandse-universiteiten/wetenschappelijke-integriteit

[2] "University of Groningen discussion paper transparent research environment/open science (2016)"

[3] Netherlands Code of Conduct for Research Integrity (2018), pp. 13-14

1.3.2 Responsibility

Primary responsibility for data management and for designing a project research data management plan according to the institute's template rests with the researchers themselves.

1.3.3 Open, unless…

To ensure that research results are disseminated as widely as possible following the primary publication process, the University of Groningen has adopted the principle that research data must be made openly available, unless ethical, legal or contractual obligations prevent this, such as:

- there is a reasonable chance that results can be commercially or industrially exploited;
- licensing and copyright protect the software developed by the BI researchers;
- data is personal/private and must be protected;
- claims are made for access to the larger collection of data from which the experimental results originate. Such claims usually cannot be granted as a research group may have its own planned publication schedule of future experiments on such a large (huge) data set. It is also common that data for international benchmarking events is made accessible fractionally, in, e.g., annual releases (in machine learning and pattern recognition);
- third parties may have rights concerning the data, prohibiting open access. In such instances, claimants must be first referred to such third parties;
- coauthors from other institutions may have their own data access policy and may ask that data is placed under embargo. The same goes for publishers that may ask for an embargo.

It is in the interest of the university that the data are accompanied by an explicit request to users (claimants) to always cite the article in which the open data were used and made available.

1.4. General Data Protection Regulation (GDPR)

The GDPR is a European regulation on data privacy that controls the processing by an individual, a company or an organisation of personal data relating to individuals in the EU. Personal data refers to any information that relates to an identified or identifiable living individual.
Different pieces of information, which collected together can lead to the identification of a particular person, also constitute personal data. The personal data to which GDPR is applicable is data that has been de-identified, encrypted or pseudonymised but can be used to re-identify a person.

Personal data that has been rendered anonymous in such a way that the individual is not or no longer identifiable is no longer considered personal data and does not fall under this rule. For data to be truly anonymised, the anonymisation must be irreversible. If identification is possible indirectly, through the connections between datasets, the data is regulated by the GDPR.

In case the Bernoulli researchers work with personal data, they should be able to demonstrate compliance with the regulation. To determine whether their data entails privacy issues, the Bernoulli researchers can perform a Data Privacy Impact Assessment (DPIA) before data are actually processed. The Research Data Office (RDO) at the UG offers assistance and courses with respect to the DPIA.

2. Bernoulli guidelines for research data management

2.1 Types of data and data format

The types of data collected by the Bernoulli Institute researchers include but are not limited to:

computer and source code, (human) behavioral data, self-report measures (questionnaires + open questions), EEG, eye-movements and pupil dilation, fMRI, video (human behavior, robotic, indoor scenes, outdoor scenes), other physiological data (e.g., heart rate), image data (external public, standard and adapted data sets, in-house data sets); annotation data, e.g., for visual recognition (handwriting, face recognition, scene text, including segmentation data); data produced by artificial systems (Robot behavior, AI or machine-learning model generators, Monte-Carlo type optimization processes, Cognitive models); data produced by simulation for experiments; simulation results, etc.

This data must be stored and retrievable. Therefore, researchers must be prepared to:[4]

- account for the data in new and ongoing research;
- develop and adopt appropriate procedures and processes for collecting, storing, processing, using and accessing research data during the research;
- document the agreements made on data management in the case of joint research projects or contract research where responsibility for data management rests in principle with the project coordinator;
- guarantee the integrity and security of their data;
- act in accordance with the Personal Data Protection Act and other legal and ethical rules;
- destroy research data in accordance under the terms of the research institution;
- develop a plan for accessing, reusing and storing research data at the end of the research project in accordance with the research institution's data management plan;
- budget the costs and time investment for data storage and management.

Data entries depend on the nature of the data, but in general they should provide:

- *Name of data collector*. The person(s) who have physically collected the data.
- *Name of data manager*. The person who initiates the data management process.
- *Access rights*. Rules describing who can access the data, beyond the default ones. Here open access and open data options can be defined.
- *Related publication*. The publication based on the data (possibly in pdf format).
- *Data descriptor*. Information on how to read and interpret the format of the raw data; scripts and programs necessary to produce the final results.
- *Program and model codes,* where applicable.
- *Raw data*. The actual raw data, if practically and legally feasible. For theoretical publications, this field can be empty.

To facilitate the organization and retrieval of data for future research as well as audit, researchers are advised to fill in Research Data Management Plans, i.e. forms that describe succinctly the research project, the data it used and produced, the type of storage, archive and access for it, etc.

[4] University of Groningen Research Data Policy, p. 5

2.2 Submission of Research Data Management Plans

A BI project that involves data should be associated to a Research Data Management Plan (RDMP).

Research groups are advised to record, using an RDMP, the way in which the group collects, stores and manages research data. A new version of the RDMP should be created whenever major changes in research occur due to inclusion of new datasets, changes in consortium policies or other factors. An RDMP template appropriate for group research will be made available as soon as possible.

Principal investigators in externally funded projects are advised to fill in an RDMP for the data used in these projects, especially if it is not recorded in the RDMP of their research group.

MSc/PhD students are responsible for submitting an RDMP for their project, to account for the data that they handle and which does not figure already in the RDMP of their research group. An RDMP template appropriate for group research will be made available as soon as possible.

2.3 Organization and data files naming

The BI does not dictate how to specifically organize the data archive nor does it provide preferred formats for data storage or naming of data files. Researchers are responsible to select the format most suitable to their research, attention being paid to the FAIR principles. Researchers are advised to choose file formats that are open format.

2.4 Data storage and archiving

Currently, the data storage alternatives for researchers at the BI range from long-term storage with automated back-up (e.g. UG servers, DANS Dataverse NL, etc.) to cloud-based storage and sharing (Unishare, GitHub, Y-drive, Dropbox, etc.), to short-term storage (PCs and laptops, external hard-drives, etc.), depending on the nature of their research and data.

Researchers are responsible with using storage and archiving options that align to the FAIR and GDPR principles. The BI staff will be informed as soon as the Research Data Management System developed through the Centre for Information Technology (CIT) will become available.

Raw data from third parties might be strictly regulated by these parties. The access privilege cannot be implicitly propagated to other parties after storing data in our repositories. A request for continued use must always be asked from the original raw data owner(s).

2.5 Data access

The author of a data entry controls sharing. It is their decision with whom the data and codes can be shared, following the "open, unless…"principle (see section 1.3.3). It is advisable that the raw and processed data should be made available to the research group leader, the BI RDMP coordinator, or the director of the institute should the author of the data entry become incapacitated. In case of conflicting research interests, a verification will be realized by an independent appointed researcher.

In the case of MSc and PhD students who are authors of data entries, they will share reading rights with their supervisors.

When an author of a data entry leaves the University of Groningen, *the group leader and the research director become responsible for the guardianship of the research data*. Access to the data can be further extended to other BI researchers, or to the world. Whether the researcher who leaves the BI can have access to the data is subject to the "open, unless…" principle and in any case requires a formal decision between the parties involved.

Access to the data by other parties at the UG for purposes of control proceeds in the following sequence: group leader, head of unit, institute director, dean, university board. While this in principle appears to be a natural order, representing the given responsibilities as regards (quality) control, there are situations in which individuals may have personal, scientific interest for accessing the data. Therefore, there needs to be a transparent system where the access is announced beforehand and logged in the system when it takes place. The BI does not yet have such a system but the Research Data Management System under development at the University of Groningen is expected to have this functionality. In cases where there are mixed interests, it may be advisable to appoint a separate committee which is allowed access for control purposes.

3. Control of research data management

Proper storage and archiving of data are part of the standards of good scientific practice and integrity and are implicit for the BI researchers. In addition, the BI will discuss with the (Under)Graduate School of Science and Engineering and the HR Department about the possibility that the submission of RDMP and storage/archiving of data become items during the Result and Development interviews for PhDs, postdocs and staff members and condition the award of diplomas to MSc students.