Research Data Management (RDM) Strategy

|  |  |
| --- | --- |
| **Research Group /**  **Scientific Unit** | e.g.**: EPFL SB IPHYS TNG** |
| **Director / Responsible**  Name, first name | **e.g.: *Jean-Luc Picard*** |
| **Strategy version**  v., Date | **e.g.: *v1.2, 2020-02-31*** |

**Template explanations (to eliminate from your copy)**

The *RDM Strategy template* is inspired by [SNSF template for NCCR RDM Strategy](https://www.epfl.ch/campus/library/wp-content/uploads/2019/10/RDM-Strategy-Coversheet-and-explanations.docx), but departs from it as it is mostly **targeted to single research groups**, even if it can be used for a consortium of groups. Divided in 5 parts related to the research data lifecycle, this templateis mostly useful in case a Research Group or Scientific Unit that wants to adopt common principles, practices and solutions for both **data and code management**.

First off, feel free to **adapt this template** to your own situation and needs. The best approach is to make it as useful as possible, by citing for instance exact tools, links to resources, addresses to persons or services responsible for specific tasks. Each section is accompanied by a guiding text and examples as an inspiration to craft your own *RDM Strategy*: you can eliminate these guiding texts once you finish adapting the template to your needs.

This is to be considered a **live document**, i.e. to be modified and improved as time passes and practices change, in order to always reflect your changes in strategic view or simply to reflect the updates to the mentioned practices and information. In case you have already a DMP or other specific documentation on data or code management for your projects, then this template becomes quite fast to compile: you might skip the parts labeled with an asterisk (\*) if you already have the specific documentation, but make sure to indicate here their location or link.

Contact the **Research Data Library Team** of EPFL at [researchdata@epfl.ch](mailto:researchdata@epfl.ch) for more information on the best ways to use this document, as well as Open Data practices, specific tools or platforms, policies on data and code, or to accompany your group in writing and later in implementing your *RDM Strategy*.

# Table of Contents

[A. Data Manager, data management organization and budget 3](#_Toc45724708)

[A1. Have you designated a Data Manager? 3](#_Toc45724709)

[A2. What are the projects of the research group affected by this RDM strategy? 3](#_Toc45724710)

[A3. How is the RDM coordinated? 3](#_Toc45724711)

[A4. What is the budget for RDM? 4](#_Toc45724712)

[B. Data collection and documentation 5](#_Toc45724713)

[B1. How is the global strategy of the group regarding data collection and documentation? 5](#_Toc45724714)

[B2.\* Which data / code do you handle? 5](#_Toc45724715)

[B3.\* Which documentation about data or code do you produce? 5](#_Toc45724716)

[B4.\* What is your file and folders naming convention? 6](#_Toc45724717)

[C. Storage and internal sharing 7](#_Toc45724718)

[C1. How is the global strategy of the group regarding storage and internal sharing? 7](#_Toc45724719)

[C2.\* How/where the data and code will be stored during research? 7](#_Toc45724720)

[C3.\* What is your internal data sharing policy? 7](#_Toc45724721)

[D. Security and Ethical issues 8](#_Toc45724722)

[D1. Are there any security or ethical issues? 8](#_Toc45724723)

[D2.\* What are the specific security issues? 8](#_Toc45724724)

[D3.\* How do you deal with the security issues? 8](#_Toc45724725)

[D4.\* What are the ethical issues? 9](#_Toc45724726)

[D5.\* How do you deal with the ethical issues? 9](#_Toc45724727)

[E. Data preservation and public sharing 10](#_Toc45724728)

[E1. How is the global strategy of the group regarding public data sharing? 10](#_Toc45724729)

[E2.\* On which repository will the data be made available? 10](#_Toc45724730)

[E3.\* Which data will be shared, retained, preserved and archived? 10](#_Toc45724731)

[E4.\* Under which conditions will the data be made available to the public? 11](#_Toc45724732)

[E5.\* Who are the owners of the copyright and Intellectual Property Right (IPR)? 11](#_Toc45724733)

|  |  |
| --- | --- |
| A. Data Manager, data management organization and budget | |
| A1. Have you designated a Data Manager? | Designate a “Data Manager”, by name, surname and email address, plus the %FTE devoted to the group’s data management. Otherwise, state if you don’t have designated one.  The Data Manager is responsible for the coordination of the RDM (Research Data Management) within the group and with possible external collaborators or services. |
| A2. What are the projects of the research group affected by this RDM strategy? Table 1   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Project  No** | **Project  Name** | **Research Fund** | **RDM Responsible**  **Name** | **Connected Project  No** | **Exists a  DMP** | | 1 | *Constellation* | *H2020* | *Geordi La Forge* | 2 | YES (link) | | 2 | *WP 5* | *SNSF* | *Data Manager* | 1,3 | YES (link) | | 3 | *Tritium* | *none* | *…* | 2 | NO | | 4 | *…* |  | *…* | … |  | | … | *…* |  | *…* | … |  | | List the projects of the research group:   * For each project, a person responsible for the RDM is designated: this can be the Data Manager or other members of the group or external collaborator. * For each project, all connected projects must be listed. Two projects are considered connected iff sharing some RDM tools or practices (ex. same file naming convention, same cloud storage, same ELN, etc.).   Delineation of RDM responsible persons for the projects can be dictated by the scientific organization of the group or the data management coherence demanded by the projects (tools / procedures / people).  **Note 1**: Instead of projects, you might also refer to possible work packages (WP) as submitted to funders: in this case, link here to the specific documentation.  **Note 2**: Even if other individuals are responsible for RDM of some projects, the Data Manager is finally responsible to supervise the execution and overall fulfilment of this RDM Strategy. |
| A3. How is the RDM coordinated? Table 2   |  |  | | --- | --- | | **Task** | **RDM Responsible Name(s)** | | 1. Information and training of the group members regarding RDM policy and requirements | *Data Manager* | | 1. Maintenance of the data management infrastructures, data backup | *Data Manager* | | 1. Maintenance, development or acquisition of software solutions for RDM | *Geordi La Forge,* ***Jean-Luc Picard*** | | 1. Preparation, curation, and documentation of datasets, quality control/self-assessment | *…* | | 1. Timely submission of datasets on repositories | *…* | | 1. Update of the projects’ DMPs | *…* | | 1. Update of the group’s RDM Strategy | ***Jean-Luc Picard,*** *Data Manager* | | Describe how data management will be coordinated between the researchers on different projects or with external collaborators.  Tasks can be distributed between the *Data Manager* [A1](#_A1._Have_you), other RDM responsible persons [A2](#_A2._What_are), group’s members or external collaborators or services. More than one person can be designated for a task.  **Note**: The individual responsibilities should be clearly delineated, but the Data Manager is finally responsible to supervise the execution and overall fulfilment of this RDM Strategy. |
| A4. What is the budget for RDM? | Provide the budget planned by the group for RDM, for each year of the projects. Cost categories might include for instance:   * Salaries (e.g. Data Manager) * Data processing (e.g. storage, software, facilities) * Data documentation (e.g. ELN license) * Data publication (e.g. paid data repository).   Costs are not bound to be already described in the projects’ submissions budgets to research funders.  When possible, refer to RDM costs included in the time / resources allocated for work packages (WP) submitted to funders, and link here to the specific documentation. |

|  |  |
| --- | --- |
| B. Data collection and documentation *\* B2, B3 and B4 should be completed only if the corresponding answer in B1 is “centralized” or “centralized with exceptions”. Otherwise, indicate the specific DMP(s).* | |
| B1. How is the global strategy of the group regarding data collection and documentation? | Choose among these 4 possible strategies:   * **Centralized**: all data generated in the frame of the group is stored centrally (e.g. server dedicated to the group accessible to all members and projects collaborators, institutional IT service provided to the group, …) * **Centralized with exceptions**: generally centralized but 1-2 projects cannot adhere to the general group’s strategy and have their own. List the exception(s). * **Project-specific**: every project defines its own data storage and preservation strategy (e.g. storage at the researcher level). * **Autonomous**: the practices are nowhere specified nor are they shared to everyone in the lab, apart from the supervision of the Data Manager.   **Note**: The organization can follow different strategies depending on the subject, e.g. “centralized” storage strategy and “individual” data preservation strategy. In this case, you can also say “Centralized with exceptions”. |
| B2.\* Which data / code do you handle? Table 3   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **Data / Code Type** | **Acquisition** (simulation, instrument, 3rd party, …) | **Format** (raw) | **Format** (processed) | **Metadata** (readme, logs, parameters file, …) | **Estimated volume** (total) | **Project No** | | Type 1: *Optical micrographs* | *Leica DM2700M with Suite v3.8* | *TIFF* | *PNG* | *XML file containing scale settings, resolution, acquisition setup, possible annotations.* | *300 GB* | *3, 4* | | Type 2: … | *3rd party collaboration* | *…* | *…* | *…* | *…* | *1* | | Type 3: … | *…* |  |  |  |  | … | | Type … | *…* |  |  |  |  | … | | Briefly describe the data or code you collect, observe, generate or reuse. The descriptions include the acquisition type, format, metadata, and volume.  **Note**: If no data nor code are generated, collected nor reused for a/some project/s, please state it and explain why. |
| B3.\* Which documentation about data or code do you produce? | Describe the documentation that is provided to enable secondary users to understand and reuse the data and code. Refer to [*Table 3*](#_B2.*_Which_data) whenever possible. The documentation might include for instance:   * README file(s) * Information on the used metadata standards * Record management (e.g. file-naming convention) * Log files or parameter files * Laboratory notebooks or protocols * Licenses   **Note**: If specific tools are needed to re-use the data, this needs to be documented and, if possible, the tools made available. |
| B4.\* What is your file and folders naming convention? | Describe the record management policy you want to use. This might include the way folders are nested and named, as well as any other convention on the name of files in those folders.  Try to provide general examples of these conventions.  **Note**: In case of databases, you can use this section to explain the the databases structure and the databases objects naming. |

|  |  |
| --- | --- |
| C. Storage and internal sharing *\* C2 and C3 should be completed only if the corresponding answer in C1 is “centralized” or “centralized with exceptions”. Otherwise, indicate the specific DMP(s).* | |
| C1. How is the global strategy of the group regarding storage and internal sharing? | Define (chose among the 4 possible strategies:   * **Unified**: all data generated in the frame of the group is stored and shared according to the same strategy (use of same repository, …) * **Unified with exceptions**: generally unified but 1-2 projects cannot adhere to the general group’s strategy and have their own. List the exception(s). * **Project-specific**: every project uses its own storage and internal sharing strategy * **Autonomous**: the practices are nowhere specified nor are they shared to everyone in the lab, apart from the supervision of the Data Manager.   **Note**: The organization can follow different strategies depending on the subject. You might also adopt a strategy during the projects and a different one after the projects end. |
| C2.\* How/where the data and code will be stored during research? Table 4   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **Storage** | **Capacity** (max) | **Data / Code Type** (Table 3) | **Project No** | **Link / Place** | **Back-up freq.** | **Responsible** | | Storage 1: *External SSD drive* | 4 TB | Type 1 | 3, 4 | PI desk | 1 / 2 weeks (manual) | PI | | Storage 2 | *…* | Type 2 | 1 | [www.cloud...com](http://www.cloud...com) (U.S.A.) | 1/ 20 sec (automatic) | Data Manager | | Storage … |  | Type … | … | … | … | … | | Describe the dedicated server, institutional IT service, laptops, cloud services, …. If available, give an estimate of your data storage capacity.  If external services are used, it is important that this does not conflict with the RDM Strategy or project-specific DMPs, especially concerning the issue of sensitive data.  Describe the back-up procedures (frequency of updates, responsibilities, automatic/manual process, etc.) |
| C3.\* What is your internal data sharing policy? | Describe the access for the group’s members or collaborators to the data handled in the frame of the group’s projects.  Be specific on (1) who has access to (2) what data, on (3) which storage, with (4) which access rights, at (5) what point in time and (6) how. Refer to previous sections when possible.  **Note:** Any security measures and the handling of sensitive data can be addressed later in section [D](#_D._Security_and). |

|  |  |
| --- | --- |
| D. Security and Ethical issues *\* D2, D3, D4 and D5 should be completed only if any “ethical” or “security” issue are listed in D1 that are not already documented in any DMP or authorization request.* | |
| D1. Are there any security or ethical issues? Table 5   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Project  No** | **Data / Code Type**  (Table 3) | **Security issue** | **Ethical issue** | **Ethical  Authorization** | | 1 | Type 3 | *Commercial cloud* | *Personal data* | *Requested (req. n. XXX, date XXX)* | | 2 | Type 4 | *None* | *Sensitive data* | *Obtained (auth. n. XXX, date XXX)* | | 3 | Type 1, 2 | *External HD* | *None* | *…* | | … | Type … | *…* | *…* | *…* | | Write here if the project(s) involve any possible security or ethical issues, in reference to the data / code types of [*Table 4*](#_C2.*_How/where_the).  It isn’t always clear whether one is even processing personal or sensitive data or not, which can happen in disciplines not usually related to sensitive, personal data. In the same way, one can be unaware of some issues due to data security. See possible examples of security and ethical issues in [D2](#_D2.*_What_are) and [D4](#_D4.*_What_are), respectively.  **Note 1**: security and ethics might be correlated but are not equal. Moreover, if no ethical issues are foreseen, it is good practice to just state it (use *None*). |
| D2.\* What are the specific security issues? | Along with the synthesis in [*Table 5*](#_D1._Are_there), describe all potential security issues while collecting, sharing, storing, re-using or analyzing the data.  Potential security issues might relate (but not limited) to the following risks:   * Data tampering * Data loss * Data theft or Eavesdropping * Theft of user identities or Password threats * Remote-Connection problems * Unauthorized access to data * Lack of accountability * Cumbersome User Management requirements * Uncertain data provenance |
| D3.\* How do you deal with the security issues? | For each of the potential security issue, describe whether and how you mitigate it, or if no solution is in place for risk mitigation. Include any processes, software solutions or hardware used to deal with the security issues. When possible, refer to [C3](#_C3.*_Who_has).  Methods to manage data security may include, for instance:   * Centralized authentication system * Data synchronization * Redundant backup strategy * Reduction of online content * Recurrent auditing. |
| D4.\* What are the ethical issues? | Along with the synthesis in [*Table 5*](#_D1._Are_there), describe all potential personal or other sensitive data to be collected, shared, re-used or analyzed within the project and the corresponding level of risks. Include the main processes, software or facilities for storage and processing of data subject to ethical issues.  These issues might be categorized as follows:   * **Personal data**: any piece of information that someone can use to identify, with some degree of accuracy, a living person, e.g. name and surname, age, spoken languages, home address, email address, identification card number, location data, Internet Protocol (IP) address, advertising identifier of personal smartphone, etc. * **Sensitive data**: special categories of personal data, e.g. racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, genetic data, biometric data, data concerning health or data concerning a natural person's sex life or sexual orientation, etc. * **Others**: organic tissues, experiments on animals, dangerous materials, military applications, etc.   In case of doubt, contact the [EPFL Research Ethics](https://www.epfl.ch/research/ethic-statement/). |
| D5.\* How do you deal with the ethical issues? | Describe how the data and code are handled within the project to address the potential ethical issues. Outline the measures taken to comply with data protection and, whenever possible, refer to C3. regarding how data access, sharing and security will be managed.  Methods to manage data ethics depend on the issue itself, and may include for instance:   * Anonymization * Hashing * Encryption * Physical location security |

|  |  |
| --- | --- |
| E. Data preservation and public sharing *\* E2, E3, E4 and E5 should be completed only if answer to E1 is “unified” or “unified with exceptions”. Otherwise, indicate the specific DMP(s).* | |
| E1. How is the global strategy of the group regarding public data sharing? | Define (chose among the 4 possibilities:   * **Unified**: all data generated in the frame of the group is published or externally shared according to the same strategy (use of same repository, …) * **Unified with exceptions**: generally unified but 1-2 projects cannot adhere to the general group’s strategy and have their own. List the exception(s). * **Project-specific**: every project uses its own data sharing strategy * **Autonomous**: the practices are nowhere specified nor are they shared to everyone in the lab, apart from the supervision of the Data Manager. |
| E2.\* On which repository will the data be made available? | In many cases (especially with SNSF funds) the chosen repository must conform to the [FAIR Data Principles](https://www.epfl.ch/campus/library/wp-content/uploads/2019/09/EPFL_Library_RDM_FastGuide_All.pdf#page=2) and/or must be maintained by a non-profit organization. If these conditions cannot be fulfilled, please explain why.  Tell whether or not you use data or code repositories or even archives that:   * Provide a DOI or other PID * Support standardized metadata (e.g. for keywords) * Are harvested and indexed by the major funders and search digital platforms * Use crosslinking to published articles or other data repositories   **Note**: the research groups are usually not expected by funders to use funds to create their own repositories, but rather to use already existing services. |
| E3.\* Which data will be shared, retained, preserved and archived? | Specify which data will be retained, shared, preserved and archived after the completion of the projects and the corresponding criteria (e.g. long-term value, potential value for reuse, obligations to destroy some data, etc.). While some criteria might be general, this might differ for different projects: please specify depending on your choice in sec. B1.  Please outline a long-term preservation plan for the datasets beyond the lifetime of the group. Comment on the choice of file formats and the use of community standards.  **Note**: explain if some data cannot be published due to specific condition, e.g. legal, ethical, copyright, confidentiality or other issues. |
| E4.\* Under which conditions will the data be made available to the public? | Usually, data and code need to be shared as soon as possible, but at the latest at the time of publication of the respective scientific output.  Apart from timing, check whether you want to make some restriction to the accessibility to some dataset (ex. for ethical reasons, embargoes agreed with funders, possible IPR issues, etc.). |
| E5.\* Who are the owners of the copyright and Intellectual Property Right (IPR)? | Please specify the licensing scheme you use for publishing your data and code collected or generated in the frame of the group.  For each project involving external collaborators or service providers, indicate IPR ownership and authorship agreements. In case of doubt, contact the [Technology Transfer Office](https://www.epfl.ch/research/services/units/technology-transfer-office/).  **Note**: EPFL owns the original data and code created while affiliated with it [[ETH Act, art. 36](http://www.admin.ch/opc/en/classified-compilation/19910256/index.html#a36)], but the authors can use it for research and commercial IPR claims. |