CNRS: DMP CNRS (english)

Data description and collection or re-use of existing data

Research output description

Recommandations:

- Give details of the type of data: for example digital (databases, spreadsheets), textual (documents), images, audio, video and/or composite media.
- Give details of the objectives for which the dataset was produced.
- Indicate whether you will be considering having a persistent identifier (PID) assigned to the data. PIDs should be assigned so data can be located and referenced reliably and efficiently. These also help keep track of citations and re-uses. A trustworthy persistent repository will typically assign persistent identifiers. The Inist is Datacite's French representative for assigning Digital Object Identifiers (DOIs). If a repository does not offer persistent identifiers for your data, it is possible to obtain one from the OPIDoR PID (*French link*) platform.

Will existing data be reused?

Recommandations:

- Reusing existing research data can be a real advantage, particularly if data are expensive to produce in a given context. You can consult directories like Cat OPIDoR (*French link*), re3data and FAIRsharing to find repositories where such data may be archived. You are recommended to contact a local data management cluster for help with this task if required at this stage.
- Indicate the nature and characteristics of the data. If you have been re-using data, give details of the source of this data and, if applicable, indicate the URL or identifier the data originated from.
- Indicate whether there are any constraints on the re-use of pre-existing data such as restrictive licences, costs and so forth.
- If applicable, give a brief account of the reasons why using existing data was considered but not taken further.
- To help with the legal aspects of re-using data you can consult the flowchart provided by the Institut Pasteur (*French link*).

How new data will be collected or produced?

Recommandations:

- Explain the methodologies or software to be used to collect or produce new data. Indicate the latest version of any software used. If appropriate, give details of how the data will be generated or collected and the tools and processes to be used.
- Explain how the provenance of the data will be documented.
- Data acquisition and collection must be well documented to be trustworthy. In some cases, this can be made easier by using tablets or electronic laboratory notebooks (ELNs) in the field. The CNRS offers an ELN service based on the eLabFTW solution and further information can be found on the CNRS intranet page (*French link*). Other tools like Collec-Science can also be used to record observed readings and metadata.

Documentation and data quality

What metadata and documentation (for example way of organising data) will accompagny the data?

Recommandations:

- Particular attention needs to be paid to the choice and quality of standards and metadata. These are essential to make sure the data can be understood, trusted and re-used. It is recommended that you contact data management clusters and thematic reference centers for all such tasks. Scientific teams can also benefit from the skills and expertise developed by the CNRS's transversal technology networks (*French link*).
- Indicate the metadata to be provided to help search for, identify, understand and re-use the data. Specify whether such metadata will be generated automatically or manually.
- Indicate the metadata standards to be used (DDI, TEI, EML, MARC, CMDI, etc.). It is preferable to use the community metadata standards that are available. You can access and download thesauri, ontologies, controlled vocabularies and other terminology repositories with the FAIRsharing tool and via the Loterre (*French link*) platform.
- Indicate how data will be organised during the project with reference to naming conventions, version control and folder structures, etc. For guidance on the consistent management of your files, you can consult the 'Comment bien nommer ses fichiers' (*French link*) guide on the DoRANum platform.
- Give thought to the documentation required to enable the data to be re-used. This could include information on the data collection methodology, the analysis procedures and methods used, the definition of variables, units of measurement and so forth.
- You should take into account how this information will be obtained and stored. This could be in a database with a link to each file, in a 'read me' text file, in file headings, in a code book or laboratory notebook.

What methods will be used to ensure their scientific quality?

Recommandations:

- Demonstrate that all work on obtaining data is controlled. For example, this could include controlling the functional chain of an analysis (pipettes, balance, analysis equipment) or compliance with a survey protocol (representative sample).
- Explain how the quality and conformity of data collection will be controlled and documented. For example, you could give

details of processes like calibration, repetition of samples or measurements, standardised data capture, validation of data entries, peer review or representation based on controlled vocabularies.
- It is also important to ensure the data are properly structured (file structure and format), comply with recognised standards and benchmarks and contain additional information (documentation). All of this also guarantees the quality of the data.
- The 'Traçabilité des activités en recherche et gestion des connaissances' (*French link*) guide gives details of best practices in this area and you can also contact the CNRS's Quality In Research network (QeR, *French link*).

Legal and ethical requirements, codes of conduct

If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

Recommandations:

Please refer to the CNRS Data Protection Department's recommendations on processing personal data.

How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

Recommandations:

- Explain the planned conditions of access to the data. Will the data be freely accessible or will there be restrictions? Give details of the latter if applicable. This flowchart (*French link*) from *Ecole des Ponts ParisTech* or the INRAe's equivalent (*French link*) will help you answer these questions.
- The deadlines for filing a patent justify the implementation of an embargo. Specify the reasons for any embargo and also the duration thereof.
- Make sure your consortium agreement covers such issues of data access control rights in the framework of multi-partner projects and/or involving shared data ownership. The responsibilities of each partner as regards all research products should be clearly defined.
- Indicate whether intellectual property rights (for example Database Directive, *sui generis* rights) apply. If this is the case please explain which rights and how this will be managed. To find out more about the legal framework for data, see the 'Guide d'application de la Loi pour une république numérique pour les données de la recherche' (*French link*).
- Indicate whether any restrictions apply to the re-use of data supplied by third parties.

What ethical issues and codes of conduct are there, and how will they be taken into account?

Recommandations:

- Determine whether ethical issues will impact how data are stored and transferred, who can visualise or use them, and what retention periods will be applied. Demonstrate that these aspects are taken into account and planned for.
- Adopt national and international codes of conduct and institutional codes of ethics. You should also check whether a review of practices (for example by an ethics committee) is required as regards data collection in the framework of a research project.
- To find out more about ethical and responsible practices, you can contact the CNRS ethics committee (*French link*) and consult reference guides like 'Pratiquer une recherche intègre et responsable' (*French guide*) and 'The European Charter for Researchers'.
- For guidance on best practices in terms of ethics and scientific integrity, consult the 'Charte déontologique du CNRS' (*French link*) and 'The European Code of Conduct for Research Integrity'.

Data processing and analysis

How and with what resources will the data be processed / analyzed?

Recommandations:

- Explain the processing carried out on raw data from acquisitions and collections.
- Specify whether format conversions were carried out.
- Indicate the tools or software infrastructures used.

Storage and backup during the research process

How will data be stored and backed up during the research?

Recommandations:

- Describe where the data will be stored and backed up during the research process and how often the data will be backed up (daily, weekly, monthly). The data should be stored in at least two separate locations. The basic principle regarding this is that data should be copied onto different media from those used when producing them. If possible, there should be 2 copies on 2 different media including 1 at a remote location.
- It is advisable to use robust storage systems with an automatic back-up function in shared infrastructures like data centres, mesocentres or the IT services of the host institution involved. It is advisable not to store data on laptops, external hard drives or storage devices like USB sticks.
- Evaluate the frequency at which data is updated and accessed to determine the right frequency for back-ups to be carried out.
- Evaluate the requirements and resources as regards data storage and backup. The possible environmental or financial impacts and costs generated by data processing and handling should be taken into account. It is advisable to work with an IT team on this before the project starts.

- Make sure that only data that deserves to be stored is actually stored as it can sometimes be cheaper to regenerate data on demand than to store them.
- Explain how data will be recovered in the event of an incident.
- Explain who will have access to the data during the research process and how access to the data is controlled, particularly in the framework of collaborative research projects.
- Data protection should be fully taken into account particularly if the data involved are sensitive, for example personal data, politically sensitive information or trade secrets. Describe all main risks involved and how these are to be managed.
- Explain the institutional data protection policy to be implemented. The CNRS information systems security policy (*French link*) gives information on data hosted at the CNRS.

Data sharing and long-term preservation

How will data be shared?

Recommandations:

- Give details of data formats i.e. the way data is encoded for storage which is usually reflected by the file name extension (e.g. PDF, CSV, DOC, TXT, HDF or RDF).
- Explain the choice of formats. Certain choices may be determined by the expertise of the organisation's staff members, a preference for open formats, the standard format accepted by data repositories, widespread use of given formats in a research community or the software or equipment to be used.
- Standard and open formats should be given priority because they facilitate the sharing and long-term re-use of data. Several catalogues provide lists of these 'preferred formats'. For long-term conservation purposes a list of formats eligible for archiving (*French link*) at the CINES is available on its institutional website.
- Give details of volumes expressed in terms of the storage space required (bytes) and/or the number of objects, files, rows and columns.
- Explain how the data will be made available and shared. For example, the data could be deposited in a trusted data repository, indexed in a catalogue, accessed using a secure data service or through direct processing of data requests and indeed any other suitable mechanisms.
- Explain when the data will be made available by indicating the expected publication deadlines. Explain whether use of the data will be made exclusive and, if so, for what reason and for how long. Indicate whether sharing the data will be postponed or restricted for example for publication reasons or to protect intellectual property or patent rights.
- Indicate who will be able to use the data. If access needs to be restricted for certain communities or an agreement for data sharing needs to be imposed, you should explain how and why. Also the measures that will be taken to circumvent or minimise these restrictions should be given.
- Data access and re-use licences should be considered if required. This list (*French link*) provides help in identifying the most suitable licence for your requirements. The License Selector and Choose an open source license tools are also helpful for making a choice on this point.
- Describe the foreseeable uses and/or users of the data in a research context.
- Indicate where the data is to be deposited. Catalogues like Cat OPIDoR (*French link*), re3data and FAIRsharing help researchers identify data repositories. Care should be taken in choosing your repository based on the data you will be depositing and any recommendations from the organisation providing your funding. You should also choose a repository that complies with CoreTrustSeal criteria (*French link*). The data management clusters mentioned above can help you with this process. You should give details of how the repository policies and filing procedures (including metadata standards and costs) were verified.
- If no trusted thematic repository is found, it is recommended that you deposit your data in the CNRS institutional space on the Research Data Gouv platform. The 'Where to publish your data' flowchart (*French link*) provides help with making sure your data is eligible.
- Indicate whether potential users require specific tools to access and (re)use the data. The lifespan of the software required to access the data should be taken into account.

How will data be long-term preserved? Which data?

Recommandations:

- Data created in the framework of a public establishment like the CNRS are public archives and thus should be sorted and selected with a view to their conservation in the national or regional public Archives.
- Give information on how long the data should be permanently archived. The DoRANum training material (*French link*) and the 'Référentiel de gestion des archives de la recherche' (*French link)* give details of the recommended data retention periods. The planned funding arrangements should also be indicated.
- Identify the data destined for long-term storage. Some data must be kept to ensure traceability, justify the anteriority of a research discovery or to preserve their heritage and scientific value. Conversely, other data should be deleted in compliance with the current regulations.
- Repositories do not provide long-term preservation facilities so it is advisable to choose a long-term archiving platform like the CINES for this aspect. If the data are not deposited on this type of platform then specify how it will be preserved over the long-term.
- A few main data selection principles should be applied to permanent archiving and you are recommended to contact the CNRS archive services regarding this.
- Give details of how the data to be archived will be selected and specify the selection criteria used for this.