# Data Management Planning

(Check <https://confluence.csiro.au/display/RDM/Data+Management+Planning> for updates and guidance.)

**Introduction**

This is a guide to help you to consider the data management requirements for your project. Data management planning will help you to manage your data for your team's use, meet funder requirements, and allow data sharing in the future. Managing and documenting your data throughout its life cycle will assist you to understand your data and help others to use your data. Planning your data management requirements at the beginning of a project and updating it during the project will ensure that infrastructure requirements are there when needed. Increasingly journals are requiring that research data is published. Undertaking data management planning will make sure it's easier to publish it when you need to.

Every project should have a person or role nominated to be the **data custodian** for the project. Make sure your planning documents this decision.

Consider from the beginning of the project the intention for the dissemination or long term archival of the data and software outputs of the project after the project concludes.

The factors that you should consider have been divided into 5 broad areas.

Share and discuss each factor with your project team to communicate and set in place the processes and guidelines for your team. Review and update the plan regularly in consultation with your research partners during the course of the project.

Consider these data management planning aspects:

* 1. Constraints and obligations
* 2. Access
* 3. Description
* 4. Processes
* 5. Storage and compute.

Contact the Research Data Support (RDS) Team via [researchdatasupport@csiro.au](mailto:researchdatasupport@csiro.au) for any assistance.

## About this project:

Project O2D Number: OD-201788

Project Name: New methodology for increasing yield in canola

Business Unit: Agriculture and Food

Project leader: Ian Greaves

Data Custodian: Anyu Zhu

## 1. Constraints and obligations

**These aspects will impact your ability to do your research and the way that you intend to publish or share the results. They will influence the answers to the other aspects in your project's data management planning.**

From the beginning of your project you need to consider your intentions for the dissemination or long term archival of the data and software outputs of your project after the project concludes.

* Data produced by CSIRO should be open and published by default in accordance with [Australian Government policy](https://www.pmc.gov.au/sites/default/files/publications/aust_govt_public_data_policy_statement_1.pdf)[[1]](#footnote-1). Data should be published unless there are legal, contractual or privacy requirements that prevent this.
* You need to plan for the constraints and conditions that will have an impact on whether you can publish your data as open access.

**Use this checklist to document your project's constraints and obligations. Choose those that apply to your project**

|  |
| --- |
| 1.1 Using data that you are bringing in from outside the project? |
| You need to document the ownership and conditions of use.  Create and keep an updated list of data that the project is planning to bring in or acquire. For each item include:   * + Which organisation owns the data?   + Will the acquired data need attribution and how should it be cited?   + Are there intellectual property considerations associated with hosting, using or sharing the data?   + What are the licence conditions? Document any conditions of use. Do they permit your intended reuse? Are there restrictions on publishing any derived data? When negotiating use conditions for acquired data, try to negotiate open conditions and document the results.   + Will any inherited licences restrict your licence choices on any derived data? |
| **Which organisation owns the data?**  CSIRO owns the data, however the canola lines used were provided by Nuseed. We are free to use the material for a transcriptome analysis and publication however we cannot distribute the physical material or produce new cultivars for distribution.  **Will the acquired data need attribution and how should it be cited?**  I have no idea what this means however transcriptome data is normally cited on the Gene Expression Omnibus (GEO).  **Are there intellectual property considerations associated with hosting, using or sharing the data?**  There are no constraints on how we use the data and are free to publish what we find. However as stated before the material is covered by an MTA which restricts our ability to share the physical material with others or use it to produce new cultivars  **What are the licence conditions? Document any conditions of use. Do they permit your intended reuse? Are there restrictions on publishing any derived data? When negotiating use conditions for acquired data, try to negotiate open conditions and document the results.**  See above. We may be asked by the company to remove cultivar names and replace with NX\_\_\_\_. This would not impact our work at all if this was required.  **Will any inherited licences restrict your licence choices on any derived data?**  Not that I am aware of. |
|  |

|  |
| --- |
| **1.2 Using code libraries, tools or third party code acquired from outside the project?** |
| You need to document the conditions of use.  Keep a list for each item. For each one include:   * Any conditions of use. Do they permit your intended use and reuse? * All licences and conditions, including any compatibility issues such as sharing and access by third parties. |
| We were given permission to use the material for 2 years. As stated above CSIRO has ownership over the data but not the material itself. |

|  |
| --- |
| **1.3 Is this project generating data and software?** |
| You need to plan the ownership and proposed conditions of use.  Include:   * Which of the collaborating organisations for this project will own the data and software generated by this project? * Which organisation will host published data and software after the project is finished? Will they assign a persistent identifier? * Who will be responsible for releasing software or data? * Is there any reason why this data and software will not be made available publicly at the end of the project? * What licensing conditions will need to be in place for sharing the generated or derived data or software or code? * What licence is intended to be used for:   + released data? See [Data Deposit checklist](https://confluence.csiro.au/display/daphelp/Data+Deposit+Checklist)[[2]](#footnote-2)   + released software (including code)? See [Software publication release process](https://confluence.csiro.au/display/OSS/Software+Release+Process+Home)[[3]](#footnote-3) * Which people, organisations or groups will be given credit for generated data or software? Consider how you want your generated data and software attributed.   You may need to consult with:   * [Legal](http://my.csiro.au/Support-Services/Legal.aspx)[[4]](#footnote-4) for Legal advice * [Business Development & Commercial](http://my.csiro.au/business-units/development/business-centre)[[5]](#footnote-5) for Contracts advice |
| CSIRO will own the data and all software used in this project is freely available. Published data will be hosted by the GEO database which is free and enables the data to be available to the public. The data will also be uploaded onto CSIRO’s DAP. CSIRO will be responsible for releasing the data. However Nuseed will be consulted before this happens. We will use the most open licensing agreement for the data. |

|  |
| --- |
| 1.4 Ethics and sensitive data**.** |
| Ethics: Does your project involve animal or human research including social and food and health research? See [Ethics](http://my.csiro.au/Support-Services/Research-Ethics-in-CSIRO.aspx)[[6]](#footnote-6)   * If so, have you applied for/or received ethics clearances? Document your ethics clearance documentation number. * Document any ethical clearances conditions on your research data. Are there access constraints? For example will the data need to be de-identified prior to publication? Will it only be able to be shared with specific parties?   Sensitive Data: document any issues that will need to limit access or otherwise affect your ability to share or publish your data due to confidentially or other issues that have not required ethics clearances. See [ANDS sensitive data guidance](http://www.ands.org.au/__data/assets/pdf_file/0010/489187/Sensitive-data.pdf)[[7]](#footnote-7) |
| Not Applicable |

|  |
| --- |
| **1.5 Other issues** |
| Document other issues that will affect the data and software management for this project. Check for:   * Any other terms in any project contracts that will affect data management that was not covered above. * Any policies relating to any instruments you use. * The data and software sharing or publication requirements of your funding provider. * Any other issues to do with patents, copyright and other IP restrictions. * The submission guidelines of preferred target journals for requirements regarding making data and software available. Any CSIRO or government policies or legislation or policies of collaborating organisations that apply. |
|  |

## 2. Access

**THIS ASPECT WILL INFLUENCE YOUR STORAGE AND TRANSMISSION CHOICES.**

Given the constraints and obligations documented in the previous section consider who will need access to the project research data at each phase of the project.

|  |
| --- |
| **2.1 During the project** |
| Document who will need to be able to access the data storage, computing facility or code repository including CSIRO colleagues and/or external collaborators during the project and their level of access (eg. read only or read/write).   * What level of security is required, for example password protection or encryption for confidential data? * Will a data sharing agreement be required?   Suggestion: List project members, specify if they are CSIRO or external and specify their required access levels. This will be useful when requesting storage. |
| Ian Greaves - Unrestricted CSIRO  Anyu Zhu – Unrestricted CSIRO |

|  |
| --- |
| **2.2 After the project** |
| Document who will need to be able to access the data storage, computing facility or code repository including CSIRO colleagues and/or external collaborators after the project ceases and their level of access (eg. read only or read/write)   * What level of security will be required if any, for example password protection or encryption for confidential data?   If the data is to be made publicly available which repository will be used? When will it be available? Will a persistent identifier be available? |
| This data will be made publically available once it has been published. This will be done using the Gene Expression Omnibus database (GEO) |

|  |
| --- |
| **2.3 During any embargo** |
| After the project will any embargo need to be applied?  Document who will need to be able to access the data storage, computing facility or code repository including CSIRO colleagues and/or external collaborators during the embargo period and their level of access (eg. read only or read/write) |
| No |

|  |
| --- |
| **2.4 Other access issues** |
| Note any other access considerations. For example document the location of existing data that you plan to acquire. How will it be transmitted to where it will be used? If not transmitted how will you be accessing it? |
|  |

## 3. Description

**This aspect will influence your project ability for each project member to find and use the data during the project and then easily publish or otherwise share your data at the end of a project.**

Your project team will need to decide whether you will be using files and folders for the project's data or some sort of database. Whatever is decided you will need to also plan the description or metadata for the data at some level.

Use the following checklist for metadata planning:

|  |
| --- |
| **3.1 Data collection description:** |
| If not using a database how are you going to make your data collection discoverable? Answering this will help you share your data and publish your data later.   * Document any decisions relating to description of this data or software as a collection; where known, include any metadata standards that you need to comply with, and/or vocabularies that you need to use. |
|  |

|  |
| --- |
| **3.2 Folder/directory naming:** |
| If not using a database how are you going to help each other find the relevant sections of your data?   * Document decisions regarding folder and directory naming and structures if relevant. * Document any standards that apply. Include any metadata standards that you need to comply with, and/or vocabularies that you need to use. |
|  |

|  |
| --- |
| **3.3 File naming:** |
| If not using a database how are you going to help each other find relevant data files and the relevant versions?   * Document decisions regarding file names if relevant. * Document any standards that apply. Include any metadata standards that you need to comply with, and/or vocabularies that you need to use. * Document how versions will be indicated and tracked in your naming conventions (if using). |
|  |

|  |
| --- |
| **3.4 Fields and variables:** |
| How are you going to help each other make your data interoperable?   * Document joint decisions regarding field names and variable formats. * Document any standards that apply. Include any metadata standards that you need to comply with, and/or vocabularies that you need to use. |
|  |

|  |
| --- |
| **3.5 File formats:** |
| This will affect your storage and/or service selection and interoperability.   * What file formats will be provided or are intended to be generated– database, jpeg, xls, etc? Take into account the end users of your data, the tools that you and they will be using and any domain norms. * Are there any issues relating to the formats such as special software or hardware requirements – GeoServer, MapServer, OpenDAP, THREDDS? * Are the formats being provided or generated according to relevant standards? |
| File formats we are intending to provide will be the raw .fastq Files and processed files which will be outputted from a deseq2 analysis. A .CSV file will also be provided with information on files and samples. |

## 4. Processes

**These aspects will influence the reproducibility of your research and your projects outcomes**

|  |
| --- |
| **4.1 Version control:** |
| How will versions be managed? A joint understanding of version control processes will aid communication, save time and aid reproducibility. Decisions about managing versions will impact storage and access decisions during and after the project.   * Document how you will distinguish between versions of generated data or software. How will contributors or users know that the data or software has been updated? How will you get updates for acquired data and software? * For software what revision control process will be used? * Document which versions need to be retained and for how long. Have you ethics clearance compliance factors to consider? * How will retained versions be managed and preserved after the project has finished, and by whom and where? * If there is a service which delivers the data how will this be managed after the project finishes? How will it be funded? |
| Github will be used for managing all files and metadata files. In the scripts we will also comment on versions of software used. |

|  |
| --- |
| **4.2 Reproducibility:** |
| A joint understanding of work flow will aid communication, save time and aid reproducibility. Document:   * What are the proposed data collection and analysis workflows? * How are you going to document the processes used (or the provenance of your worked data)? * How will you document which software and version was used for analysis? * What will be the process for keeping an up to date list of software assets and dependencies? |
| Analysis of software:  Samples were be checked with fastqc on the hpc using a .bash script. This script will include comments to help with readability of the script  Samples will be aligned using STAR and the latest assembled genome for brassica napus. The bash script will include what version of the brassica napus assembled genome this is.  STAR will output read counts to genes along with reads not considered for mapping (unmapped, repetitive).  The .bam output will be used for SNP calling either using SAMTOOLS or FREEBAYES. Both available on the HPC  output .txt files will be loaded into R and used for an analysis pipeline using DESeq2. This will be done using R markdown which will allow both outputs and information to be placed in a notebook for records. |

|  |
| --- |
| **4.3 Resourcing:** |
| Document the resources and responsibilities that are estimated to be required for data management processes such as:   * data cleaning acquired data * standardising data from varied project inputs * preparing the data for publication * de-identifying or otherwise processing sensitive data * providing access to a service/database/data/software after the project has completed.   Estimating these factors early will ensure the project is adequately resourced. |
|  |

## 5. Storage and compute.

**Data loss will severely impact delivering your research.**

It is recommended that all participants in a project use the same storage location for ease of management. Using [CSIRO IMT managed storage](https://confluence.csiro.au/display/RDM/Storage+choices)[[8]](#footnote-8), if suitable, will reduce your data management resourcing requirements. The location of the storage selected needs to be planned and documented so the project participants are informed. The suitability of a storage location will depend on the project stage and will change. Consideration of the destination of the data and software outputs of a project after the project concludes needs to be planned.

|  |
| --- |
| 5.1 Requirements: |
| Document the project's requirements for storage and compute not already covered by the rest of your plan. Include:   * An estimation of the volume of data storage required. * An estimation of its rate of growth. * The type of data storage required. See [Data Categorisation into Storage Block Types](https://confluence.csiro.au/display/SC/Data+Categorisation+into+Storage+Block+Types)[[9]](#footnote-9). * Whether significant compute power (ie HPC High Power Computing) will be required. * Any other special needs (eg if there will be a lot of small files). * If you will be developing code and so need a project code repository. |
|  |

|  |
| --- |
| 5.2 Existing Storage: |
| Having a joint decision and information about which storage solution to use will save moving data around and losing data. Using IMT managed storage will reduce the potential for data loss.  Do you have **existing managed storage** which fully fits your project's requirements above (taking into account size requirements, access, backups, compute requirements, special format requirements) at all stages of your project?  If Yes. Where? List the locations with path(s) if applicable including project stage (eg data acquired from outside the project, raw data, working data, data to be published).   * Document how the data will be stored, backed up and secured if not on managed CSIRO IM&T storage. * Do the proposed storage and compute locations comply with access and governance and format considerations documented in your plan? |
|  |
| **5.3 Existing Processing:** |
| Thinking about computing resources early will enable those resources to be available when required.  If you require significant computing (HPC) do you already have access to HPC that fully fits your projects requirements?  If yes please specify where. Document available HPC facilities. . |
|  |

|  |
| --- |
| **5.4 Existing Code Repository:** |
| Deciding on and using a joint code repository will enable good code development practices.  Do you have an existing code repository which fully fits your project's requirements at all stages of your project? If Yes. Where?  If you need a code repository see <https://bitbucket.csiro.au/> |
|  |

|  |
| --- |
| **5.6 If no to any of above- use the decisions in your project's Data Management Plan to request storage and compute via** <https://bowen.it.csiro.au/> |

1. <https://www.pmc.gov.au/sites/default/files/publications/aust_govt_public_data_policy_statement_1.pdf> [↑](#footnote-ref-1)
2. <https://confluence.csiro.au/display/daphelp/Data+Deposit+Checklist> [↑](#footnote-ref-2)
3. <https://confluence.csiro.au/display/OSS/Software+Release+Process+Home> [↑](#footnote-ref-3)
4. <http://my.csiro.au/Support-Services/Legal.aspx> [↑](#footnote-ref-4)
5. <http://my.csiro.au/business-units/development/business-centre> [↑](#footnote-ref-5)
6. <http://my.csiro.au/Support-Services/Research-Ethics-in-CSIRO.aspx> [↑](#footnote-ref-6)
7. <http://www.ands.org.au/__data/assets/pdf_file/0010/489187/Sensitive-data.pdf> [↑](#footnote-ref-7)
8. <https://confluence.csiro.au/display/RDM/Storage+choices> [↑](#footnote-ref-8)
9. <https://confluence.csiro.au/display/SC/Data+Categorisation+into+Storage+Block+Types> [↑](#footnote-ref-9)