H2020 DMP Template

This is a data management plan (DMP) template for a H2020 project in computational materials science using the Materials Cloud Archive as a data repository and AiiDA to collect the full provenance of the calculations performed.
If you have any feedback or questions feel free to contact us at info@materialscloud.org.

Further H2020 guidelines on DMPs can be found here.

Please adapt the template to your project!

1. Data Summary

- What is the purpose of the data collection/generation and its relation to the objectives of the project?
- What types and formats of data will the project generate/collect?
- Will you re-use any existing data and how?
- What is the origin of the data?
- What is the expected size of the data?
- To whom might it be useful ('data utility')?

The *ab initio* calculations performed in this project are expected to generate large volumes of data, most of it is in the form of intermediate binary files, which can be straightforwardly regenerated from the original inputs.

In this project, calculations will be performed through the *Automated Interactive Infrastructure and Database for Computational Science* (AiiDA), a python framework for high throughput calculations and provenance tracking (www.aiida.net). AiiDA automatically stores all information required to reproduce the result of each calculation. On the output side, AiiDA is designed to strike a balance between the cost of storage and the cost of recomputing a piece of data. For example, by default, total energies, electronic band structures and log files are stored, while Kohn-Sham wave functions and charge densities are not.

Before submitting a calculation through AiiDA, all inputs (+ metadata, see 2.1) are automatically stored in a local database. AiiDA then transmits the necessary information to the target computer, which can be a remote supercomputer, a local cluster or the workstation of the researcher. AiiDA adds the calculation to the computer's job queue, monitors the status, and retrieves the results once the calculation is finished.

In this model, all data is generated on the target computer, while only the data intended for preservation is transferred back to the workstation of the researcher.

AiiDA records this data in the form of directed acyclic graphs. For a full specification of the data (+ metadata) stored, see the documentation of the ORM API: https://aiida-core.readthedocs.io/en/stable/apidoc/aiida.orm.html

Given that a simulation may take many hundreds (if not thousands) of CPU hours it is beneficial to the materials science community to be able to access these without having to recompute them.

2. FAIR Data

2.1 Making data findable, including provisions for metadata

- Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?
- What naming conventions do you follow?
- Will search keywords be provided that optimize possibilities for re-use?
- Do you provide clear version numbers?
- What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

The inputs and outputs of all calculations, including their full provenance, will be stored in the form of an AiiDA graph. The AiiDA graphs contain the full provenance of the calculations performed. They will be preserved in the form of AiiDA export files.
AiiDA export files contain the AiiDA database in JSON format as well as all files associated with individual nodes in compressed form. Their specification can be found here: https://aiida-core.readthedocs.io/en/stable/import_export/main.html#export-file-format
The AiiDA export files will be archived on the Materials Cloud Archive, a non-profit data repository for computational materials science which guarantees preservation of data sets for at least ten years after deposition, irrespective of future funding for the repository.

Records of the Materials Cloud Archive are versioned and a citable DOI is assigned to each record.

Researchers can find the deposited data via persistent links from the publications or by searching the Materials Cloud Archive.
Researchers can interact with the data online through

- Materials Cloud EXPLORE sections, which provide automatically generated interfaces for browsing of the uploaded AiiDA databases
- Materials Cloud DISCOVER sections, which provide user-defined interfaces to the uploaded AiiDA databases

AiiDA automatically generates the following standardized metadata for every calculation:

- Which inputs were used
- Which code was used
- Which computer was used
- Which outputs were produced
- Who submitted the calculation
- When the calculation was submitted

For a full specification of the metadata stored, see the documentation of the ORM API: https://aiida-core.readthedocs.io/en/stable/apidoc/aiida.orm.html

Data can be queried based on their content, metadata and history using

- the AiiDA Query builder, an intuitive python API https://aiida-core.readthedocs.io/en/stable/querying/querybuilder/usage.html
- the AiiDA REST API https://aiida-core.readthedocs.io/en/stable/restapi/index.html

2.2 Making data openly accessible

- Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.
- How will the data be made accessible (e.g. by deposition in a repository)?
- What methods or software tools are needed to access the data?
- Is documentation about the software needed to access the data included?
- Is it possible to include the relevant software (e.g. in open source code)?
- Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.
- Have you explored appropriate arrangements with the identified repository?
- If there are restrictions on use, how will access be provided?
- Is there a need for a data access committee?
- Are there well described conditions for access (i.e. a machine readable license)?
- How will the identity of the person accessing the data be ascertained?

No confidentiality agreements apply.
Where commercially licensed crystal structures were used as input, appropriate identifiers have been put in place to preserve provenance.

The data will be made accessible by deposition on the open-access Materials Cloud Archive repository which complies with the FAIR Data Principles. The Materials Cloud Archive is registered on re3data, and recommended by FAIRsharing and Nature Scientific Data.

Data and metadata will be licensed under the Creative Commons Attribution (CC-BY) 4.0 license.
AiiDA is licensed under the permissive MIT open source license, which puts very limited restrictions on reuse (namely: including the license and copyright notice).
Code written within this project will also be released under an open source license (MIT, BSD or GPL).

2.3 Making data interoperable

- Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?
- What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?
- Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?
- In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

The data produced in AiiDA - such as crystal structures, k-points, arrays of data, band structures, pseudopotentials - are stored following standardised formats. Data and metadata formats are described in the AiiDA documentation (see for instance [1]), and make data easy to reuse in other codes.

Efforts are in progress to increase the standardisation of metadata and ontologies. In particular, AiiDA is integrated with the Theoretical Crystallography Open Database (TCOD) and is able to export data to the TCOD format [2].
Moreover, AiiDA is part of the Open Databases Integration for Materials Design (OPTiMaDe) consortium, which aims to make materials databases interoperational by developing a common REST API. For more details on the API metadata specifications see [3].

Where specific formats are indispensable, a README file with the relevant mappings to more commonly used vocabularies will be added to the Archive record.

[1] https://aiida.readthedocs.io/projects/aiida-core/en/latest/datatypes/index.html
[2] A. Merkys, N. Mounet, A. Cepellotti, N. Marzari, S. Gražulis and G. Pizzi, *A posteriori metadata from automated provenance tracking: integration of AiiDA and TCOD*, Journal of Cheminformatics **9**:56 (2017)
[3]https://github.com/Materials-Consortia/OPTiMaDe

2.4 Increase data re-use (through clarifying licences)

- How will the data be licensed to permit the widest re-use possible?
- When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.
- Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.
- How long is it intended that the data remains re-usable?
- Are data quality assurance processes described?

Data and metadata will be licensed under the Creative Commons Attribution (CC-BY) 4.0 license, guaranteeing that everybody can find, access and reuse the data without restrictions.
AiiDA is licensed under the permissive MIT open source license, which puts very limited restrictions on reuse (namely: including the license and copyright notice).
Code written within this project will be released under an open source license (MIT, BSD or GPL).

Data deposition and publishing should happen ideally when the corresponding paper is published (or even earlier when appropriate, e.g. together with the paper pre-print), but in any case at the latest within one year from the publication of the corresponding paper.

3. Allocation of resources

- What are the costs for making data FAIR in your project?
- How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).
- Who will be responsible for data management in your project?
- Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

4. Data security

- What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?
- Is the data safely stored in certified repositories for long term preservation and curation?

[Please adapt if this does not apply to how you manage your data]
During research, data is stored on supercomputer premises, protected by access control lists (ACLs) and transferred to workstations using the secure copy protocol (scp).
No personal or other sensitive data are collected in this project.

Data will be stored on the /project file system of Piz Daint at the Swiss National Supercomputing Centre (CSCS) in Lugano. Its features include:

- redundancy provided by the GPFS file system
- daily snapshots for the last seven days
- versioning of files for the last 90 days

For details, see https://user.cscs.ch/storage/recovery/

AiiDA databases and file repositories stored locally are backed up weekly to a central backup server.

The AiiDA export files will be archived on the Materials Cloud Archive repository, which guarantees preservation of data sets for at least ten years after deposition, irrespective of future funding for the repository.

Temporary files residing on the (super)computers, where calculations were performed, will not be preserved.

5. Ethical aspects

- Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).
- Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

No confidentiality agreements apply.
Where commercially licensed crystal structures were used as input, appropriate identifiers have been put in place to preserve provenance.
No personal or other sensitive data are collected in this project.

6. Other issues

- Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

Acknowledgement
This form is based on a template provided by the Participant Portal H2020 Online Manual. In particular, the text in blue is taken from the template. For more details, see the Guidelines on FAIR Data Management in Horizon 2020.