



## D7.7

# Revised Data Management Plan

Cristiano Malica, Nicola Marzari

Due date of deliverable 30/06/2025 (**month 06**)

Actual submission date 30/06/2025

Lead beneficiary UBremen (participant number 6)

Dissemination level PU - Public



## Document information

Project acronym	MAX
Project full title	Materials Design at the Exascale
Research Action Project type	Centres of Excellence for HPC applications
EC Grant agreement no.	101093374
Project starting/end date	01/01/2023 (month 1) / 31/12/2026 (month 48)
Website	<a href="http://www.max-centre.eu">http://www.max-centre.eu</a>
Deliverable no.	D7.7

Authors	Cristiano Malica, Nicola Marzari
To be cited as	C. Malica, and N. Marzari (2025): Revised Data Management Plan (DMP) for MAX Centre of Excellence. Deliverable D7.7 of the HORIZON-EUROHPC-JU-2021-COE-01 project MAX (final version as of 30/06/2025). EC grant agreement no: 101093374, UBremen, Germany.

## Disclaimer

This document's contents are not intended to replace consultation of any applicable legal sources or the necessary advice of a legal expert, where appropriate. All information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user, therefore, uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission has no liability in respect of this document, which is merely representing the authors' view.

## Contents

<b>1</b>	<b>Executive Summary</b>	<b>4</b>
<b>2</b>	<b>Introduction</b>	<b>5</b>
2.1	About this document . . . . .	5
2.2	Materials Cloud, latest achievements and upcoming updates . . . . .	5
<b>3</b>	<b>Description of the data</b>	<b>9</b>
3.1	Types of data . . . . .	9
3.2	Format and scale of the data . . . . .	10
<b>4</b>	<b>Data collection/generation</b>	<b>10</b>
4.1	Methodologies for data collection/generation . . . . .	10
4.2	Data quality and standards . . . . .	11
<b>5</b>	<b>Data management, documentation and curation</b>	<b>11</b>
5.1	Managing, storing and curating data . . . . .	11
5.2	Metadata standards and data documentation . . . . .	12
5.3	Data preservation strategy and standards . . . . .	13
5.4	Data sovereignty: Materials Cloud Archive at CINECA . . . . .	13
<b>6</b>	<b>Templates for data management plans for projects using the Materials Cloud</b>	<b>14</b>
<b>7</b>	<b>Data security and confidentiality</b>	<b>14</b>
7.1	Formal information/data security standards . . . . .	14
7.2	Main risks to data security . . . . .	15
<b>8</b>	<b>Data sharing and access</b>	<b>15</b>
8.1	Suitability for sharing . . . . .	15
8.2	Discovery by potential users of the research data . . . . .	15
8.3	Data access and reusability . . . . .	16
8.3.1	REST interface and OPTiMaDe support . . . . .	16
8.3.2	Sharing of plugins workflows . . . . .	16
8.4	Governance of access . . . . .	16
8.4.1	Governance of data access . . . . .	16
8.4.2	Governance of code access . . . . .	17
<b>9</b>	<b>Relevant institutional, departmental or study policies on data sharing and data security</b>	<b>17</b>
	<b>Acronyms</b>	<b>17</b>
	<b>References</b>	<b>18</b>



## 1 Executive Summary

The MAX **Centre of Excellence (CoE)** addresses the needs of the materials modelling community by providing data, codes, expertise, and integrated services to support the development of novel materials in the exascale computing era. This document revises and updates the strategies adopted — see MAX **Data Management Plan (DMP)**, D7.2 — to implement a robust materials informatics framework for curating, preserving, and sharing data generated by the MAX flagship codes.

At the core of this framework is **AiiDA**, a Python-based infrastructure that enables the automation of complex workflows, full provenance tracking via a **Directed Acyclic Graph (DAG)**, and standardized data storage through code-specific plugins. **AiiDA** manages input/output parsing, HPC execution, and provenance archiving using structured data formats.

The **AiiDA** ecosystem also facilitates open sharing of data and workflows through plugin and workflow updates, a **REpresentational State Transfer (REST) Application Programming Interface (API)** powering the **Materials Cloud** platform, and export tools compatible with common data formats and ontologies. Long-term data preservation is ensured via the **Materials Cloud** Archive, which guarantees storage for at least 10 years and integrates with the Explore and Discover sections to support FAIR-compliant data dissemination.

MAX promotes open research data practices. Unless restricted for legal or ethical reasons, all data supporting publications will be deposited on the **Materials Cloud** Archive under open licenses, enabling unrestricted reuse. To assist researchers, DMP templates tailored to **AiiDA** and **Materials Cloud** users are provided (see Section 6 for details).

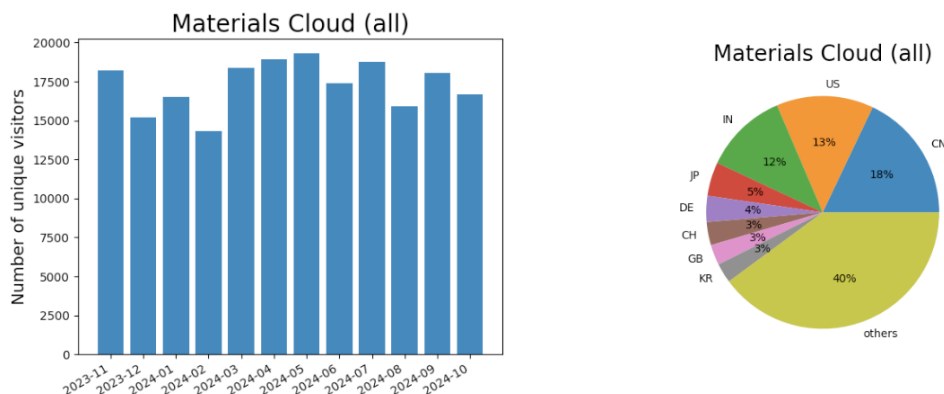


Figure 1: Number of unique page visits per month and per country, as on Oct 25, 2024. Materials Cloud uses Plausible (<https://plausible.io/data-policy>) to count its page views without tracking users or storing cookies.

## 2 Introduction

### 2.1 About this document

This document is deliverable D7.7 of the MAX project and briefly describes the types of data produced in the project, standards used for data, and how the data is being curated, preserved and shared. The basic types of data and the way they are organised follow the spirit of AiIDA. This is a living document and will be updated continuously in the course of the project. The acronyms used are summarised in the glossary at the end.

In this document, we first present a summary of the most recent advancements in data management practices within the MAX CoE. We then illustrate these developments through selected examples of scientifically relevant cases where the structured approach to data cataloguing and provenance provided by AiIDA and the Materials Cloud platform are expected to play a fundamental role. Finally, we revisit the core concepts of data generation, provenance tracking, and data sharing, building upon the general description of data handling provided in the previous version of the DMP. For the sake of completeness, the original and unmodified parts of the DMP from D7.2 are also explicitly kept in this document.

### 2.2 Materials Cloud, latest achievements and upcoming updates

The Materials Cloud platform (described in more detail in the following sections) is dedicated to sharing resources for computational materials science. It offers services for the archiving and dissemination of both raw and curated data, as well as tools for data analytics and educational resources. Over time, the platform has continuously expanded its functionalities and improved the overall user experience.

Materials Cloud currently receives nearly 20,000 visits per month from users across the globe, including most European countries, China, the United States, and India (see Fig. 1). The Archive section of the platform hosts over 30 million crystal structures, more than 13 million of which are associated with density-functional theory (DFT) self-consistent field (SCF) calculations and are freely available for download.

Figure 2: MC3D has a new and improved web interface and contains 33674 materials in the PBEsol dataset (in addition to the previous PBE-v1 dataset). Additionally, powder X-ray diffraction analysis is available for all structures.

Some of the recent updates include the QUANTUM ESPRESSO input generator<sup>1</sup> tool with improved k-point sampling/smearing protocols, as well as the addition of new examples to the interactive phonon visualizer<sup>2</sup> featuring common 3D crystals. A new cell2mol<sup>3</sup> tool has been introduced to help users interpret crystallographic data and retrieve various properties of molecular complexes and their components. Some of the Discover sections have also been updated. In particular, following the release of the version 1.3.0 of the Standard Solid-State Pseudopotentials (SSSP) in 2023, the Materials Cloud 3D (MC3D<sup>4</sup>) crystals database has been updated and the web interface has been redesigned. MC3D now features more than 33,000 materials in the PBEsol-v1 dataset, calculated using the PBEsol DFT functional, supplementing the prior PBE-v1 dataset (see Fig. 2).

**Materials Cloud** gives great importance to educational outreach and major improvements in this area have been achieved with the Lhumos<sup>5</sup> learning hub. The Materials Cloud Learn section will be replaced by Lhumos, making this educational section more intuitive and user-friendly. On April 3, 2024, the **Materials Cloud Archive** repository celebrated reaching the milestone of 1,000 publications. Furthermore, an automated service

<sup>1</sup><https://www.materialscloud.org/work/tools/qeinputgenerator>

<sup>2</sup><https://interactivephonon.materialscloud.io/>

<sup>3</sup><https://cell2mol.materialscloud.io/>

<sup>4</sup><https://www.materialscloud.org/discover/mc3d>

<sup>5</sup><https://alpha.lhumos.org/>

has been established to launch the OPTIMADE APIs<sup>6</sup> directly from Archive entries.

On June 23, 2025, the new version of the Materials Cloud Archive officially went into production (see Fig. 3), marking the completion of a major upgrade from Invenio<sup>7</sup> to InvenioRDM,<sup>8</sup> a turn-key data management repository developed by CERN. InvenioRDM is the same software used by Zenodo,<sup>9</sup> which currently hosts over 4 million records. The upgrade of the Archive to InvenioRDM introduces a range of new functionalities, improves interoperability and makes the platform more easily maintainable and scalable. Key new functionalities include the ability to share records with colleagues and peer reviewers before publication, monitor usage statistics by tracking the number of views and files downloads for each record, and the option to add authors and affiliations using ORCID and ROR IDs.

Additionally, the REST APIs available for all operations within the repository enhances its interoperability, allowing for smoother integration with other systems and tools. To further improve data accessibility and interoperability within repositories, efforts have been made to set recommendations for data repositories of the Swiss ETH domain on how to interpret and implement the metadata DataCite schema 4.5, released in January 2024. Finally, on January 2025, a New Discover app for the Materials Cloud 2D crystals database was released<sup>10</sup> and a new tool<sup>11</sup> was published that allows users to predict molecular electron density with machine learning.

---

<sup>6</sup><https://www.materialscloud.org/optimade>

<sup>7</sup><https://inveniosoftware.org/products/framework/>

<sup>8</sup><https://inveniosoftware.org/products/rdm/>

<sup>9</sup><https://zenodo.org/>

<sup>10</sup><https://mc2d.materialscloud.org/>

<sup>11</sup><https://rho-prediction.materialscloud.io/>

The screenshot displays the Materials Cloud Archive interface. The top navigation bar includes 'LEARN', 'WORK', 'DISCOVER', 'EXPLORE', and 'ARCHIVE'. A search bar and user profile 'valeria.granata@epfl.ch' are visible. The main content area shows a record titled 'materialscloud:2023.65' with a subtitle 'A Standard Solid State Pseudopotentials (SSSP) library optimized for precision and efficiency'. The record is published on April 13, 2023, version v11. The authors listed are Prandini, Gianluca<sup>1\*</sup>; Marrazzo, Antimo<sup>1</sup>; Castelli, Ivano E.<sup>1,2</sup>; Mounet, Nicolas<sup>1</sup>; Passaro, Elisa<sup>1</sup>; Yu, Jusun<sup>1</sup>; Marzari, Nicola<sup>1\*</sup>. The abstract describes a universal standard protocol for verifying pseudopotential libraries. The 'Files' section lists 'CHANGELOG.txt' and 'Files (35.6 GB)'. The 'Details' section shows a DOI of 10.24435/materialscloud-23-ym, resource type 'Dataset', publisher 'Materials Cloud', and language 'English'. The 'Citation' section provides the full citation text. A 'Share access' dialog box is open, showing a table of links with columns for 'Link title', 'Created', 'Expires', and 'Access'. Two links are listed: one created 3 seconds ago with 'Can view drafts' access, and another created 0 seconds ago with 'Can edit' access. Both links have 'Delete' and 'Copy link' buttons. At the bottom of the dialog, there are input fields for 'Short name for your link (optional)', 'Expiration date format: YYYY-MM-DD', and 'Can view drafts', along with a 'Create a new link' button and a 'Close' button.

Figure 3: The new version of Materials Cloud Archive using InvenioRDM provides various functionalities, including the ability to share access to records through links for editing or viewing.

Table 1: The primary codes that will be used as part of this project.

Name	License	Main Developers
AiiDA [1, 2]	MIT	EPFL, Robert Bosch
BIGDFT [3, 4]	GNU-GPL	CEA, UNIBAS
FLEUR [5, 6]	MIT	Jülich
QUANTUM ESPRESSO [7, 8]	GNU-GPL	SISSA, CNR, CIN
SIESTA [9, 10]	GNU-GPL	ICN2, BSC
YAMBO [11, 12]	GNU-GPL	CNR, CIN

### 3 Description of the data

Within this project, various open-source first principles simulation codes such as Fleur, SIESTA, Quantum ESPRESSO, YAMBO, BigDFT are being developed (see table 1). The materials informatics framework AiiDA has been designed for the support of many different codes through a plugin interface. Plugins for the codes within this CoE are already available in most cases. A general overview of the present status of the plugins can be found at <https://aiidateam.github.io/aiida-registry/> together with the contact of the reference person for each plugin and a link to the code repository. The page collects also numerous other plugins, for codes outside this CoE, which are being currently developed as part of other collaborations or by individual contributors.

AiiDA promotes advanced programming models leveraging python abstraction layers to disseminate advanced functionalities to arbitrary quantum engines (i.e., simulation codes). It provides a model of automatic data generation and storage, to guarantee provenance, preservation, reproducibility, and reuse. This platform is used to organise and coordinate thousands of simulations, it allows one to acquire and store a variety of heterogeneous microscopic data from the calculations that can be subsequently queried for the desired material properties. Details on the calculation execution such as the parallelisation scheme and execution time are also retained to empower performance optimisation. Furthermore, AiiDA allows for an automated design and implementation of complex workflows and task tracking, based on a python interface for job creation and submission. Leveraging the AiiDA engine, the inputs, results and computational procedures at each step of the workflows are collected and stored in a database. All plugins, workflows and the data produced through them are thought and designed to be openly shared as discussed afterward.

#### 3.1 Types of data

AiiDA provides automated solutions and various plugins for computer codes without a need for tuning code specific parameters. It stores the calculations, their inputs and their results (either parsed, extracted from Extensible Markup Language (XML) outputs or from text files with the appropriate dictionaries) in a database and its associated file repository. This data is generated from open-source electronic-structure material simulations codes that encompass key technologies such as all-electron, pseudopotentials, and localised basis sets, density-functional theory, time-dependent density-functional theory,

and many-body perturbation theory, multiscale/multiphysics modelling with a focus on quantum mechanics/molecular mechanics, solvation and electrochemistry, thermal/electrical transport, and complex magnetic properties.

The specific data to be stored in the database within the input and output nodes of a calculation and the files to be retrieved and stored in a local repository are determined by each code plugin. This follows the specific characteristic of each simulation code and the physics of the problem under study. The design of each calculation node is documented by the plugin developers.<sup>12</sup> AiiDA also allows one to use data from external open access databases of crystal structures for organic and inorganic compounds such as [Crystallography Open Database \(COD\)](#) [13], [Theoretical Crystallography Open Database \(TCOD\)](#) [14, 15] and [Inorganic Crystal Structure Database \(ICSD\)](#) [16] to obtain the input atomic coordinates of crystalline materials. The platform also offers the possibility, at the workflow level, to copy large files (for example charge densities) to a data storage facility for later reuse and to save into the database a symbolic link to such remote folder.

### 3.2 Format and scale of the data

AiiDA parses the input and output files mostly stored as text or XML and runs the calculations/codes on high performance computing platforms. The full provenance of each data object (inputs, outputs, calculations) is automatically stored in database in a format that enables the simulation results to be fully reproduced. The database has an associated repository with text and binary (machine-independent) files. We have developed uniform formats to define the most common raw and analysed data irrespective of the different plugins. These standards contain data in dictionary format, exportable for instance to plain [JavaScript Object Notation \(JSON\)](#) (for example, StructureData, Dict data types in AiiDA store metadata in python dictionaries within a database).

AiiDA however allows for the flexibility to define new data structures and formats which might be strongly code dependent. Currently we are using the PostgreSQL [17] open-source relational database to store our data. The size of the databases depends on the individual research project, with some existing examples counting of the order of tens of millions of records (with hundreds GB of occupied disk space). Currently we use applications like Jmol, Visual Molecular Dynamics (VMD), PyMOL, VESTA, XCrySDen and Blender to visualise 2D and 3D structures, and Matplotlib, Gnuplot and Mathematica for plotting the data.

## 4 Data collection/generation

### 4.1 Methodologies for data collection/generation

Data for the project are created and collected by using the AiiDA framework for the management of the simulations. AiiDA plugins and workflows have been written for different simulation codes in order to support the simulation with at least the codes used within this CoE, but also to support other codes available in the community.<sup>13</sup>

---

<sup>12</sup>As an example see <https://aiida-quantumespresso.readthedocs.io/en/latest/index.html>

<sup>13</sup>See <https://aiida-team.github.io/aiida-registry/>

By using AiiDA, the full provenance of all calculations is preserved from initial inputs to final outputs, as well as all steps along the way, in the form of a **DAG**. This allows any output data to be retrospectively checked for quality if there are questions about how it was generated. Workflows on the other hand provide a means of proactive quality assurance whereby a series of steps is designed and implemented by a domain expert and packaged as a workflow. A workflow can then be executed by experts and non-experts alike, with internal checks and heuristics that attempt to ensure the quality of data with respect to convergence and other relevant simulation parameters. Furthermore, by having a standard way of running particular calculations, it becomes much easier to compare and validate results.

Raw inputs and outputs from computer simulation codes are stored directly so that they may be re-parsed or manually inspected if necessary. Otherwise data are stored as standard, code-independent objects in the AiiDA framework (e.g., crystal structures, band structures, pseudopotentials,  $k$ -point paths, etc.) allowing for easy querying and manipulation of results from a variety of simulation software packages. All naming of these input and output files are handled internally by AiiDA and files can be retrieved for particular calculations by either issuing a query to match specific search criteria or directly by using the **Universally unique identifier (UUID)** of a known simulation.

## 4.2 Data quality and standards

As mentioned previously, the combination of persistent provenance and workflows is used to maintain consistency and quality. Our provenance model also acts as a form of documentation storing all the steps that lead to any result in the database.

One aspect of the project involves the dissemination of a library of, so called, pseudopotentials which contain information about the quantum mechanical properties of the outermost electrons (those relevant in chemical bonds) for the elements of the periodic table. Currently there may be many different pseudopotentials for each element, however this makes it difficult to compare calculations, and worse, some pseudopotentials are not accurate enough to give reliable results for some of the calculations they are being used in. By providing a standard set of pseudopotentials that have been thoroughly tested we alleviate both these problems. The data standards adopted and used in AiiDA are described in section 3.2. We are also involved in working on ontologies in collaboration with the **TCOD** team. During this collaboration, we have implemented exporters for calculations managed using AiiDA to the domain-specific ontology that is being built within the **TCOD** project, so that also calculation results (such as crystal energies or atomic forces) can be stored in a code-independent format [15].

## 5 Data management, documentation and curation

### 5.1 Managing, storing and curating data

In AiiDA, all data (calculations, their inputs and their outputs) generated by running high-throughput simulations on local or remote servers are naturally stored on those computers. Moreover, relevant inputs and outputs are persisted in the AiiDA repository, composed both of a folder-like structure and of a database. For the latter, we use PostgreSQL,

a powerful open source object-relational database. The format for storing data (depending on the specific type of data) is defined by the specific AiiDA data plugins, described in detail in the code documentation.

The data format of common objects (crystal structures, band structures, force constants, etc.) is the same for all objects of the same type, even if generated by different computer codes, to facilitate data exchange, queries, and the bridging of different simulation tools. Moreover, each data format is accompanied by data import and export functions from/to standard formats (for example [Crystallographic Information File \(CIF\)](#) files for crystal structures). Importers and exporters for commonly used formats such as ASE [18] (Atomic Simulation Environment) and PyMatgen [19] format have been developed. Other import/export capabilities can be transparently added through upon necessity.

Every data object is a node in the [DAG](#) where links between nodes keeps track of the data provenance (who generated the data, with which parameters, etc.), allowing for easy regeneration of the same data with the same inputs. Moreover, beside common metadata (user/owner, creation and last modification date, etc.) any further metadata can be attached to any node of the database (data and calculations). Also, AiiDA provides data sharing capabilities, both to share portions of the calculations database with selected groups of users and collaborators, and to export the data to public repositories like the [Materials Cloud Archive](#).

During research the databases are stored in the universities data centres with periodical backups stored on supercomputer premises. The implementation of the periodical backups is responsibility of the individual research project. As an example, the database used for the pilot 3 (*Dissemination of highly-curated computational materials data*) of MAX-2 WP5 and part of its file repository (small files) are stored on a server at [École Polytechnique Fédérale de Lausanne, Switzerland \(EPFL\)](#) while the remaining part of its file repository is stored on a [Centro Svizzero di Calcolo Scientifico - Swiss National Supercomputing Centre, Switzerland \(CSCS\)](#) server. The policy defining what a large file is depends on the application and is defined within the AiiDA workflows used to generate the data.

## 5.2 Metadata standards and data documentation

There are several key simulation codes that will be used for this [CoE](#) as described in [table 1](#). Typically a simulation is run by supplying one or more input files which along with the primary data of interest (be it a configuration of atoms in space, the electronic structure, a material property, etc.) will typically produce auxiliary data (metadata) which can vary greatly from software to software. To interface software with AiiDA, a plugin is written that converts AiiDA nodes (used as input) to the actual input files required by the code, and parses outputs allowing these to be stored in the database in a standard way, such that it can later be queried using the AiiDA [API](#).

The AiiDA code itself can be considered to be *data* in this context. AiiDA is fully documented both in the form of descriptions of functions that make up the [API](#) and as guides describing steps such as the installation procedure, configuring users, setting up codes, etc. The documentation is shipped with the code and can also be found online. <sup>14</sup>

---

<sup>14</sup>See <http://aiida-core.readthedocs.org/en/latest/>

To ease the deployment procedure a dockerfile [20] for AiiDA installation with default parameters has been realised.<sup>15</sup> This method will be also extended to each code with a plugin interface with AiiDA in order to readily install a working AiiDA and simulation code instance.

### 5.3 Data preservation strategy and standards

As part of the MARVEL National Centre of Competence in Research (NCCR) project [21], funded by the Swiss National Science Foundation (SNSF), a large storage allocation has been purchased at CSCS which is used to store and retain large files for a long-term period. The agreement with CSCS guarantees that the data will be stored for at least 10 years after their deposition. This storage service was already paid upfront, meaning that even if the MARVEL project were to end (see timeline in Fig. 4), CSCS would still store the existing data for 10 years after deposition. In the meantime we expect to obtain other funding opportunities to enable further preservation of the data. The Materials Cloud platform has been fully operative since February 2018. Entire AiiDA graphs or part of them could be directly shared through the Materials Cloud portal.

### 5.4 Data sovereignty: Materials Cloud Archive at CINECA

The Materials Cloud Archive database and curated data are stored at the Swiss National Supercomputing Centre (CSCS) in Lugano. To minimise the risks associated with data loss, and to ensure data sovereignty at the European level, the database and curated data will be duplicated and stored at the National Supercomputing Centre (CINECA) in Bologna, Italy on the HPC ADA cloud infrastructure. ADA cloud is an Infrastructure as a Service (IaaS) based on OpenStack Wallaby. Duplication and synchronisation will be performed using Rclone, an open source, multi-threaded command line program to manage or migrate content on cloud and other high-latency storage. Synchronised copies are checked for completeness and integrity by checking the files in the source and destination match. Rclone, combined with our own scripts, compares sizes and MD5 and logs a report of files that don't match. Synchronised data will be backed up in accordance with the CINECA backup procedure. The CINECA HPC infrastructure will guarantee preservation of data at least for the duration of the MAX 3 project.

The data retained, being generated with AiiDA, will include full provenance of all simulations carried out. Some large output files may not be preserved if they are judged to be easy to reproduce and unlikely to be needed after the simulation has completed. This policy is code-dependent and sensible defaults are defined within each AiiDA plugin, but can easily be changed by the user or group who runs the simulations and generates the data.

Taken together, the Materials Cloud and AiiDA aim at enabling researchers to trivially adopt aspects of the Findable, Accessible, Interoperable and Re-usable (FAIR) principles as laid out by Force11. Any calculation run through AiiDA is automatically interoperable (we define an open standard and provide tools to convert data to other libraries and are adding Open Databases Integration for Materials Design (OPTiMaDe) [22] compatibility) and re-usable (by virtue of the provenance tracking and workflows). The Ma-

---

<sup>15</sup>See [https://hub.docker.com/r/aiidateam/aiida\\_core\\_base/](https://hub.docker.com/r/aiidateam/aiida_core_base/)

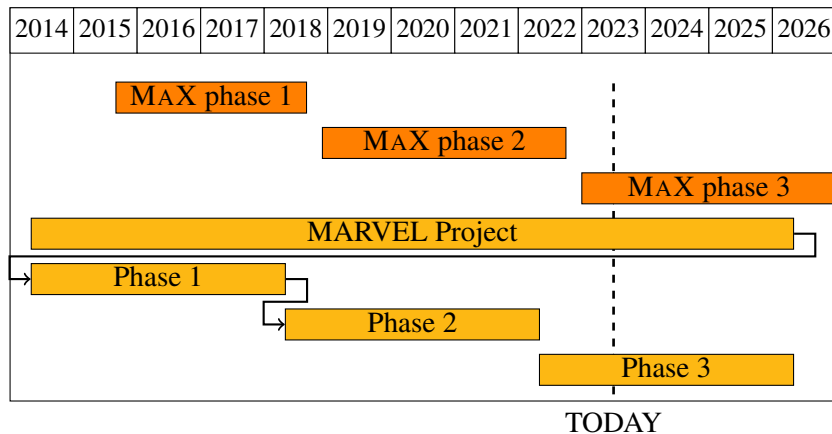


Figure 4: MAX project timeline, compared with the Swiss NCCR MARVEL project. The current project is indicated as MAX phase 3.

terials Cloud on the other hand ensures that the data is both findable and accessible. Section 8 outlines further details.

## 6 Templates for data management plans for projects using the Materials Cloud

As part of making research data FAIR, HE projects that produce, collect or process research data need to provide a DMP. A DMP describes every step of the data lifecycle, starting from data creation, to transformation, analysis, storage, sharing and reuse.

Researchers who use AiiDA and/or share their data on the Materials Cloud Archive can easily redact a data management plan using the templates provided on the Materials Cloud platform. Those templates contain a set of questions that the authors should address with a level of detail appropriate to the project, and data management details for projects using AiiDA and/or Materials Cloud are provided.

## 7 Data security and confidentiality

### 7.1 Formal information/data security standards

AiiDA adopts a distributed approach whereby an AiiDA instance (the code plus the associated database) can be hosted on an individual’s machine, a group server or a national or international server. Instances within a group should be managed and secured by the group itself or an appointed administrator. We provide a means of sharing results either with collaborators or the public at large via the materialscloud.org website which runs a full AiiDA instance. In this case, EPFL and CSCS will be responsible for maintaining data security.

Data transport and access will be carried out over secure communication channels, i.e. Secure Shell (SSH), and access to the database will be restricted to authorised users only.

The AiiDA database does not store users' private **SSH** keys and therefore any possible compromise of the database does not lead to a security breach that extends beyond the data stored in the database itself.

## 7.2 Main risks to data security

Access to data and the execution of simulations is typically initiated by opening an **SSH** connection. This protocol itself is considered to be highly secure and is widely used. Also, **SSH** keys are used to connect rather than passwords. Moreover, these keys are not stored in the AiiDA database; instead, AiiDA uses the keys of the Linux user under which the AiiDA daemon is running, therefore there is no additional security risk with respect to standard **SSH** connections. In any case, should a private key be obtained by an individual other than the authorised user there are typically system logs that keep track of all accesses, and the specific **SSH** key can be disabled to stop further activity.

## 8 Data sharing and access

### 8.1 Suitability for sharing

The data generated in this project is highly suitable for sharing. Given that a simulation may take many hundreds (if not thousands) of **Central Processing Unit (CPU)** hours it is beneficial to the community to be able to access these without having to recompute them. Raw data produced by AiiDA can be shared on the **Materials Cloud Explore** section, which uses directly AiiDA as a backend and allows users to easily browse the full data provenance. In addition to that, for high profile research projects there will be an entry on the **Materials Cloud Discover** section, that will contain curated data and results condensed from many simulations in a form that gives an overview of a particular property or area of interest.

AiiDA provides a social ecosystem where the simulation results, materials and provenance data can be shared. It provides plugins to import crystal structures from many common formats and directly from external databases such as **ICSD** [16] or **COD**. It has also **COD** and **TCOD** exporters to export data to these external databases. Moreover it is fully interoperable with commonly used data formats for crystal structures such as **XSF** [23], **ASE** [18] and **Pymatgen** [19].

### 8.2 Discovery by potential users of the research data

Data will be discoverable by the following means:

- The **materialscloud.org** website will host a public facing frontend enabling access to publicly shared results.
- A persistent **Digital Object Identifier (DOI)** is assigned to all data entries published on the **Materials Cloud Archive** such that data entries can be cited in scientific articles.
- Publications will contain references to the database (including **UUIDs**) indicating where the data used for that study can be found.

- Publications that use results from the AiiDA repository will be encouraged to cite the paper describing the software infrastructure.

## 8.3 Data access and reusability

### 8.3.1 REST interface and OPTiMaDe support

A key feature to enable large-scale data access is represented by the development of the AiiDA [REST API](#), an interface that assigns Universal Resource Identifiers (URIs) to the objects stored in an AiiDA instance, thus making them obtainable via HTTP requests in a programmatic manner. The [REST API](#) can be coupled to authentication/authorisation modules so that the AiiDA administrator can choose the degree of accessibility of the resources. Moreover the AiiDA developers signed-on to provide a URI syntax compliant with the standards defined in the [OPTiMaDe \[22\]](#) protocol to let the users interrogate heterogeneous databases with the same syntax. Finally, the AiiDA [REST API](#) can be used as an essential building block for on-line services that expose a repository of data persisted in AiiDA. A paramount example is the [Materials Cloud](#) whose back-end largely relies on the AiiDA [REST API](#).

### 8.3.2 Sharing of plugins workflows

Point A1.1 of the [FAIR](#) principles requires that “the protocol is open, free, and universally implementable”. This is facilitated with a flexible plugins system which is part of AiiDA and enables developers to extend AiiDA to support their own codes and formats and share this functionality with the community via the Materials Cloud. Similarly, workflows that encode scientific expertise on how to carry out a series of steps to arrive at a result can be disseminated via the Materials Cloud. The [AiiDA plugin registry](#) currently contains 62 code executables from 25 simulation packages, and 50+ workflows (most of which are output of MAX phase 1 [24]), which leverage the AiiDA workflow engine that is part of the “aiida\_core” package.

## 8.4 Governance of access

### 8.4.1 Governance of data access

In MAX, we believe that sharing research data in an open format is crucial to guarantee reproducibility, increase transparency and impact of research and accelerate discovery. To this aim, data needed to reproduce a given scientific publication should be published in a curated format in an open repository, together with the corresponding scientific paper. Furthermore, to guarantee reuse, sharing of data should follow the principles of Open Science and ensure the four [FAIR](#) pillars of Findability, Accessibility, Interoperability and Reusability. For this reason, except when data is bound to confidentiality by legal, ethical or copyright reasons, MAX researchers will deposit data needed to reproduce a scientific paper published within the scope of the MAX CoE on the [Materials Cloud Archive](#) (and, when appropriate, also on the [Explore](#) and [Discover](#) sections of the Materials Cloud), with open licenses (e.g., with one of the Creative Commons variants) guaranteeing that everybody can find, access, and reuse the data without restrictions. 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.

#### 8.4.2 Governance of code access

The lighthouse codes of MAX CoE are released under open-source licenses (MIT or GNU-GPL, see table 1). The core of the AiiDA API (“aiida\_core”) is released under an open-source MIT license and is available to download for free. The full code, including the full content of “aiida\_core”, plus a set of useful additional plugins, are available on GitHub. Finally, the AiiDA plugins for all MAX lighthouse codes are released under open-source MIT licenses.

## 9 Relevant institutional, departmental or study policies on data sharing and data security

Data will be generated by different institutions, and the data policies of the respective institutions will apply. For data shared on the **Materials Cloud**, the policies of **EPFL** and **CSCS** will apply (as the data is stored at these two institutions). In particular, **EPFL** provides a combined document “**Directive concerning research integrity and good scientific practice at EPFL (LEX 3.3.2)**” [25] for all data policies. The data policies from **CSCS** are instead explained on the **User Portal of CSCS website** [26].

## Acronyms

- API** Application Programming Interface. 4, 7, 12, 16, 17
- CIF** Crystallographic Information File. 12
- COD** Crystallography Open Database. 10, 15
- CoE** Centre of Excellence. 4, 5, 9, 10, 12, 16, 17
- CPU** Central Processing Unit. 15
- CSCS** Centro Svizzero di Calcolo Scientifico - Swiss National Supercomputing Centre, Switzerland. 12–14, 17
- DAG** Directed Acyclic Graph. 4, 11, 12
- DMP** Data Management Plan. 4, 5, 14
- DOI** Digital Object Identifier. 15
- EPFL** École Polytechnique Fédérale de Lausanne, Switzerland. 12, 14, 17
- FAIR** Findable, Accessible, Interoperable and Re-usable. 13, 14, 16

**ICSD** Inorganic Crystal Structure Database. 10, 15

**JSON** JavaScript Object Notation. 10

**NCCR** National Centre of Competence in Research. 13

**OPTiMaDe** Open Databases Integration for Materials Design. 13, 16

**REST** REpresentational State Transfer. 4, 7, 16

**SNSF** Swiss National Science Foundation. 13

**SSH** Secure Shell. 14, 15

**TCOD** Theoretical Crystallography Open Database. 10, 11, 15

**UUID** Universally unique identifier. 11, 15

**XML** Extensible Markup Language. 9, 10

## References

- [1] Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. AiiDA: automated interactive infrastructure and database for computational science. *Comp. Mat. Sci.* **111**, 218 – 230 (2016).
- [2] AiiDA: Automated Interactive Infrastructure and Database for Computational Science. URL <http://www.aiida.net>.
- [3] Genovese, L. *et al.* Daubechies wavelets as a basis set for density functional pseudopotential calculations. *J. Chem. Phys.* **129**, 014109 (2008).
- [4] The BigDFT project. URL <http://www.bigdft.org>.
- [5] Blügel, S. & Bihlmayer, G. Full-potential linearized augmented planewave method. In Grotendorst, J., Blügel, S. & Marx, D. (eds.) *Computational Nanoscience: Do It Yourself! - Lecture Notes*, vol. 31, 85 (NIC Series, 2006). ISBN: 3-00-017350-1.
- [6] The Juelich FLEUR project. URL <http://www.flapw.de>.
- [7] Giannozzi, P. *et al.* QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Cond. Matt.* **21**, 395502–(2009).
- [8] Quantum ESPRESSO. URL <http://www.quantum-espresso.org>.
- [9] Soler, J. M. *et al.* The SIESTA method for ab initio order-N materials simulation. *J. Phys. Cond.Matt.* **14**, 2745 (2002).

- [10] SIESTA. URL <http://departments.icmab.es/leem/siesta/>.
- [11] Marini, A., Hogan, C., Grüning, M. & Varsano, D. yambo: An ab initio tool for excited state calculations. *Computer Physics Communications* **180**, 1392 – 1403 (2009).
- [12] The Yambo project. URL <http://www.yambo-code.eu>.
- [13] Gražulis, S. *et al.* Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research* **40**, D420–D427 (2012).
- [14] TCOD: Theoretical Crystallography Open Database. URL <http://www.crystallography.net/tcod/>.
- [15] Merkys, A. *et al.* A posteriori metadata from automated provenance tracking: integration of aiida and tcod. *JOURNAL OF CHEMINFORMATICS* **9** (2017).
- [16] ICSD: Inorganic Crystal Structure Database. URL <http://www.fiz-karlsruhe.com/icsd.html>.
- [17] PostgreSQL: The world's most advanced open source database. URL <http://www.postgresql.org/>.
- [18] ASE file format. URL <https://wiki.fysik.dtu.dk/ase/>.
- [19] PYMATGEN file format. URL <http://pymatgen.org/>.
- [20] Docker documentation. URL <https://docs.docker.com/engine/reference/builder/>.
- [21] MARVEL NCCR Project. URL <http://nccr-marvel.ch>.
- [22] OPTiMaDe: Open Databases Integration for Materials Design. URL <http://www.optimade.org/>.
- [23] XSF file format. URL <http://www.xcrysden.org/doc/XSF.html>.
- [24] Huber, S. *et al.* Release of a web open repository of data, workflows and turnkey solutions. Deliverable D3.5 of the H2020 CoE MaX. EC grant agreement no: 676598, EPFL, Lausanne, Switzerland (2017).
- [25] Directive concerning research integrity and good scientific practice at EPFL (LEX 3.3.2). URL [http://research-office.epfl.ch/files/content/sites/research-office/files/Research%20Ethics/3.3.2\\_principe\\_integrite\\_recherche\\_an%2811%29.pdf](http://research-office.epfl.ch/files/content/sites/research-office/files/Research%20Ethics/3.3.2_principe_integrite_recherche_an%2811%29.pdf).
- [26] CSCS Data storage policy. URL [http://www.cscs.ch/fileadmin/user\\_upload/customers/CSCS\\_Application\\_Data/Files/Data\\_Storage\\_policy\\_V2.pdf](http://www.cscs.ch/fileadmin/user_upload/customers/CSCS_Application_Data/Files/Data_Storage_policy_V2.pdf).