D7.2

Data Management Plan

Nicola Marzari

Due date of deliverable	30/06/2023 (**month 06**)
Actual submission date	30/06/2023
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	http://www.max-centre.eu
Deliverable no.	D7.2

Authors	Nicola Marzari, Fabio Affinito, Andrea Ferretti, Luisa Neri
To be cited as	Nicola Marzari (2023): Data Management Plan (DMP) for MAX Centre of Excellence. Deliverable D7.2 of the HORIZON-EUROHPC-JU-2021-COE-01 project MAX (final version as of 30/06/2023). EC grant agreement no: 101093374, UBremen, Germany.

Disclaimer

Contents

1 Executive Summary **4**

2 Introduction **5**
2.1 About this document 5

3 Description of the data **5**
3.1 Types of data 6
3.2 Format and scale of the data 6

4 Data collection/generation **7**
4.1 Methodologies for data collection/generation 7
4.2 Data quality and standards 7

5 Data management, documentation and curation **8**
5.1 Managing, storing and curating data 8
5.2 Metadata standards and data documentation 8
5.3 Data preservation strategy and standards 9
5.4 Data sovereignty: Materials Cloud Archive at CINECA 9

6 Templates for data management plans for projects using the Materials Cloud **10**

7 Data security and confidentiality **10**
7.1 Formal information/data security standards 10
7.2 Main risks to data security 11

8 Data sharing and access **11**
8.1 Suitability for sharing 11
8.2 Discovery by potential users of the research data 11
8.3 Data access and reusability 12
8.3.1 REST interface and OPTiMaDe support 12
8.3.2 Sharing of plugins workflows 12
8.4 Governance of access 12
8.4.1 Governance of data access 12
8.4.2 Governance of code access 13

9 Relevant institutional, departmental or study policies on data sharing and data security **13**

Acronyms **13**

References **14**

1 Executive Summary

The MAX Centre of Excellence (CoE) aims at supporting the needs of all the stakeholders involved in the field of materials modelling, simulation and design by providing new instruments and services in the form of data, codes, expertise and turnkey solutions to efficiently address the crucial challenges of novel materials development in the exascale computing era. This document provides a description of the strategies and solutions adopted within the MAX CoE to establish a high level materials' informatics framework to curate, preserve and share all the data produced by the flagship codes.

The core technology behind this objective is the AiiDA code, a python infrastructure designed to support different codes through a plugin interface, allow for an automated design and implementation of complex workflows and task tracking and able to store the full provenance of each object in a tailored database. AiiDA parses the input and output files and runs the calculations on high performance computing platforms, stores the data using uniform formats based on python dictionaries and preserve the full provenance in the form of a Directed Acyclic Graph (DAG).

AiiDA also enables a social ecosystem where the simulation workflows and results can be openly shared, on one hand with the update of the AiiDA plugin and workflow systems and on the other, with the development of the AiiDA REpresentational State Transfer (REST) Application Programming Interface (API) which also constitutes the backbone of the Materials Cloud portal and finally, with the implementation of various exporters and converters to the most commonly used data formats and ontologies. Long-term sharing and preservation is supported thanks to the Materials Cloud Archive open repository, guaranteeing storage for at least 10 years after publication. The Archive is integrated with the Materials Cloud Explore section, that guarantees a Findable, Accessible, Interoperable and Re-usable (FAIR)-compliant sharing of data produced by AiiDA, and with the Materials Cloud Discover section, for sharing highly curated data.

Statement on Open Research Data: In MAX , we believe that sharing research data in a FAIR format is crucial to guarantee reproducibility, increase transparency and impact of research and accelerate discovery. 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 with open licenses, guaranteeing that everybody can find, access and reuse the data without restrictions. For more details, see section 8.4.1.

Support to researchers writing Data Management Plans (DMPs): We provide DMP templates to researchers who use AiiDA and/or share their data on the Materials Cloud Archive. The templates are described in more detail in section 6.

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

2 Introduction

2.1 About this document

This document is deliverable D7.2 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.

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 simulations code and the physics of the problem under study. The design of each calculation node is documented by the plugin developers.[1] 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.

[1] As an example see https://aiida-quantumespresso.readthedocs.io/en/latest/index.html

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. [2]

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].

[2] See https://aiidateam.github.io/aiida-registry/

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. [3] To ease the deployment procedure a dockerfile [20] for AiiDA installation with default parameters has been realised. [4] 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. 1), 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.

[3] See http://aiida-core.readthedocs.org/en/latest/
[4] See https://hub.docker.com/r/aiidateam/aiida_core_base/

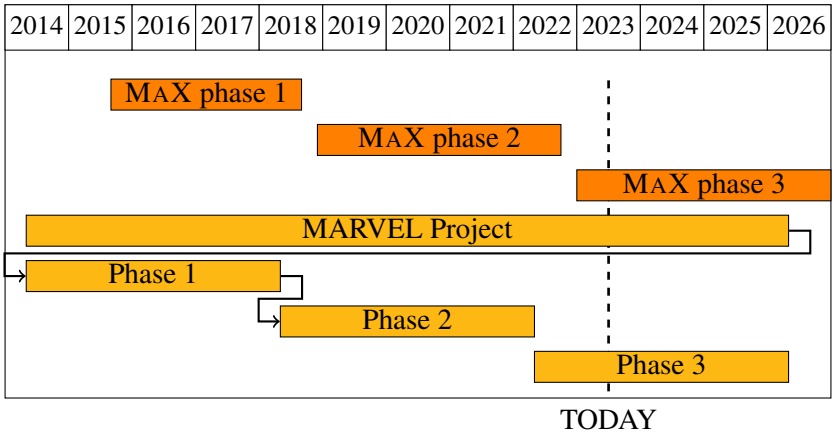

Figure 1: MAX project timeline, compared with the Swiss NCCR MARVEL project. The current project is indicated as MAX phase 3.

Taken together, the Materials Cloud and AiiDA aim at enabling researchers to trivially adopt aspects of the 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 Materials 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+ worflows (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, 9, 12, 13

CIF Crystallographic Information File. 8

COD Crystallography Open Database. 6, 11

CoE Centre of Excellence. 4, 5, 7, 8, 13

CPU Central Processing Unit. 11

CSCS Centro Svizzero di Calcolo Scientifico - Swiss National Supercomputing Centre, Switzerland. 8, 9, 11, 13

DAG Directed Acyclic Graph. 4, 7, 8

DMP Data Management Plan. 4, 10

DOI Digital Object Identifier. 12

EPFL École Polytechnique Fédérale de Lausanne, Switzerland. 8, 11, 13

FAIR Findable, Accessible, Interoperable and Re-usable. 4, 10, 12

ICSD Inorganic Crystal Structure Database. 6, 11

JSON JavaScript Object Notation. 6

NCCR National Centre of Competence in Research. 9

OPTiMaDe Open Databases Integration for Materials Design. 10, 12

REST REpresentational State Transfer. 4, 12

SNSF Swiss National Science Foundation. 9

SSH Secure Shell. 11

TCOD Theoretical Crystallography Open Database. 6, 7, 11

UUID Universally unique identifier. 7, 12

XML Extensible Markup Language. 6

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.

[26] CSCS Data storage policy. URL http://www.cscs.ch/fileadmin/user_upload/customers/CSCS_Application_Data/Files/Data_Storage_policy_V2.pdf.