D4.2

First Data Management Plan

Nicola Marzari, Daniele Tomerini, Andrea Padovani,

Pablo Ordejón, and Arrigo Calzolari

Document information

Project acronym:	INTERSECT
Project full title:	Interoperable Material-to-Device simulation box for disruptive electronics
Research Action Project type:	Accelerating the uptake of materials modelling software (IA)
EC Grant agreement no.:	814487
Project starting / end date:	1st January 2019 (M1) / 31st January 2022 (M37)
Website:	www.intersect-project.eu
Final version:	30/07/2019
Deliverable No.:	D4.2
Responsible participant:	CNR (participant number 1)
Contributing Consortium members:	CNR, EPFL, ICN2, IMEC, FMC, AMAT
Due date of deliverable:	31/07/2019
Actual submission date:	31/07/2019
Dissemination level:	PU – Public

Authors:	Nicola Marzari, Daniele Tomerini, Andrea Padovani, Pablo Ordejón, and Arrigo Calzolari
To be cited as:	N. Marzari, D. Tomerini, A. Padovani, P. Ordejón, and A. Calzolari: First Data Management Plan. Deliverable D4.2 of the H2020 project INTERSECT (final version as of 17/07/2019). EC grant agreement no: 814487, CNR, Modena, Italy

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.

D4.2 First Data Management Plan

Content

1 Executive Summary

INTERSECT aims at supporting the needs of all the stakeholders involved in the field of materials and device modeling, simulation and design by providing new instruments and services in the form of data, codes, expertise and interoperable solutions to efficiently address the crucial challenges of development of disruptive electronics, such as synaptic devices for neuromorphic computing. This document provides a description of the strategies and solutions adopted within the INTERSECT project to establish a high level materials' informatics framework to curate, preserve and share all the data produced by the Interoperable Materials-To-Device (**IM2D**) simulation box. The core technology behind this objective is the *simulation hub* which provides modelling service and stems from the interconnections of original engine codes, namely Quantum ESPRESSO (QE), SIESTA^{TM} and Ginestra^{TM}. QE and SIESTA^{TM} are software tools for electronics and atomistic models, based on DFT, for quantum mechanical materials modelling. Ginestra^{TM} is a commercial software for atomistic and continuum model, specifically oriented to advanced device simulation. IM2D conjugates the advantages of both material- and device-driven software solutions connecting the properties of materials at the atomistic level to the electrical performance of the devices. The code interconnection and management of data exchange and storage is controlled by the *interoperability hub* which is based on the AiiDA and SimPhoNy codes. AiiDA is a Python infrastructure designed to support different codes through a plugin interface that allows 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. SimPhoNy is a framework for the integration of a suite of tools. It includes the SimPhoNy-common which is a simulation platform providing seamless integration of disparate simulation, pre/post processing software tools, advanced workflow management and data curation.

2 Introduction

2.1 About this document

This document is deliverable D4.2 of the INTERSECT project [1] 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. 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

One goal of the INTERSECT project is setup and implementation of the IM2D modelling infrastructure, which results from the upscaling and the interoperable integration of existing codes for materials and device modelling: QE [2], SIESTA^{TM} [3] and Ginestra^{TM} [4]. The three codes will be at the core of the *simulation hub* of IM2D, which is conceived as an integrated modeller for the simulation of complex materials (hereafter materials cycle) and electron devices (hereafter device cycle). QE and SIESTA^{TM} are open-source computer codes for electronic-structure calculations and materials modeling, while Ginestra^{TM} is a commercial, proprietary software designed to simulate the electrical response of a selected synaptic devices (e.g. FeFET, selectors, PCM, RRAM, etc). Materials and device cycles generate, respectively, materials and device data type. Interoperable connection among modelling engines is provided by AiiDA [5] and SimPhoNy [6] infrastructures.

The materials informatics framework AiiDA has been designed to support different codes, and plugins are already available or are being currently developed to support all codes involved in the INTERSECT project, irrespective of the cycle (materials vs device) that generated or read the data. AiiDA promotes advanced programming models through 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 will be used to organise and coordinate materials and device simulations, searching for optimal properties and performance by acquiring a variety of heterogeneous microscopic data from the calculations. It should ideally allow for an automated design and implementation of complex workflows and task tracking, based on a scripting interface for job creation and submission. This will implement, for instance, the linked model workflows (namely materials-to-device and device-to-materials) within INTERSECT, where the electronic model is treated by QE and SIESTA^{TM} codes, while the atomic and continuum ones by Ginestra^{TM}. The results will feed a database of structures and properties that will in turn drive further simulations. The data thus generated may be used for instance for data-mining and machine-learning, or to build classical neural networks to further ramp up the time and length scales accessible to numerical modeling.

Code Name	License	Main Developers
Quantum ESPRESSO [2]	GNU-GPL	CNR, EPFL, SISSA
SIESTA^{TM} [3]	Siesta License	ICN2, BSC
Ginestra^{TM} [4]	Commercial	Applied Materials Italia (formerly MDLab)
AiiDA [5]	MIT	EPFL, Robert Bosch
SimPhoNy [6]	GNU-GPL	FRA

Table 1: The primary codes that will be used as part of IM2D infrastructure.

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 for from text files with the appropriate dictionaries) in a database and its associated file repository. In the case of materials cycle, this data is generated from open-source electronic-structure material simulations codes that encompass key technologies such as pseudopotentials, and localised basis sets, (time dependent) density-functional theory, multiscale/multiphysics modelling with a focus on structural and electronic properties, and thermal/electrical transport of complex defective systems. We are also using the Crystallography Open Database (COD) [7] and Theoretical Crystallography Open Database (TCOD) [8] external open access databases of crystal structures for organic and inorganic compounds to obtain the input atomic coordinates of crystalline materials. In the case of device cycle, this data is generated from Ginestra^{TM} code that encompasses technologies such as kinetic Monte Carlo for defects and ion diffusion/recombination at finite temperature as well as multiscale/multiphysics modelling (such asd drift-diffusion, Fowler-Nordheim (FNT) and direct tunnelling (DT), and the defect-assisted charge transport) for the description of carrier transport across the active layers.

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 used a uniform format to define the main raw and analysed data irrespective of the different plugins (e.g. QE or SIESTA^{TM}). These formats contain data in dictionary format, exportable for instance to plain JavaScript Object Notation (JSON): for example, StructureData, ParameterData data types in AiiDA store metadata in python dictionaries within a database.

Currently we are using the PostgreSQL [9] open-source relational database to store our data. 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. The SimPhoNy code provides also interoperable solutions for data visualization.

AiiDA provides a social ecosystem where the simulation results, materials and provenance data and scientific workflows can be shared. It provides plugins to import crystal structures from many common formats and directly from external databases such as the Inorganic

Crystal Structure Database (ICSD) [10] or the COD. It has also COD and TCOD exporters to export data to these external databases. This allows us to share data easily and also ensures their long-term preservation.

4 Data collection/generation

4.1 Methodologies for data collection/generation

Data for the project will be created and collected by using the AiiDA framework for the management of the simulations. AiiDA plugins and workflows are being written for different simulation codes in order to support the calculations with at least the codes used within this project, but also to support other codes available in the community (e.g. VASP, CP2K). 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 Directed Acyclic Graph (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 materials and device simulations codes will be stored directly so that they may be re-parsed or manually inspected if necessary. Otherwise data will be stored as standard, code-independent objects in the AiiDA framework (e.g. crystal structures, band structures, pseudopotentials, k-point paths, etc) allowing 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 will be used in combination 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 uptake of involved codes to a semantic interoperability level that does not restrict to the coupling and linking of models and the generation of a data pipeline between existing codes, but it requires the description of the information meaning in a formal and machine-readable and processable way (metadata and schema based on semantics, i.e., meaning). To this end INTERSECT adopted a shared ontology (including core vocabularies and standards) to introduce a formal and explicit specification of shared concepts (e.g. material entities, models, materials relations) as well as to represent knowledge as a set of concepts related by hierarchical relations. The implementation is based on SimPhoNy-common that provides implementation of a proper ontology, compliant with the European Materials Modelling Ontology (EMMO), into a specific data structure, Common Universal Unified Data Structures (CUDS), and a common Application Programming Interface (API). In particular, SimPhoNy CUDS provide means to store all aspects of a workflow using a hierarchical data structure based on ontologies. In INTERSECT the CUDS are upscaled to support the entire workflow relevant to synaptic electronics, including device modelling. CUDS provide a simplistic API for data manipulation and management as well as a Hierarchical Data Format (HDF5) based data structure, namely h5CUDS. The CUDS API and h5CUDS provide IM2D with an EMMO based semantic layer both to the outside applications and internally for communicating between different components.

5 Data management, documentation and curation

5.1 Managing, storing and curating data

All data (calculations, their inputs and their outputs) generated by running both materials and device cycle simulations on local or remote servers is 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 (e.g. crystal structures, band structures, density of states, etc.) is the same for all objects of the same type, even if generated by different computer codes (e.g. QE or SIESTA^{TM}), 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 instance Crystallographic Information File (CIF) files for crystal structures). Further export functions can be added transparently. 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 (user discretional) to export the data to public repositories.

Currently the full materials database and part of the file repository (small files) are stored on a server at École Polytechnique Fédérale de Lausanne, Switzerland (EPFL) while the remaining part of the 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. On the EPFL server, there are backup scripts running every day performing a full backup of database and an incremental backup of the file repository. The data stored at CSCS server is also backed up daily.

Device data, will be stored in the Ginestra™ workspace using the hierarchical data format HDF5. Saved data can be retrieved anytime through the GUI and exported in proprietary file formats that can be imported in other Ginestra™ installations. This enables an easy exchange of the results between the project partners using Ginestra™. Ginestra™ material database is stored using XML format and can also be exported in a proprietary file format.

5.2 Metadata standards and data documentation

There are several key pieces of simulation software that will be used for this project as described in Table 1. Typically a simulation is run by supplying one or more input files that, along with the primary data of interest (be it a configuration of atoms in space, the electronic structure, a material property, electrical test specifications, etc.) will 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 (e.g. QE, SIESTA^{TM}, Ginestra^{TM}), 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, SimPhoNy and the single engine codes themselves can be considered to be data in this context. For example, 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 at http://aiida-core.readthedocs.org/en/latest/. Further documentation can be found on the web pages of single codes [2-6].

5.3 Data preservation strategy and standards

As part of the MARVEL National Centre of Competence in Research (NCCR) project [11], funded by the Swiss National Science Foundation (SNSF), a large storage allocation is being purchased at CSCS which will be used to store and retain large files until at least 2022. Later, data will be handled until 2026. In the meantime, we expect to obtain other funding opportunities for further future preservation of the data. This applies also to all the data on the **materialscloud.org** platform that is being developed at EPFL, where consortium members will be free to share data publicly and privately (with selected collaborators) under the condition that data are made publicly available within approximately a year from depositing.

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 be easily changed by the user or group who runs the simulations and generates the data.

6 Data security and confidentiality

6.1 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, able to run a full AiiDA instance. In this case, EPFL and CSCS (for the AiiDA repository and the large files, respectively) 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.

6.2 Main risks to data security

Access to data and the execution of simulations usually is 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 the password. 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 system logs that keep track of all access, and the specific SSH key can be disabled to stop unauthorized activity.

In addition, AiiDA keeps an extensive log of the activities carried out which can be examined retrospectively if necessary.

7 Data sharing and access

7.1 Suitability for sharing

The data generated in this project are 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.
In addition to raw data there will be a curated section of **materialscloud.org** that will contain results condensed from many simulations in a form that gives an overview of particular properties or areas of interest.

7.2 Discovery by potential users of the research data

Data will be discoverable by way of the following means:
- The **materialscloud.org** website which will host a public facing frontend enabling access to publicly shared results.
- A private section of the website will allow dissemination with selected (authorised) collaborators.
- Publications will contain references to the database (including UUIDs) indicating where the data used for that studies are located and can be found.
- Publications that use results from the AiiDA repository will be encouraged to cite the paper describing the software infrastructure.

7.3 Governance of access

The ultimate decision about sharing the data will lie with the PI and the authors of the data, however in general it is expected that the research data produced within this project will be made *findable, accessible, interoperable and reusable* (FAIR). This, however, does not imply any obligation by project partners or other external users to share their own data or make them public beyond this project.
We are in the process of preparing a data sharing policy for the data added to the **materialscloud.org** web portal, that will require users to make their data available under a Creative Commons license after a fixed term, which is likely to be one year.
Notably, this holds only if the users willingly upload their AiiDA databases to the Materials Cloud - it is expected that for industrial users and consultants these would remain within the company firewalls. Data documentation is independent from user’s distribution policies: both open and confidential data will be produced, traced and curated with the same standards. Furthermore, INTERSECT shall use an internal registry of all data exchanged between partners or produced, especially on confidential data and either open or proprietary data with potential for exploitation.

The core of the AiiDA APIs (“aiida_core”) is released under an open-source MIT license and is available to download for free (Tab. 1). SimPhoNy, QE and SIESTA^{TM} codes are open source, distributed with GNU-GPL or SIESTA-licence. Ginestra$^{\text{TM}}$ is a proprietary code distributed under commercial licence.

8 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 **materialscloud.org**, the policies of EPFL and CSCS will apply (as the data are expected to be 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)"*[12] for all data policies. The data handling from CSCS is described on their website [13].

Acronyms

API Application Programming Interface. 9, 11
CIF Crystallographic Information File. 9
COD Crystallography Open Database. 6, 7
CPU Central Processing Unit. 10
CSCS Centro Svizzero di Calcolo Scientifico - Swiss National Supercomputing Centre, Switzerland. 8–11
CUDS Common Universal Unified Data Structures. 9
DAG Directed Acyclic Graph. 7, 8
EPFL École Polytechnique Fédérale de Lausanne, Switzerland. 8–11
EMMO European Materials Modelling Ontology. 8, 9
HDF5 Hierarchical Data Format. 9, 10
ICSD Inorganic Crystal Structure Database. 7
IM2D Interoperable Materials-To-Device. 4, 5, 9
JSON JavaScript Object Notation. 7
NCCR National Centre of Competence in Research. 9
QE Quantum ESPRESSO. 4, 5, 7, 9, 10
SNSF Swiss National Science Foundation. 9
SSH Secure Shell. 10
TCOD Theoretical Crystallography Open Database. 6–8
UUID Universally unique identifier. 7, 11
WP Work Package. 4
XML Extensible Markup Language. 6, 7, 10

References

[1] INTERSECT project www.intersect-project.eu

[2] 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). Quantum ESPRESSO URL: www.quantum-espresso.org

[3] Soler J. M., et al. The SIESTA method for ab initio order-N materials simulation. J. Phys. Cond.Matt. 14, 2745 (2002). SIESTA^{TM} URL http://departments.icmab.es/leem/siesta

[4] Larcher L., et al. A simulation framework for modeling charge transport and degradation in high-k stacks, J. Comput. Electron.12, 658–665 (2013). Ginestra^{TM} URL: www.mdlsoft.com

[5] Pizzi G., Cepellotti A., Sabatini R., Marzari N., and Kozinsky B. AiiDA: automated interactive infrastructure and database for computational science. Comp. Mat. Sci. 111, 218 – 230 (2016). AiiDA URL: http://www.aiida.net

[6] Adler J., et al. Visualization in the integrated SimPhoNy multiscale simulation framework, Comp. Phys. Comm. 231, 45-61 (2018). SimPhoNy URL: www.simphony-project.eu

[7] 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).

[8] TCOD: Theoretical Crystallography Open Database. URL www.crystallography.net/tcod

[9] PostgreSQL: The world's most advanced open source database. URL www.postgresql.org

[10] ICSD: Inorganic Crystal Structure Database. URL www.fiz-karlsruhe.com/icsd.html

[11] MARVEL NCCR Project. URL http://nccr-marvel.ch

[12] Directive concerning research integrity and good scientific practice at EPFL (LEX 3.3.2). URL http://polylex.epfl.ch/files/content/sites/polylex/files/recueil_pdf/ENG/3.3.2_principe_integrite_recherche_an.pdf

[13] CSCS Data storage policy. URL https://user.cscs.ch/storage/file_systems/