



# D4.6

## Data Management plan assessment and revision

Francesco Aquilante, Flaviano José Dos Santos, Matthias Büschelberger,  
Arrigo Calzolari, and Nicola Marzari

## 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:	1 <sup>st</sup> January 2019 (M1) / 31 <sup>st</sup> January 2022 (M37)
Website:	<a href="http://www.intersect-project.eu">www.intersect-project.eu</a>
Final version:	29/01/2021
Deliverable No.:	D4.6
Responsible participant:	CNR (participant number 1)
Contributing Consortium members:	EPFL, CNR, FRA
Due date of deliverable:	31/01/2021
Actual submission date:	29/01/2021
Dissemination level:	PU - Public

**Authors:** Francesco Aquilante, Flaviano José Dos Santos, Matthias Büschelberger, Arrigo Calzolari, and Nicola Marzari

**To be cited as:** F. Aquilante, et al. (2021): Data Management Plan Assessment and Revision D4.6 of the H2020 project INTERSECT (final version as of 29/01/2021). 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.

### Versioning and Contribution History

Version	Date	Modified by	Modification reason
v.01	05/01/2021	Francesco Aquilante	First Version
v.02	18/01/2021	Arrigo Calzolari	Revision
v.03	25/01/2021	Matthias Büschelberger	Add contribution of SimPhoNy
V.04	27/01/2021	Arrigo Calzolari	Final Version

## D4.6 Data Management plan assessment and revision

### Table of Contents

<b>Executive Summary</b>	<b>5</b>
<b>1. Introduction</b>	<b>6</b>
1.1 About this document	6
1.2 First data management plan	6
<b>2. Integration to external databases</b>	<b>7</b>
2.1 AiiDA-OPTIMADE REST API	8
<b>3. Update due to the contribution of SimPhoNy to IM2D</b>	<b>11</b>
3.1 Data format, data source and data type	12
3.2 Data managing and data storing	12
3.3 Data sharing and access	13
<b>4. Security measures for data and code resources</b>	<b>13</b>
<b>Conclusion</b>	<b>14</b>
<b>References</b>	<b>15</b>
<b>ACRONYMS</b>	<b>16</b>



## Executive Summary

INTERSECT [1] is conceived to provide new instruments and services to the community of materials and device modeling in the form of data, codes, expertise and interoperable solutions for the development of disruptive electronics. To this end, INTERSECT aims at establishing a high-level informatics infrastructure to manage the data produced by the *Interoperable Materials-To-Device (IM2D)* simulation box through the interconnection of three main engine codes: *Quantum ESPRESSO (QE)* [2] and SIESTA™ [3], both being software tools for quantum mechanical modeling of materials, and Ginestra™ [4], for atomistic and continuum modeling of electronic devices. IM2D combines these tools to achieve a predictive level of accuracy for the performance of the actual device from first-principles quantum mechanical simulations. The AiiDA [5] and SimPhoNy [6] softwares are at the core of such *interoperability Hub (iHub)*. The former is a Python infrastructure supporting different codes through plugins for automated design and implementation of complex workflows and task tracking, while capable to store the full provenance of each object in a tailored database. The latter is a Python-framework offering semantic interoperability to third party tools, such as simulation-, data-storage- and data-transformation-backends. Ontology domains in common *Resource Description Framework (RDF)*-formats can be installed in the core component (*osp-core*), so that their entities can be instantiated on a script-basis, interconnected among each other and deliver information to syntactic data structures through *osp*-wrappers. The connection to public repositories and to a repository of repositories (a catalogue) is handled by a gateway to the *Application Programming Interface (API)* known as OPTIMADE [7], briefly described in the present deliverable.

Within this multilevel computational framework, data and data pipeline are of paramount relevance. Most relevant definitions, formats and data schema implemented within the INTERSECT project have been described in the First Data Management Plan (FDMP), Deliverable D4.2, submitted at M7 (July 2019). The present deliverable focuses on updates and new implementations not included in the first report. In the following, we first describe the aspects related to the access of data from public repositories by means of the OPTIMADE API gateway-client through AiiDA. Furthermore, the details of interfacing SimPhoNy and Ginestra™ have been worked out more in detail in order to semantically enrich the IM2D toolbox. By differentiating the workflow parameters into multiple degrees of difficulty, the intersection will also provide features for upscaling the Ginestra™- Graphical User Interface (GUI) in terms of higher flexibility and usability for operators with varying scientific background. Finally, we briefly describe the recent actions adopted to assure the security of data and code sources in the development of the IM2D code.

## 1. Introduction

### 1.1 About this document

This document represents the deliverable D4.6 of the INTERSECT project, and it is part of the activity of WP4, under Task 4.1. The content of this document is intended to complement D4.2, where data management within the INTERSECT project was described. No relevant Task deviations from the original DoA plan are to be mentioned.

### 1.2 First data management plan

The FDMP describes the format and the scale of the data generated/used by the IM2D and its interconnection with external data repositories, the methodologies for data collections, the data quality and standards of curated metadata, and data preservation strategies. Particular attention has been dedicated to data security and confidentiality as well as to data sharing and access. Results of IM2D may be shared either with collaborators or the public at large via **materialscloud** [8], an Open Science Platform managed by EPFL and designed to enable seamless sharing of resources in computational materials science. Confidentiality solutions have also been planned, in order to allow industrial users and consultants to keep their sensitive data privately stored within their company firewalls.

These definitions and data organization have not changed since publication of FDMP and are still the basis of the project data management. We summarize in the following the main aspects of the original plan:

#### Data type, format and standards

- Materials and device cycles (simulation hub) generate, respectively, **materials and device data type**. Input crystalline coordinates from *Crystallography Open Database* (**COD**) and *Theoretical Crystallography Open Database* (**TCOD**);
- Input and output files (**raw data**) stored as text or (**XML**) format;
- Full provenance of each data object (inputs, outputs, calculations) is automatically stored in a database (materialscloud), in a format that enables the simulation results to be fully reproduced;
- **PostgreSQL** open-source relational database is used to store our metadata with formats which contain data in dictionary format, exportable to plain *JavaScript Object Notation* (**JSON**);

- Full provenance of all calculations is preserved in the form of a Directed Acyclic Graph (**DAG**), managed by the AiiDA infrastructure, by using *Universal Unique Identifier* (**UUID**) for each node.
- Semantic upscaled data (i.e., metadata and schema based on semantics) are organized in *Common Universal Unified Data Structures* (**CUDS**) and managed by SimPhoNy, an instantiation of Python-classes representing entities of imported ontologies in an *Integrated Development Environment* (**IDE**).

### Governance of access

- Research data produced within the project will be made public, findable, accessible, interoperable and reusable (FAIR);
- **No obligation** by project partners or other external users to share their own data or make them public beyond this project;
- IM2D users have the option to **download entire AiiDA databases** for importing and reusing in their personal AiiDA instances;
- Industrial users may **decide to protect their data** for exploitation and commercialization purposes by storing data in the proprietary database section of the data hub of IM2D.

The rest of this document focuses on the new aspects not included in the FDMP.

## 2. Integration to external databases

The possibility to access (retrieve and populate) public databases for materials modeling is central to the scope of the INTERSECT project and is a part of the realization of the *data hub*, which constitutes the IM2D infrastructure. The interaction with databases and repositories is twofold: structural and electronic properties of functional materials can be retrieved as input to be used by the computational tools in both materials and device cycles (*simulation hub*), or vice versa, the output resulting from Density Functional Theory (**DFT**) simulations can populate the databases for exploitation from third parties. In both cases, the data flow to/from databases requires a well established format and schema, managed by the AiiDA infrastructure in the present case.

In particular, we considered the integration of IM2D with the OPTIMADE database and repository system. OPTIMADE (Open Databases Integration for Materials Design) is an **open database** consortium which aims (cit.) "(...) to make materials databases interoperational by developing a common (*Representational State Transfer*) **REST** API". The OPTIMADE

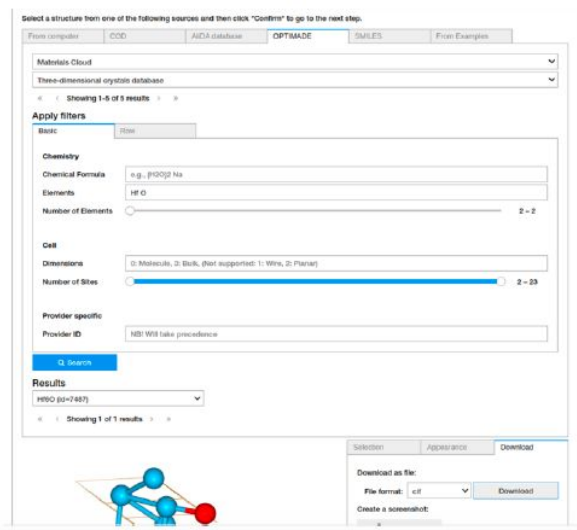
consortium collects the most relevant worldwide materials databases (e.g., AFLOW, COD, Materials Cloud, The Materials Project, NoMaD), generated by massive high throughput computer simulations. Interoperability among data and databases is pursued by means of a few semantic assets, which include among the others the *Crystallographic Information Framework* (CIF) dictionary, the Pauling vocabulary, and *European Materials Modelling Ontology* (EMMO), that have been exploited as the common semantic ground for the IM2D implementation (see WP1 activity and D1.1, D1.3, D1.4). Automatic procedures to import COD and/or CIF structures into Ginestra™, through AiiDA are discussed in detail in D2.4 (M25).

## 2.1 AiiDA-OPTIMADE REST API

The integration between IM2D and OPTIMADE has been addressed and accomplished by using the OPTIMADE API gateway, compliant with the JSON API 1.0 specification. This solution allows AiiDA to communicate with the existing databases. The actual access via AiiDA to the OPTIMADE gateway is achieved through a stand-alone tool, namely **aiida-optimade**, freely available at <https://github.com/aiidateam/aiida-optimade> and integrated within AiiDA as StructureData nodes importer, in compliance with the REST API technology. AiiDA-optimade fulfills the OPTIMADE specification about the allowed data types, the general API requirements and conventions, the response formats, the API endpoints and filters, and the entry lists. The complete description of OPTIMADE specifications can be found at [www.optimade.org/optimade](http://www.optimade.org/optimade). The aiida-optimade server is based on the test server "template" used in the [optimade-python-tools](#) package. The filter grammar and parser and [pydantic](#) models from [optimade-python-tools](#) are directly used in the present implementation.

Screenshots for its use in connection with QE are shown in Fig. 1:





```

optimade_structure_7487.in - Notesblok
Filer Rediger Formater Vis Hjælp
&CONTROL
/
&SYSTEM
ntyp      = 2
nat       = 7
ibrav     = 0
/
&ELECTRONS
/
&IONS
/
&CELL
/
ATOMIC_SPECIES
Hf 178.49 Hf_dummy.UPF
O 15.9994 O_dummy.UPF
K_POINTS gamma
CELL_PARAMETERS angstrom
2.77961590509810 1.60481199105110 5.11963394896870
0.00000000000000 1.60481199105110 5.11963394896870
0.00000000000000 0.00000000000000 5.11963394896870
ATOMIC_POSITIONS angstrom
Hf 0.0000000000 3.2158795934 1.3269293735
Hf 1.8568897994 0.8059077281 3.7927045754
Hf 4.6256706679 1.6051590743 3.7927045754
Hf 1.8562872480 3.2033683708 3.7927045754
Hf 0.9227261057 1.5989042630 1.3269293735
Hf 3.7131770474 0.0000000000 1.3269293735
O 0.0000000000 0.0000000000 0.0000000000

```

## AiiDA Lab application registry

[View on GitHub/register your app]

< Go back to the app summary



### OPTIMADE Client

Utilities

#### General information

Source code: [Go to the app source code](#)

App homepage: [Go to app homepage](#)

Documentation: Documentation not provided by the app author

#### Detailed information

Author(s): Casper Welzel Andersen

Short description: Query for and import structures from OPTIMADE providers (COD, Materials Cloud, NoMaD, Materials Project, OQMD, and more ...).

Package name (for pip): `aiidalab-aiidalab-optimade`

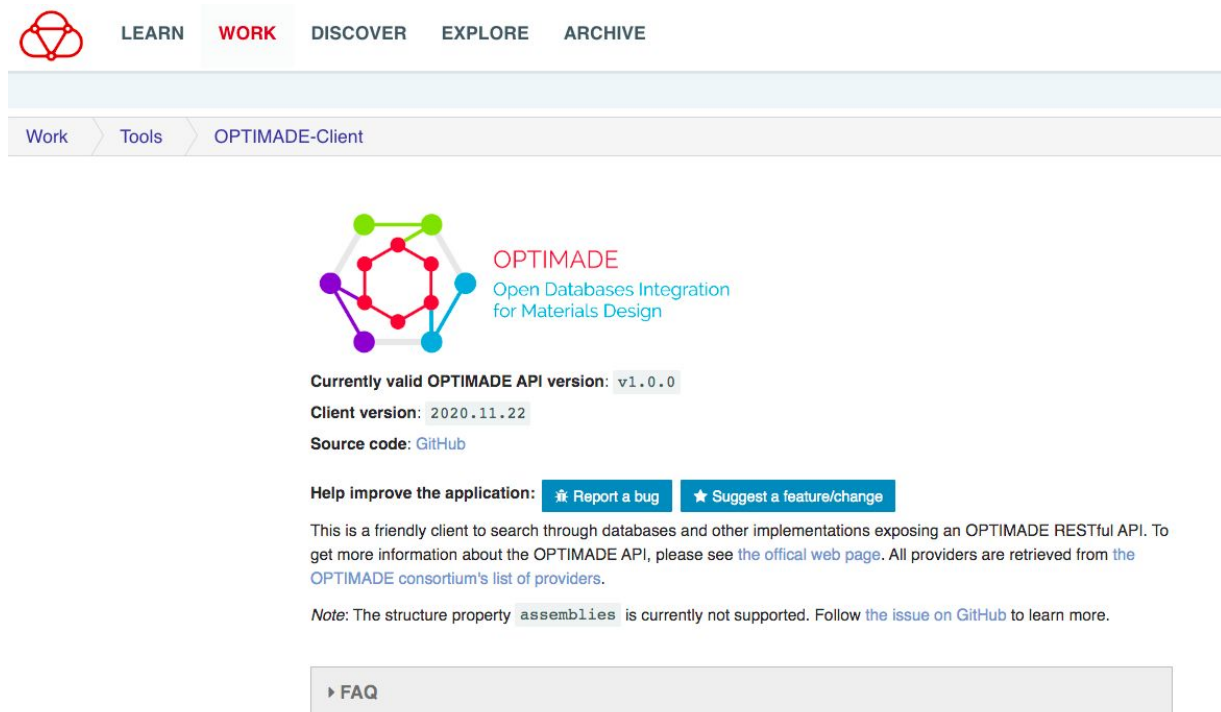
Most recent version: 1.2.3

**Figure 1** OPTIMADE connection with QE.

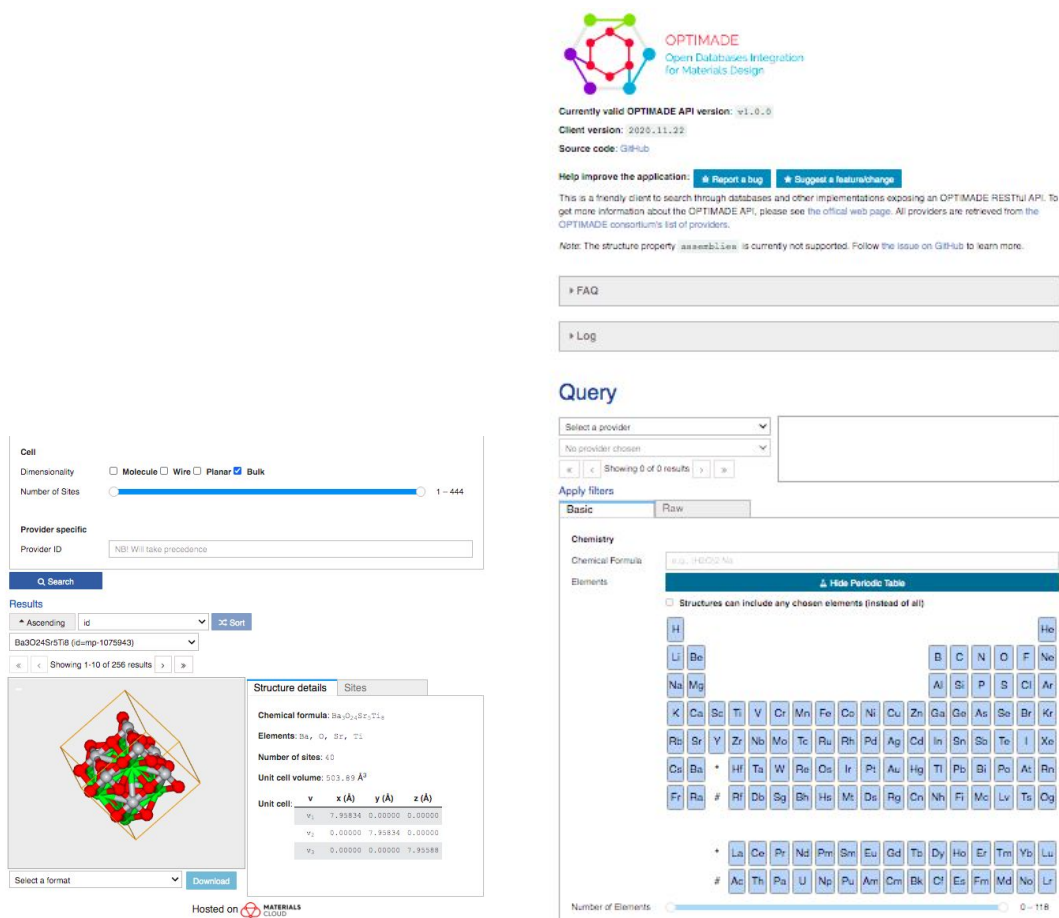
A web application of the OPTIMADE client is also available on the Materials Cloud at [materialscloud.org/optimadeclient](https://materialscloud.org/optimadeclient). The aiida-optimade client allows to browse and search materials data from several providers (e.g., materialscloud, AFLOW, open materials database, etc), and databases (e.g., standard pseudopotentials, 2d structures, 3D crystal database, etc), simply inserting the chemical formula and/or the crystallographic specification of the desired

system.

Screenshots from this client application tool are here displayed in Figs. 2 and 3:

The screenshot shows the user interface of the OPTIMADE client application. At the top, there is a navigation bar with a red logo on the left and five menu items: "LEARN", "WORK" (highlighted in red), "DISCOVER", "EXPLORE", and "ARCHIVE". Below this is a breadcrumb trail: "Work" > "Tools" > "OPTIMADE-Client". The main content area features the OPTIMADE logo, which is a colorful molecular structure, followed by the text "OPTIMADE Open Databases Integration for Materials Design". Below the logo, it displays the "Currently valid OPTIMADE API version: v1.0.0", the "Client version: 2020.11.22", and the "Source code: GitHub" link. There are two buttons: "Report a bug" and "Suggest a feature/change". A paragraph of text explains that the client is used to search through databases and other implementations exposing an OPTIMADE RESTful API. A note at the bottom states that the structure property "assemblies" is currently not supported. At the very bottom, there is a button labeled "FAQ".

**Figure 2** OPTIMADE client application tool.



OPTIMADE  
Open Databases Integration  
for Materials Design

Currently valid OPTIMADE API version: v1.0.0  
Client version: 2020.11.22  
Source code: GitHub

Help improve the application: [Report a bug](#) [Suggest a feature/change](#)

This is a friendly client to search through databases and other implementations exposing an OPTIMADE RESTful API. To get more information about the OPTIMADE API, please see the official web page. All providers are retrieved from the OPTIMADE consortium's list of providers.

Note: The structure property `assemblies` is currently not supported. Follow the issue on GitHub to learn more.

FAQ  
Log

Query

Select a provider  
No provider chosen

Showing 0 of 0 results

Apply filters

Basic Raw

Chemistry

Chemical Formula:

Elements

Structures can include any chosen elements (instead of all)

Periodic table with element selection filters.

Number of Elements: 0 - 118

Hosted on MATERIALS CLOUD

Figure 3 OPTIMADE client application tool.

### 3. Update due to the contribution of SimPhoNy to IM2D

SimPhoNy uses the existing entry-points of the AiiDA-REST-services for the DFT calculations and maps the parameters from the syntactic workflows executed by AiiDA and Ginestra™ to the ontologies developed to the guidelines of the EMMO in this project. This enhances the meaning of the quantitative parameters in the context of coherent semantic data structures which are finally representing the knowledge gained from the workflows in terms of physical quantities, physical dimensions, SI-units, user-experience levels, granularity level of the models, details of the data flow, material types, defect structures of the crystal structure, etc.

Ultimately, this has the potential to contribute to the capabilities of the semantic web which is currently supported by associated projects of the EMMC (MarketPlace, OntoTrans, etc.). The goal of the collaboration with these projects would increase the findability of the simulation tool box, usability over different types of *persona* with different scientific backgrounds, and interoperability for the most relevant tools developed within those projects [*Open Translation Environment (OTE)*, *Formulations And Computational Engineering (FORCE)*-Business Decision Support System (BDSS), etc.].

In INTERSECT, the **CUDS** are upscaled to support the entire workflow relevant to synaptic electronics, including device modelling. Furthermore, CUDS play a central role for the semantic interoperability to other simulation wrappers planned to be hosted in the framework the MarketPlace project [9].

### 3.1 Data format, data source and data type

According to the reorganization of the contribution of SimPhoNy, the data structures incoming from AiiDA and Ginestra™-computations (like mentioned in the FDMP - chapter 3) is semantically enriched without acquiring additional data, except the introduced *class hierarchy* formalized in the ontology. Basic concepts for assertion of new subclasses and individuals will be taken from the entities available in the official EMMO-GitHub DT-NMBP-09 branch , which has its origin from the version 1.0.0-beta EMMO.

The entities formalized for the data model in the ontologies will be written in a conventional RDF-format, like *Ontology Web Language (OWL)*, *Terse RDF Triple Language(TTL)* or *RDF*, since this delivers the potential to read, edit and query the conceptualization in common GUI-based applications for semantic web technologies, like e.g. Protégé.

Eventually, the formalized ontologies will also hold quantities of default values for chosen high-level workflow parameters. In correlation to the requirements for each persona from D1.1 (report on use cases), those parameters might be further formalized into differentiated accuracy levels. The details of these conceptualizations will be further discussed with the stakeholding partners in the upcoming months.

The semantic information possessed by the ontology will be interoperable with the AiiDA and Ginestra™ syntactic data structures. This means that the ontologized workflow with its quantitative properties will be transformed from CUDS into the **JSON**-format used by AiiDA, and vice versa by implementing an osp-wrapper. The interoperability between SimPhoNy and Ginestra™ is under discussion with our partners. For further information, please see D1.4 (semantic interoperability for coupling and linking).

### 3.2 Data managing and data storing

SimPhoNy provides to store the metadata of a workflow in a hierarchical data structure based on ontologies. The individuals (CUDS) created from the ontology-classes can be seamlessly stored in common third-party triplestore software for each workflow. SimPhoNy uses a dockerized version of an Allegrograph-backend available on Dockerhub [10] in order to store the semantic data in its fundamental triplets (subject, property, object) locally on the EPFL server. The Docker-Allegrograph supports a GUI-frontend, which supports *SPARQL Protocol And RDF Query Language (SPARQL)*-querying and interactive data visualisation.

The stored metadata shall represent the most important results in order to reproduce the applied techniques of the workflows. This includes, e.g., simulation in- and outputs of quantitative properties with SI-units and physical dimension, chemical formula of material, user-role and expert level, type and version of simulation software used, and representation of simulation entities from AiiDA and Ginestra™ in correlation of real-world objects.

One important concept in the semantic framework is the introduction of **expert levels**, representing the user knowledge-background, which has an individual persona-profile as identified in WP1. In parallel, SimPhoNy shall recognize and store more advanced simulation-parameters of the workflow which are actually hidden to more inexperienced users but can be made accessible and adjustable in the GUI of Ginestra™ to more experienced users and persona. This enhances the usability and reproducibility of the workflows and simulations for different user-types without losing crucial information on the data generation. More detailed information about these *knowledge levels* are treated in D1.1 (report on use cases), D1.3 (report on developed ontologies), and D1.4 (report on high level requirements).

Overall, the parallel storage of workflow-data in AiiDA PostgreSQL-backend and the Allegrograph-triplestore of SimPhoNy shall complement each other. Furthermore, this may enhance the simulation toolbox serviceability for web-technologies users and non-users of semantic, web-technologies due to the availability of syntactic and semantic data synchronously.

### 3.3 Data sharing and access

In connection to the MarketPlace project, scientific users from different industrial backgrounds in the European materials modelling community may have the potential to access data without the need to be familiar with the details of computational efforts and handling of simulations or workflows. This may give the chance to reuse generated material-data for individual production and industry demands.

Data sources and simulation tools can also be referenced, registered, and accessed in the MarketPlace by the means of ontology-driven findability for the European materials modelling community.

The details for the deployment will be further treated and discussed in D1.5 (setup and GUI).

## 4. Security measures for data and code resources

The joint on-cloud implementation of codes and workflows among partners has required to set up specific “security measures” in order to assure the protection and the confidentiality against external attacks, incidents or leakage.

We have set up a cloud server that can be accessed by developer-teams of the consortium (e.g. AMAT, EPFL, Fraunhofer, CNR, and ICN2). This allows for a shared software development on the same platform, and a quick integration between the different IM2D box components. However, having such a server could also be a risk. The two main security measures we adopted are: (i) authentication in the server via ssh-keys, a solution safer than simple password access; (ii) IP address restrictions, so that only authorized people can access the server. Security and restrictions are particularly relevant when web components, e.g. AiiDA-POST, are used to control/submit calculations on external servers/supercomputers.

Security is a fundamental issue not only for not-disclosing of sensitive data but also for the intellectual property protection of proprietary codes, such as Ginestra™. Indeed, the adopted solution allows AMAT (Ginestra™ owner) to keep the original source code and the related components within their firewalls. The AiiDA-Ginestra applet (see D2.2 and D2.4) can, in fact, connect to the remote server and communicate with AiiDa (and thus with all the other components of IM2D) via a HTTP protocol.

## Conclusion

This document updates and completes the FDMP. In particular, it focuses on the integration of IM2D to external materials databases, through the OPTIMADE REST API gateway; on the integration of the semantic interoperability through SimPhoNy as an extension to the already available syntactic layer; and on a few measurements for the security of data and code sources.

## References

- [1] INTERSECT project [www.intersect-project.eu](http://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](http://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™ 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™ URL: [www.mdlsoft.com](http://www.mdlsoft.com)
- [5] Pizzi G., et al. 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](http://www.simphony-project.eu)
- [7] Andersen et al., The OPTIMADE Specification, DOI: [10.5281/zenodo.5195050](https://doi.org/10.5281/zenodo.5195050)
- [8] MaterialsCloud <https://www.materialscloud.org>
- [8] The MarketPlace project <https://www.the-marketplace-project.eu>
- [9] The official Allegrograph server as Docker container image <https://franz.com/agraph/support/documentation/current/docker.html>



## ACRONYMS

**API** - Application Programming Interface

**BDSS** - Business Decision Support System

**CIF** - Crystallographic Information File

**COD** - Crystallography Open Database

**CUDS** - Common Universal Unified Data Structures

**DAG** - Directed Acyclic Graph

**DFT** - Density Functional Theory

**EMMC** - European Materials Modelling Council

**EMMO** - European Materials Modelling Ontology

**FDMP** - First Data Management Plan

**FORCE** - Formulations And Computational Engineering

**EMMO** - European Materials Modelling Ontology

**GUI** - Graphical User Interface

**iHub** - Interoperability Hub

**IM2D** - Interoperable Materials-To-Device

**JSON** - JavaScript Object Notation.

**OTE** - Open Translation Environment

**OWL** - Ontology Web Language

**REST** - Representational State Transfer

**RDF** - Resource Description Framework

**QE** - Quantum Espresso

**SPARQL** - SPARQL Protocol And RDF Query Language

**TCOD** - Theoretical Crystallography Open Database



HORIZON2020

Deliverable D4.6

Data Management Plan Assessment and Revision



**TTL** - Terse RDF Triple Language (Turtle)

**UUID** - Universally Unique Identifier

**XML** - Extensible Markup Language