



Agenzia nazionale per le nuove tecnologie,  
l'energia e lo sviluppo economico sostenibile

# Metadata in materials: the new path forward

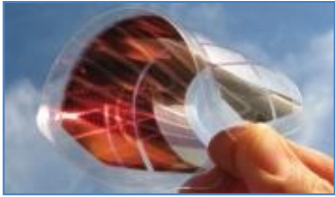
HPC & Data to run faster than COVID

*Rome, 05.11.2020*

**Massimo Celino, ENEA, TERIN-ICT**



# Materials for energy



Advanced Materials  
“enabling”  
energy technologies

 EUROPEAN COMMISSION



Brussels, 13.12.2011  
SEC(2011) 1609 final

COMMISSION STA  
Materials Roadmap Enabling I

*“With the imperative to change our energy technology mix to respond to the challenge of decarbonisation and of the security of energy supply, the need for new materials and processing routes is overriding. **We need new energy technologies – not just any low carbon technologies, but more efficient and cost-competitive low carbon technologies.** **Materials play a pivotal role in the solution, providing the means to generate and conserve energy in a more efficient and cost-competitive manner.**”*

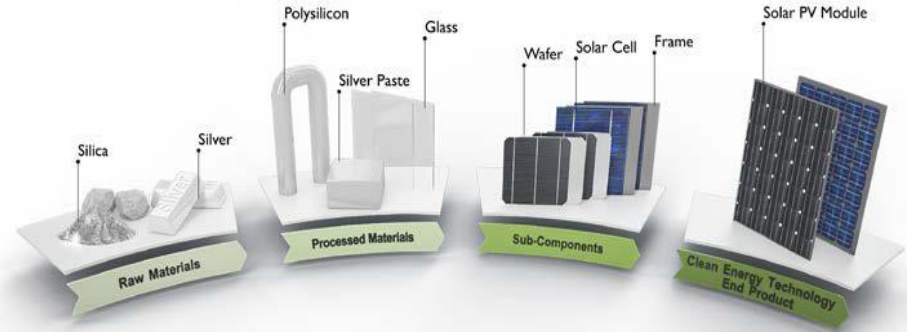



# Manufacturing Crystalline Silicon PV Modules

Raw Materials	Processed Materials	Sub-Components	Clean Energy Technology End Product
Silica, Silver	Polysilicon, Silver Paste, Glass	C-Si PV Wafer, C-Si PV Cell, Frame, Encapsulant	C-Si Solar PV Module 
Iron, Neodymium, or Dysprosium Ores	Steel, Fiberglass, Carbon Fiber, Neodymium and Dysprosium Alloys	Permanent Magnets, Generators, Gear Assemblies, Steel Components	Wind Turbine Components: Blades, Tower, Nacelle 
Lithium, Cobalt, Nickel, Graphite Ores	Cathode Materials, Anode Materials, Electrolytes	Separators, Housings, Metal Foils, Tabs	Light Duty Vehicle (LDV) LI-ion Battery Cell 
Gallium, Indium, Yttrium Ores	Sapphire Substrates, Trimethyl Gallium (TMG), Trimethylindium (TMI), YAG Phosphors	LED Chips	LED Package 

\* Processed Materials = Advanced Materials (US DOE's CEMAC reports)

Materials are the core of clean technologies



TOP 5 Solar Panel Manufacturers [Ranked by shipment guidance (GW)]	
 1. Sharp	 1. Trina Solar
 2. First Solar	 2. Yngli Green Energy
 3. Yngli Green Energy	 3. Canadian Solar
 4. Kyocera	 4. Hanwha SolarOne
 5. Trina Solar	 5. Jinko Solar
2008	2015

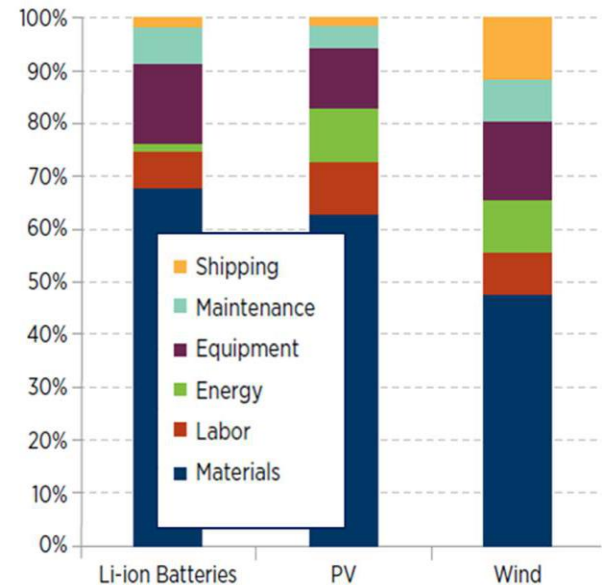
EMIRI, the Energy Materials Industrial Research Initiative, represents more than 50 organizations (industry, research, associations) active in advanced materials & nanotech for clean energy & clean mobility technologies.

Advanced materials (plastics, non ferrous metals, steel, glass, ceramics ...) represent today **beyond 50% of the cost** structure of clean energy & clean mobility technologies.

Moreover trends like **Industry 4.0 will squeeze out labor and energy costs possibly bringing the share of advanced materials in the cost structure up to 80%**

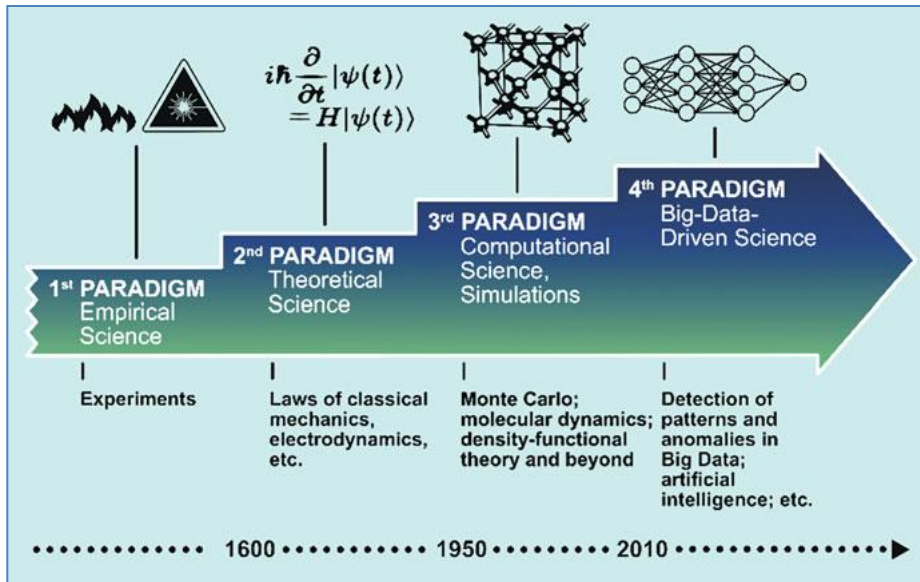
For instance, advanced materials represent today more than 60% of the cost structure of Li-ion battery cells which are key for Europe's transition to clean mobility.

<https://emiri.eu/>



# The four paradigms

Development of research paradigms of materials science and engineering



## 4V challenge:

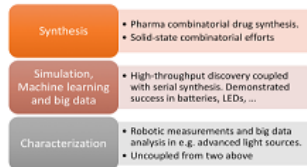
1. **Volume** (the amount of data),
2. **Variety** (the heterogeneity of form and meaning of data),
3. **Velocity** (the rate at which data may change or new data arrive)
4. **Veracity** (the uncertainty of data quality)

**NOMAD: The FAIR concept for big data-driven materials science**

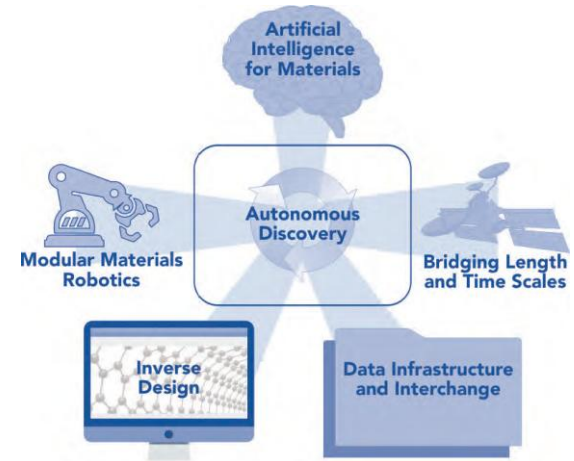
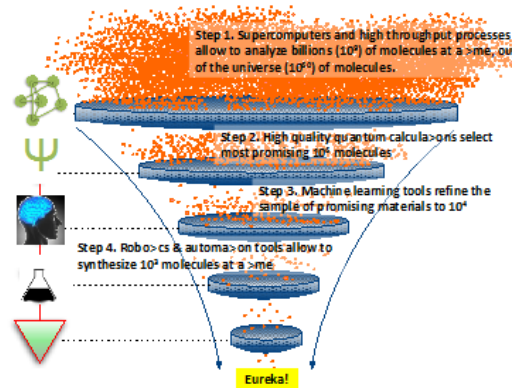
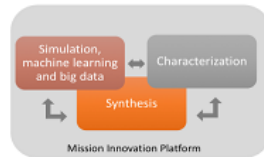
Claudia Draxl and Matthias Scheffler

MRS BULLETIN • VOLUME 43 • SEPTEMBER 2018

a) Current uncoupled high-throughput approaches



b) Proposed integrated approach

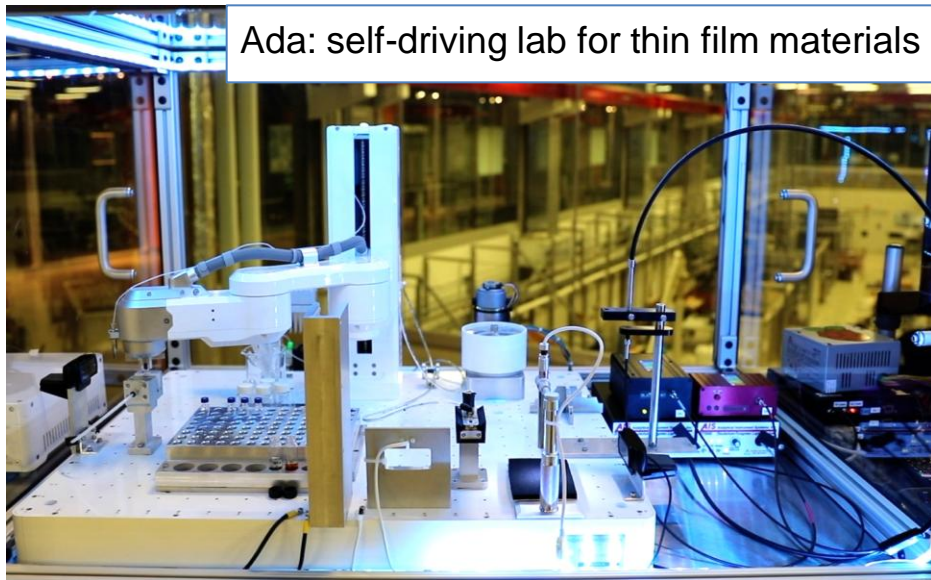


## Accelerating Advanced Energy Materials Discovery by Integrating High-Throughput Methods with Artificial Intelligence



# Data driven experiments

Ada: self-driving lab for thin film materials



Canada Natural Resources Canada Ressources naturelles Canada

Canada



UNIVERSITY OF TORONTO

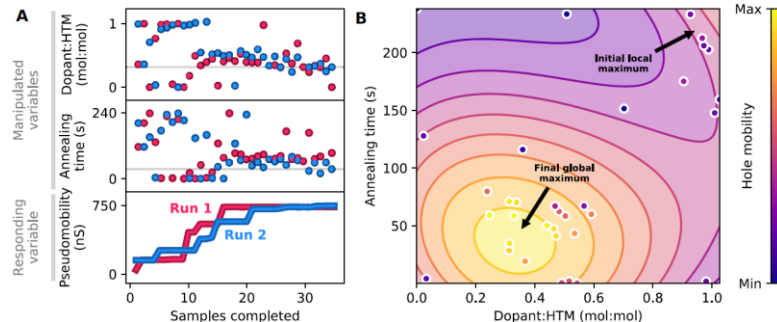
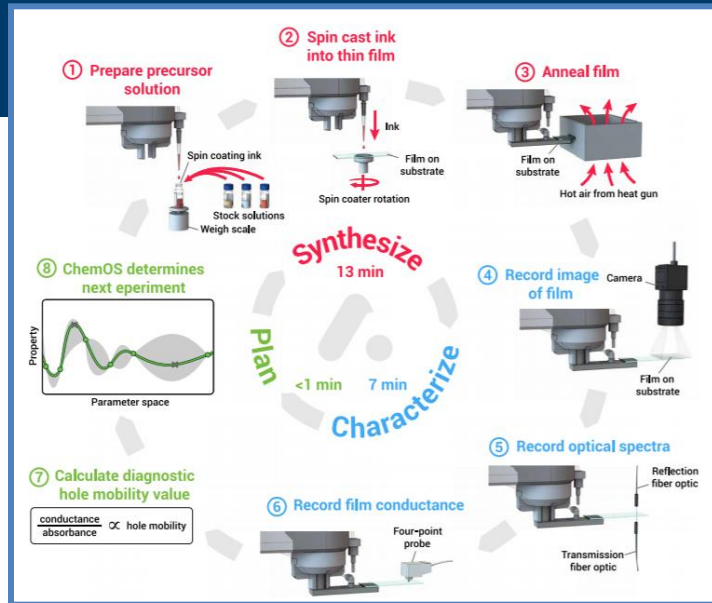
CIFAR

NORTH ROBOTICS

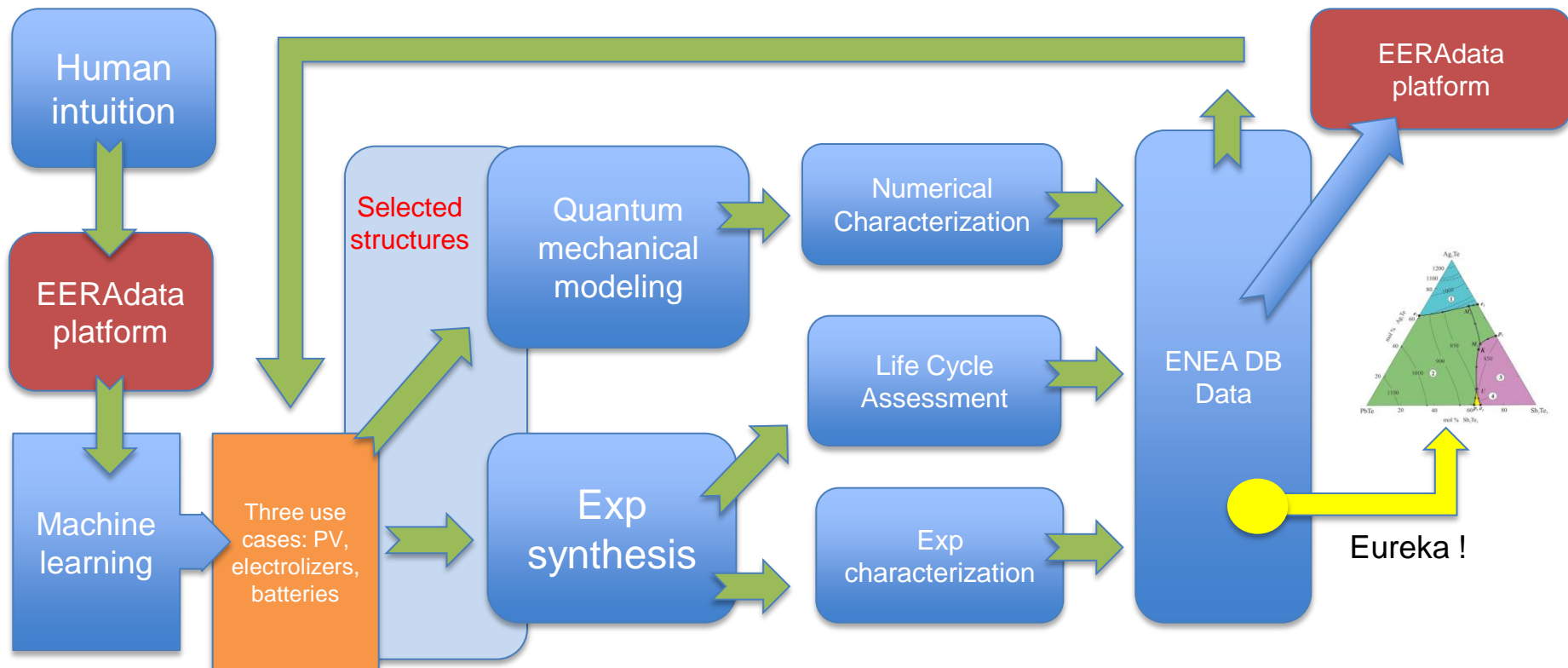


THE UNIVERSITY OF BRITISH COLUMBIA

ENEA



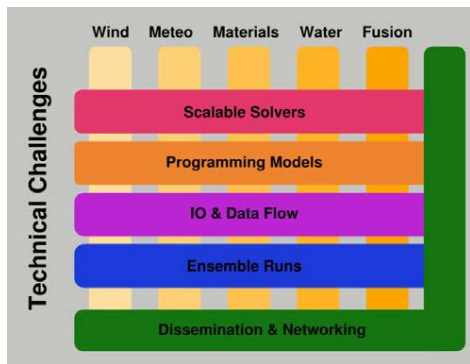
# IEMAP: Italian Energy Materials Acceleration Platform







Transversal multidisciplinary effort providing high-end expertise in **applied mathematics** and **HPC** to face energy scientific challenges and exploit the new European exascale ( $10^{18}$ ) supercomputers



- EOCOII (2<sup>nd</sup> phase): 3 years project (2019 – 2021)
- Coordinator : Prof. E. Audit (Maison de la Simulation, CEA, Saclay)
- 7 countries, 18 partners



# Data repositories

1. NOMAD, NOvel MAterials Discovery Repository (<https://repository.nomad-coe.eu>)
2. Materials Cloud ([www.materialsproject.org](http://www.materialsproject.org))
3. MATDB (<https://odin.jrc.ec.europa.eu/alcor/Main.jsp>)
4. Open Materials Database (<http://openmaterialsdb.se/>)
  
5. AFLOW, Automatic FLOWfor materials discovery (<http://aflowlib.org/>)
6. OQMD, Open quantum materials database (<http://oqmd.org/>)
7. T-COD, Theoretical Crystallography Open Database ([www.crystallography.net/tcod](http://www.crystallography.net/tcod))
8. High Throughput Experimental Materials (HTEM) Database ([htem.nrel.gov](http://htem.nrel.gov))

# Metadata

There are several examples of metadata approaches in computational materials science, the most prominent being:

- **ChemML**, the Chemical Markup Language ([www.xml-cml.org](http://www.xml-cml.org)),
- **CIF**, the Crystallography Information File ([www.iucr.org/resources/cif](http://www.iucr.org/resources/cif)),
- **VASP** (<http://cms.mpi.univie.ac.at/vasp/vasp/vasp.html>) or
- **Molpro** ([www.molpro.net/info/2012.1/doc/manual/manual.html](http://www.molpro.net/info/2012.1/doc/manual/manual.html))
- **XML** outputs
- **XYZ** ([http://openbabel.org/wiki/XYZ\\_\(format\)](http://openbabel.org/wiki/XYZ_(format))) for the storage of the configuration of a system (atomic species, coordinates, also allowing for comments).

In general, a metadata structure is built a priori, by starting from a list of names that identify the needed concepts and quantities. Code outputs and file formats are then designed to reflect the a priori metadata structure.

# MatDB

The European Commission JRC [ODIN Portal](#) hosts a number of scientific databases, one of which is [MatDB](#). MatDB is a database application designed to store mechanical test data coming from tests performed in accordance with mechanical testing standards.

The **metadata are organized into categories** relevant to materials testing, so that there are main entities:

- for source (i.e. provenance),
- material (i.e. production, heat treatment, microstructure, etc),
- specimen,
- test condition,
- documents and
- test result.

The access management model supports both open access and restricted access. With a view to FAIR compliance, the database supports data citation (using the [DataCite](#) framework) and interoperability standards for test data (hosted at [CEN](#)).



The screenshot displays the MatDB website interface. At the top, there is a navigation bar with the European Commission logo, the text "JOINT RESEARCH CENTRE", and links for "Legal Notice / Brexit Content Disclaimer / Language Policy / Privacy Policy / Cookies / Contact / Search / English (en) / Login". Below this is a blue header with "MatDB Open Access" and "European Commission / EU Science Hub / ODIN / MatDB Open". A light blue navigation bar contains "DATA RETRIEVAL", "DATA ENTRY", "SANDBOX -", and "FEEDBACK". The main content area features a large image of interlocking gears and a "DATA RETRIEVAL" button with a home icon. Below the image, a text block states: "The Open Access data retrieval module allows you to select and view MatDB data without needing to register or login. Presently, the collection of Open Access data is relatively small. To access the larger collection of Open Access and registered access data, please login." There is also a "DATA ENTRY" button with a document icon. At the bottom, a text block explains: "Data Entry allows you to enter and manage your test data in MatDB. Data entry rights are granted on a case by case basis. If you are interested to use MatDB for managing your project data, please Register for an ODIN account and then request data entry rights from your My Profile page."

# Materials Cloud

## Vision and purpose

Materials Cloud is built to enable the seamless sharing and dissemination of resources in computational materials science, offering educational, research, and archiving tools; simulation software and services; and curated and raw data. These underpin published results and empower data-based discovery, compliant with data management plans and the FAIR principles.

Share your scientific results on Materials Cloud to make them

- **Comprehensive:** Share the *entire workflows and provenance graphs* of your calculations, and not just individual *input and output files*.
- **Downloadable:** Download individual files or entire databases at the click of a button.
- **Browsable:** Browse and query calculations directly from a web browser.
- **Repurposable:** Download and import any database and start your calculations from where the original authors left off.

Materials Cloud is powered by [AiiDA](#), an open-source python infrastructure to manage and persist the ever-growing amount and complexity of workflows and data in computational science.

<https://www.materialscloud.org>



**MATERIALSCLOUD**

Built for seamless sharing of resources  
in computational materials science.

**LEARN** Lectures and tutorials in computational materials science

**WORK** Simulation tools and services - in the cloud or on your computer

**DISCOVER** Curated research data with tailored visualizations

**EXPLORE** Interactive browser for AiiDA provenance graphs

**ARCHIVE** An open-access, moderated repository for research data in computational materials science

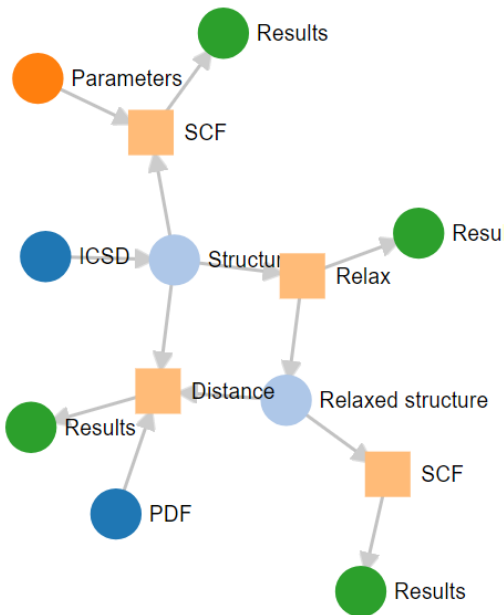
Please cite [L. Talirz et al., Sci Data 7, 299 \(2020\)](#), if you use Materials Cloud in your research.

## What is AiiDA?

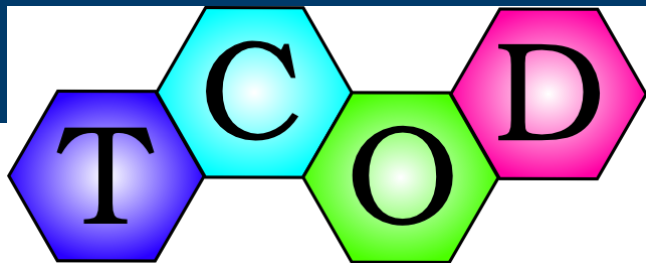
AiiDA is an open-source Python infrastructure to help researchers with automating, managing, persisting, sharing and reproducing the complex workflows associated with modern computational science and all associated data.

AiiDA is built to support and streamline the four core pillars of the [ADES model](#): Automation, Data, Environment, and Sharing. Key features include:

- **Workflows:** AiiDA allows to build and execute complex, auto-documenting workflows linked to multiple codes on local and remote computers.
- **High-throughput:** AiiDA's event-based workflow engine supports tens of thousands of processes per hour with full check-pointing.
- **Data provenance:** AiiDA automatically tracks and records inputs, outputs and metadata of all calculations and workflows in extensive provenance graphs that preserve the full lineage of all data.
- **Advanced queries:** AiiDA's query language enables fast graph queries on millions of nodes.
- **Plugin interface:** AiiDA can support via plugins any computational code and data analytics tool, data type, scheduler, connection mode, etc. (see [public plugin repository](#))
- **HPC interface:** AiiDA can seamlessly deal with heterogeneous and remote computing resources; it works with many schedulers out of the box (SLURM, PBS Pro, torque, SGE or LSF).
- **Open science:** AiiDA allows to export both full databases and selected subsets, to be shared with collaborators or made available and browsable online on the [Archive](#) and [Explore](#) sections of [Materials Cloud](#).
- **Open source:** AiiDA is released under the MIT open-source license.







Open-access collection of theoretically calculated or refined crystal structures of organic, inorganic, metal-organic compounds and minerals

RESEARCH ARTICLE

Open Access



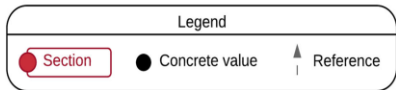
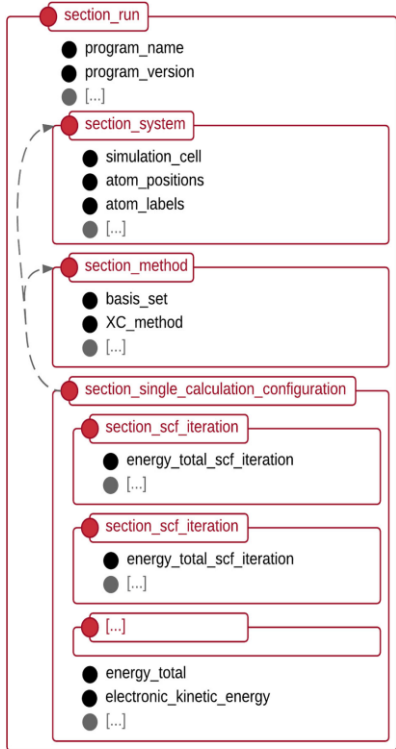
# A posteriori metadata from automated provenance tracking: integration of AiiDA and TCOD

Andrius Merkys<sup>1,2\*</sup> , Nicolas Mounet<sup>1</sup>, Andrea Cepellotti<sup>1</sup>, Nicola Marzari<sup>1</sup>, Saulius Gražulis<sup>2,3</sup>   
and Giovanni Pizzi<sup>1</sup>

## Abstract

In order to make results of computational scientific research findable, accessible, interoperable and re-usable, it is necessary to decorate them with standardised metadata. However, there are a number of technical and practical challenges that make this process difficult to achieve in practice. Here the implementation of a protocol is presented to tag crystal structures with their computed properties, without the need of human intervention to curate the data. This protocol leverages the capabilities of AiiDA, an open-source platform to manage and automate scientific computational workflows, and the TCOD, an open-access database storing computed materials properties using a well-defined and exhaustive ontology. Based on these, the complete procedure to deposit computed data in the TCOD database is automated. All relevant metadata are extracted from the full provenance information that AiiDA tracks and stores automatically while managing the calculations. Such a protocol also enables reproducibility of scientific data in the field of computational materials science. As a proof of concept, the AiiDA–TCOD interface is used to deposit 170 theoretical structures together with their computed properties and their full provenance graphs, consisting in over 4600 AiiDA nodes.

**Keywords:** DFT, Reproducibility, Provenance, Open data, Ontology, Materials science



Currently, the NOMAD Repository holds the information of over 40 million total-energy calculations (corresponding to converged single point, atomic structure, calculations), which corresponds to several billion CPU hours. NOMAD offers their users to restrict (for up to 3 years) their upload to themselves and other selected users, or to make them “open access” right away.

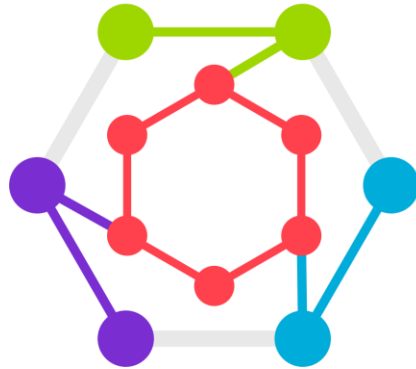
NOMAD Repository contains also all the original inputs and outputs of the data in the AFLOW, OQMD, and Materials Project databases. While this has been and is useful for its purpose of data sharing via a repository (enabling the confirmatory analysis of materials data, their reuse, and repurposing), the data is very heterogeneous, as they are provided by the different codes.

“NOMAD Meta Info”, i.e., the metadata defined and used for the NOMAD code independent data format, is defined a posteriori, by starting from the existing inputs and outputs of many different codes stemming from different scientific fields and thus with quite diverse feature sets.

Three formats to store data using the NOMAD Meta Info: the human readable JSON, HDF5, and Apache Parquet.

<https://repository.nomad-coe.eu>

Ghiringhelli et al. , npj Computational Materials (2017) 3:46 ; doi:10.1038/s41524-017-0048-5



OPTIMADE

Open Databases Integration  
for Materials Design

The **Open Databases Integration for Materials Design** (OPTIMADE) consortium aims to make materials databases interoperational by developing a common REST API.

Massimo Celino  
massimo.celino@enea.it



1101 0110 1100  
0101 0010 1101  
0001 0110 1110  
1101 0010 1101  
1111 1010 0000

