



Deliverable D5.6  
Report on the data stewardship policies and on the  
scalability and features of the Materials Cloud Archive

## D5.6

# Report on the data stewardship policies and on the scalability and features of the Materials Cloud Archive

Marco Borelli, Valeria Granata, Nicola Marzari, Elsa Passaro,  
and Giovanni Pizzi

Due date of deliverable: 30/11/2021  
Actual submission date: 30/11/2021

Lead beneficiary: EPFL (participant number 6)  
Dissemination level: PU - Public



Deliverable D5.6  
Report on the data stewardship policies and on the  
scalability and features of the Materials Cloud Archive

## Document information

Project acronym:	MaX
Project full title:	Materials Design at the Exascale
Research Action Project type:	European Centre of Excellence in materials modelling, simulations and design
EC Grant agreement no.:	824143
Project starting / end date:	01/12/2018 (month 1) / 31/05/2022 (month 42)
Website:	<a href="http://www.max-centre.eu">www.max-centre.eu</a>
Deliverable No.:	D5.6

**Authors:** M. Borelli, V. Granata, E. Passaro, N. Marzari, and G. Pizzi

**To be cited as:** M. Borelli et al., (2021): Report on the data stewardship policies and on the scalability and features of the Materials Cloud Archive. Deliverable D5.6 of the H2020 project MaX (final version as of 30/11/2021). EC grant agreement no: 824143, EPFL, Lausanne, Switzerland.

## 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.



Deliverable D5.6  
Report on the data stewardship policies and on the  
scalability and features of the Materials Cloud Archive

## D5.6 Report on the data stewardship policies and on the scalability and features of the Materials Cloud Archive

### Content

<b>Executive Summary</b>	<b>4</b>
<b>Introduction</b>	<b>4</b>
<b>Materials Cloud Archive</b>	<b>6</b>
<b>Conclusions</b>	<b>11</b>



## 1 Executive Summary

This document provides a description of the solutions adopted within the MAX Center of Excellence (CoE) to implement data stewardship policies and deploy the infrastructure that guarantees compliance with FAIR principles and European and national data management plans.

Sharing research data in a FAIR format is crucial to guarantee reproducibility, increase transparency and impact of research and accelerate discovery. The MAX CoE supports long-term sharing and preservation of the data needed to reproduce a scientific paper through 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, which guarantees a Findable, Accessible, Interoperable and Reusable (FAIR)-compliant sharing of data produced by AiiDA; in addition, together with the Materials Cloud Discover section, also for sharing highly curated data in the field of materials science.

The Materials Cloud Archive has recently been refactored to use the Invenio scalable digital library framework in order to highly improve the robustness, scalability and user-friendliness of the platform. We discuss all efforts that have taken place during the lifetime of MAX, to ensure robustness, scalability, user-friendliness and sustainability of the platform.

## 2 Introduction

This document is deliverable D5.6 of the MAX project and briefly describes the activities we carried out to accomplish task T5.4, which sets the goal of improving data stewardship practices in MAX and for the public through multiple actions.

First and foremost, we moved the Materials Cloud Archive to the Invenio 3 framework, which allowed us to make it more scalable, user-friendly, and maintainable. Invenio (<https://inveniosoftware.org>) is an open-source software framework developed at CERN, designed to implement and deploy large-scale digital repositories.

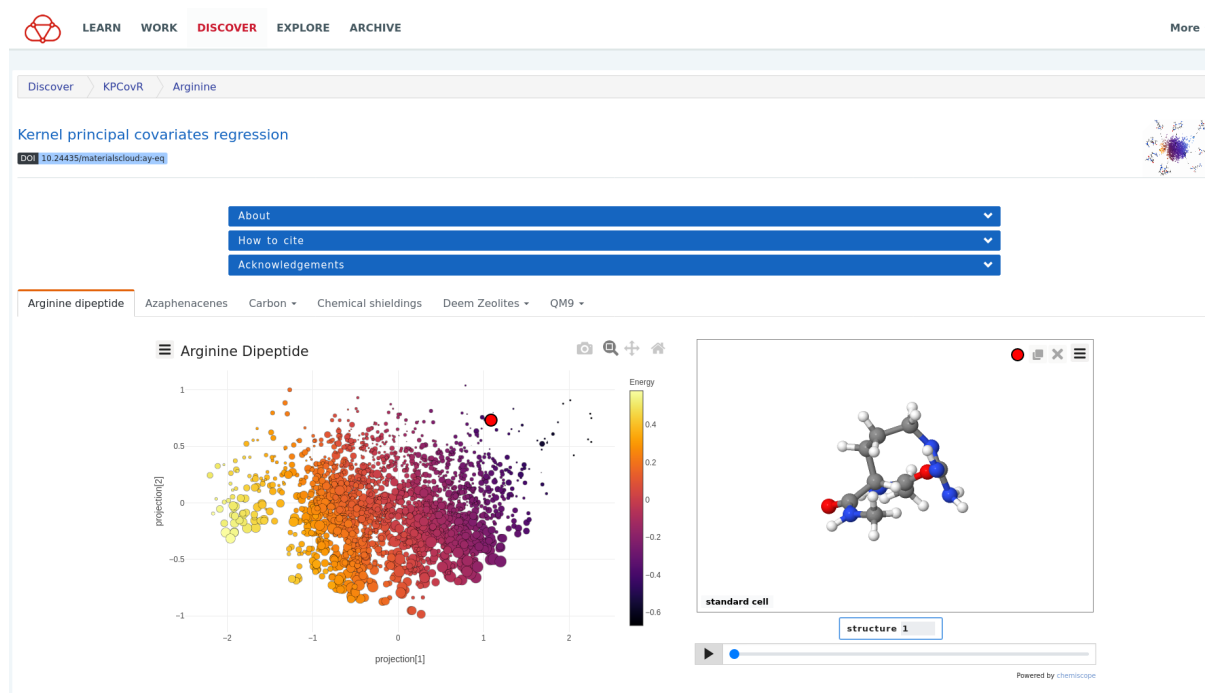
Since Materials Cloud Archive is a moderated repository, we implemented from scratch a moderation workflow inside the Invenio platform. Therefore, the submission and moderation processes are now simpler for the user, nor require technical knowledge of the platform codebase by the moderators. All published data records are automatically assigned a Datacite DOI (via a partnership and contract with the ETHZ library in Zurich), and are guaranteed to be available for at least 10 years thanks to a partnership with CSCS (Swiss Supercomputing Centre), where the platform and data are hosted. Finally, users can self-manage their data records whenever moderation is not necessary (e.g. to update the citations to scientific papers, or to update keywords).

As part of WP5, a Data Management Plan (DMP) for MAX was redacted and provided as [Deliverable D5.1](#). For researchers funded by the Swiss SNSF or by EU projects (e.g. in the Horizon 2020 program), we also offer DMP templates on the Materials Cloud (<https://www.materialscloud.org/dmp>) that help them explain how their data is compliant with requirements from their funding agencies, when they upload their data to the Materials Cloud Archive (and/or use AiiDA as a data management solution).

Deliverable D5.6

Report on the data stewardship policies and on the scalability and features of the Materials Cloud Archive

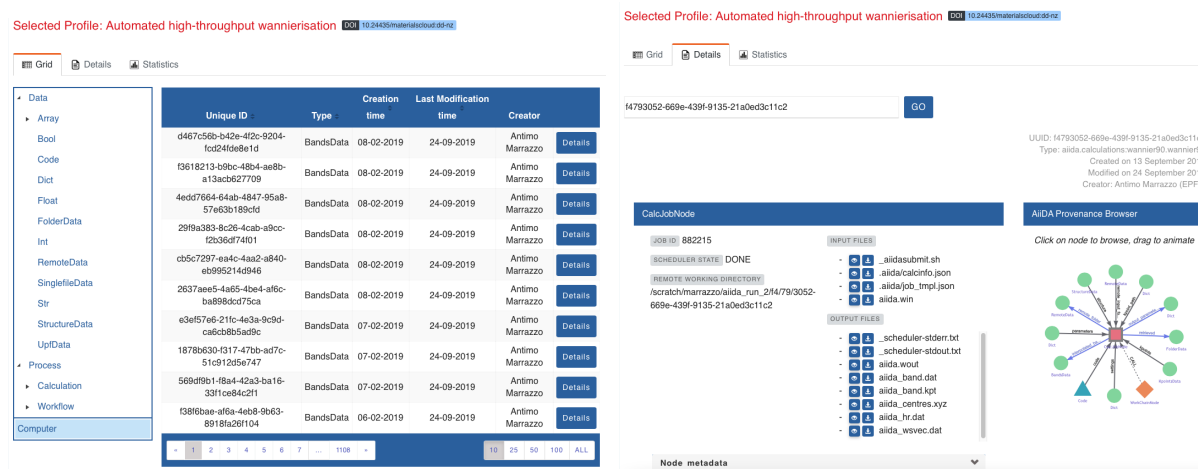
Finally, the Discover and Explore sections of Materials Cloud improve and make more accessible some of the data in the Archive section, when these sections are provided by the data authors. Materials Cloud Discover offers curated and interactive visualizations of contributed datasets, which support and enhance their authors' research. These sections are prepared in collaboration with Materials Cloud developers. Materials Cloud Explore provides a web-based provenance browser for AiiDA databases (for those datasets generated with AiiDA): this includes several records uploaded to the Archive, but we also implemented the graphical UI so that the same interface allows a user to browse their own local database, without any data leaving their computer. These features are supported by AiiDA's powerful REST API and by the successful integration of all MAX flagship codes via AiiDA plugins. All Explore and Discover entries are linked to the Archive record containing the source raw data. Figs. 1 and 2 show a couple of examples of Discover and Explore sections that are currently available online.



**Fig. 1:** Example of Discover entry: interactive visualization of curated data (in this case, exploring large datasets of crystal structures and inspecting graphically the structural similarity and the correlation with materials properties).

## Deliverable D5.6

### Report on the data stewardship policies and on the scalability and features of the Materials Cloud Archive



**Fig. 2:** Example of Explore entry: browsing through the AiiDA provenance of a dataset uploaded to the Archive (in this case, a project on automated high-throughput Wannierisation of materials).

## 3 Materials Cloud Archive

The Materials Cloud Archive is one of the five sections of Materials Cloud (<https://www.materialscloud.org>). It is a moderated data repository that allows researchers to upload files related to research data associated with scientific publications, and to obtain DOIs for them, with the guarantee that data will be preserved for at least 10 years after their publication.

Many of the data repositories currently available (with DOIs and long-term preservation), for example Zenodo (<https://zenodo.org>) or FigShare (<https://figshare.com>), are limited to storing and providing the data as files (or with simple viewers), but there is no way to represent graphically complex provenance relationships between data items (as e.g. the provenance tracked in the form of acyclic directed graphs by AiiDA). Therefore, these tools provide Findability (via DOIs) and Reusability (via open licenses), but often do not allow full FAIR sharing of complex, interconnected datasets, where also Accessibility and Interoperability are needed. Moreover, they do not provide straightforward ways to share large datasets (having hard limits for generic users). Finally, they are general repositories, containing data from multiple disciplines. We instead believe that having field-specific repositories is a booster for research and discovery, because it makes it easier to discover and find high-quality datasets that might be of interest to researchers.

In the Materials Cloud Archive, when available, links to corresponding Discover and Explore sections of Materials Cloud are displayed; data is then also presented interactively in these sections, thus providing further layers of accessibility, interoperability and reproducibility.

In May 2020 we published a completely new version of the Archive (<https://archive.materialscloud.org>) developed using the robust Invenio (v3) framework (<https://inveniosoftware.org/products/framework/>), making it scalable and open to users beyond the scientific projects currently sponsoring it. The Invenio framework is developed at CERN and it is the framework behind the well-known Zenodo repository.

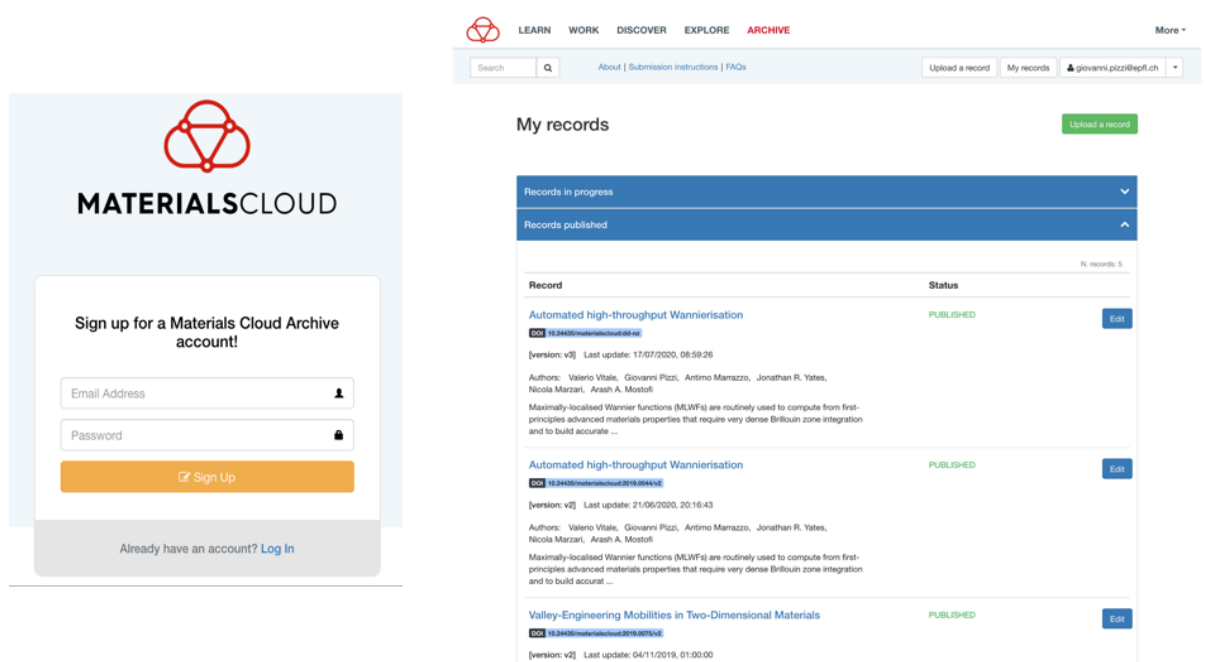
Deliverable D5.6

Report on the data stewardship policies and on the scalability and features of the Materials Cloud Archive

Our Archive leverages several of the features already provided by the framework, such as the built-in search engine, the possibility to create user accounts (see Fig. 3), and to have multiple versions of a record.

As the Archive is a moderated repository, we implemented a customised workflow for the moderation of records. Users have access to a personal area where they can browse through the records they submitted, and check the status of their records throughout the entire publication process (see Fig. 3). Once published, users can also update keywords and citations of their own records.

In addition, also moderators are now offered a custom web page where they can check and moderate (accept and publish, ask for changes, or reject) entries directly from the browser (see Fig. 4), as well as send messages to the entry owners in case communication is required (e.g. to ask for clarifications or changes).

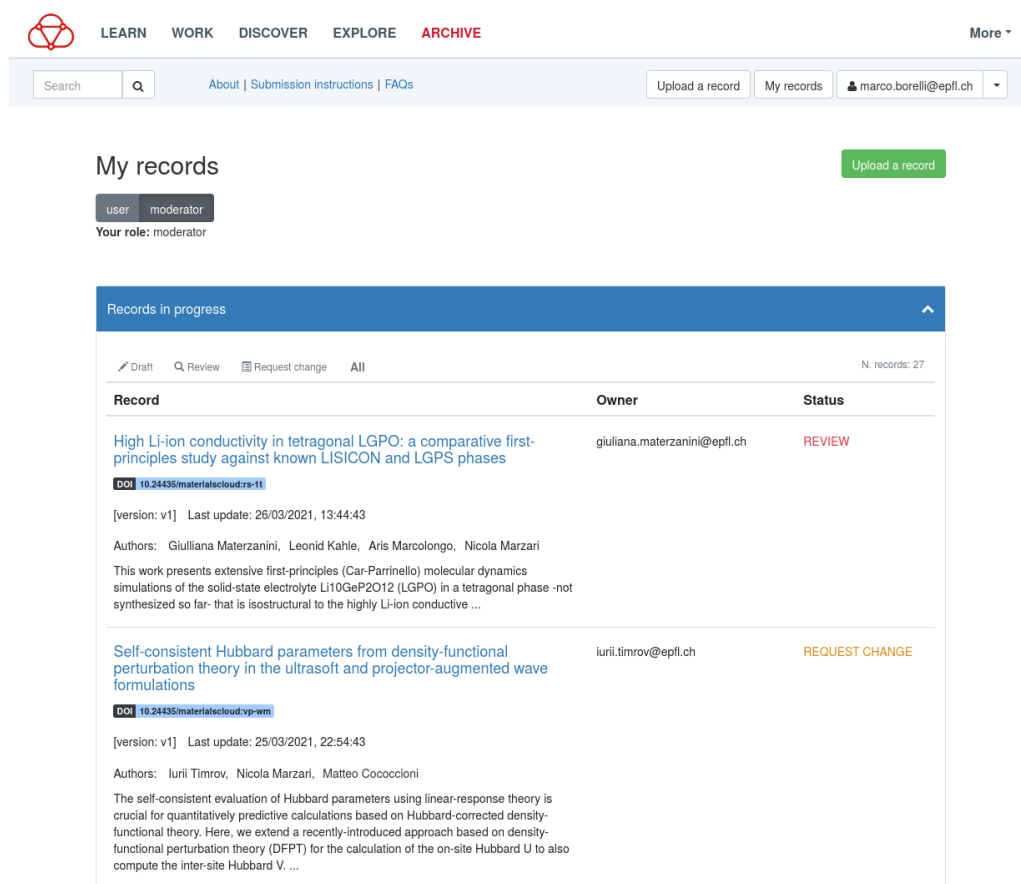


**Fig. 3:** Personal user account and work area. Left: login page; right: personal page of each user. Users can browse through their submitted records, check their status at any time, update them, or create new versions of records already published.

The Archive uses the Invenio user management and authentication module “invenio-accounts” for user registration, password reset/recovery and email verification (<https://invenio-accounts.readthedocs.io>). User passwords are stored encrypted in the Invenio postgres database using a strong cryptographic hashing algorithm (pbkdf2\_sha512). The Archive uses this module also to set the roles of moderator and administrator to a user.

Deliverable D5.6

Report on the data stewardship policies and on the scalability and features of the Materials Cloud Archive



Record	Owner	Status
<p>High Li-ion conductivity in tetragonal LGPO: a comparative first-principles study against known LISICON and LGPS phases</p> <p>DOI: 10.24435/materialscloud-rs-11</p> <p>[version: v1] Last update: 26/03/2021, 13:44:43</p> <p>Authors: Giuliana Materzanini, Leonid Kahle, Aris Marcolongo, Nicola Marzari</p> <p>This work presents extensive first-principles (Car-Parrinello) molecular dynamics simulations of the solid-state electrolyte Li10GeP2O12 (LGPO) in a tetragonal phase - not synthesized so far - that is isostructural to the highly Li-ion conductive ...</p>	giuliana.materzanini@epfl.ch	REVIEW
<p>Self-consistent Hubbard parameters from density-functional perturbation theory in the ultrasoft and projector-augmented wave formulations</p> <p>DOI: 10.24435/materialscloud-vp-wm</p> <p>[version: v1] Last update: 25/03/2021, 22:54:43</p> <p>Authors: Iurii Timrov, Nicola Marzari, Matteo Cococcioni</p> <p>The self-consistent evaluation of Hubbard parameters using linear-response theory is crucial for quantitatively predictive calculations based on Hubbard-corrected density-functional theory. Here, we extend a recently-introduced approach based on density-functional perturbation theory (DFPT) for the calculation of the on-site Hubbard U to also compute the inter-site Hubbard V. ...</p>	iurii.timrov@epfl.ch	REQUEST CHANGE

**Fig. 4:** Moderator's work area: a moderator can publish, reject, or request changes to a record. The exchanges between users and moderators are recorded and displayed.

The Archive data and metadata are stored at the Swiss National Supercomputing Centre (CSCS, <https://www.cscs.ch>) in Lugano. The files associated with records are stored in a container of the CSCS object store and the application runs on virtual machines.

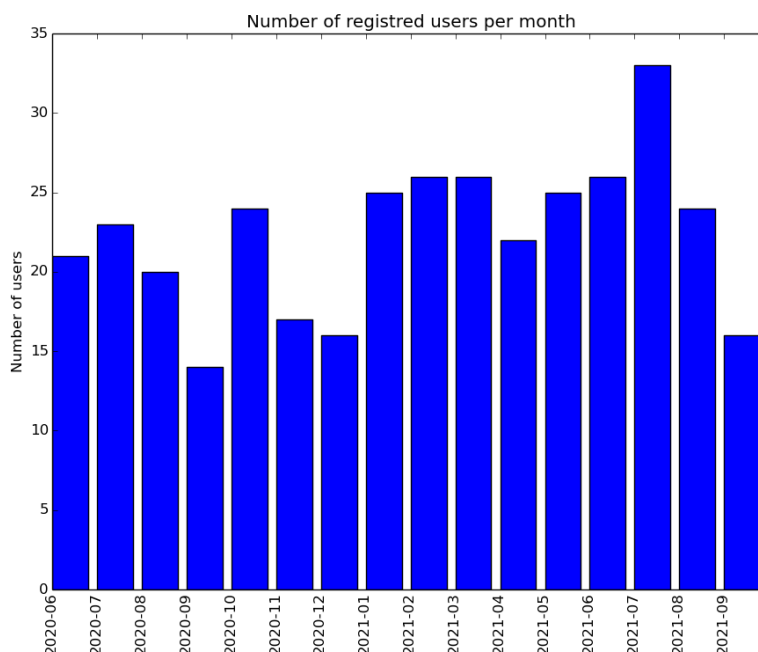
A DOI is associated with each record and is reserved the moment the record is created in the Archive. The DOIs will only resolve once the records are published; their registration is provided by the ETH library in Zurich.

Detailed guidelines and policies for submission of records on the Archive have been prepared and are available on the site (<https://archive.materialscloud.org/deposit/information>). In these documents it is clearly explained how to submit a record, what are the criteria for a record to be approved for publication, the open licences and open file formats accepted, and how records will be reviewed and finally published by the moderators.

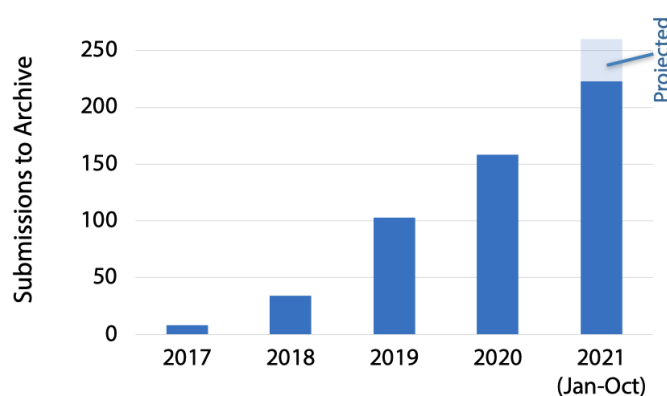
To guarantee permanent access to the Archive, a failover system on Amazon Web Services (AWS) was implemented. The Archive deployed on AWS is read-only and guarantees access to published records data and metadata even in case of (planned or unexpected) downtimes and maintenance of the CSCS infrastructure.



As of October 5th 2021, there are 471 users registered on the Archive of which 252 have uploaded at least one record (see Fig. 5 for the number of users creating a new account on the platform every month). The records published are 495 (including the versions of records); a breakdown for every year since the creation of Materials Cloud Archive is presented in Fig. 6, showing a clear growth trend of the number of entries over time.



**Fig. 5:** Number of new registered users per month since the new Archive was published in May 2020 (data as of October 2021), showing a constant trend of about 20 to 30 users per month.



**Fig. 6:** Number of records published per year since the Archive was published for the first time in 2017 (data as of October 2021; a linear projection of the submissions to the end of Dec 2021 is also shown).

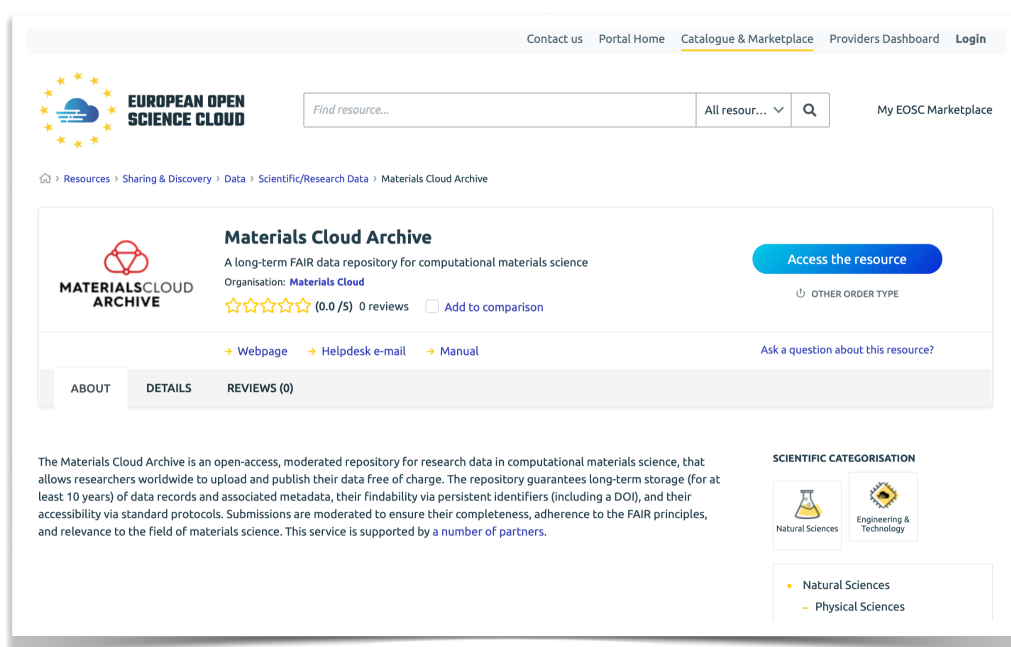
Deliverable D5.6

Report on the data stewardship policies and on the scalability and features of the Materials Cloud Archive

The Materials Cloud ARCHIVE is registered both on the re3data and FAIR sharing repository registries. It is indexed by Google Dataset Search as well as the B2FIND portal by EUDAT (<http://b2find.eudat.eu>) that is now part of EOSC.

In addition, Materials Cloud Archive is one of the services that is offered also via the EOSC Marketplace portal (see Fig. 7).

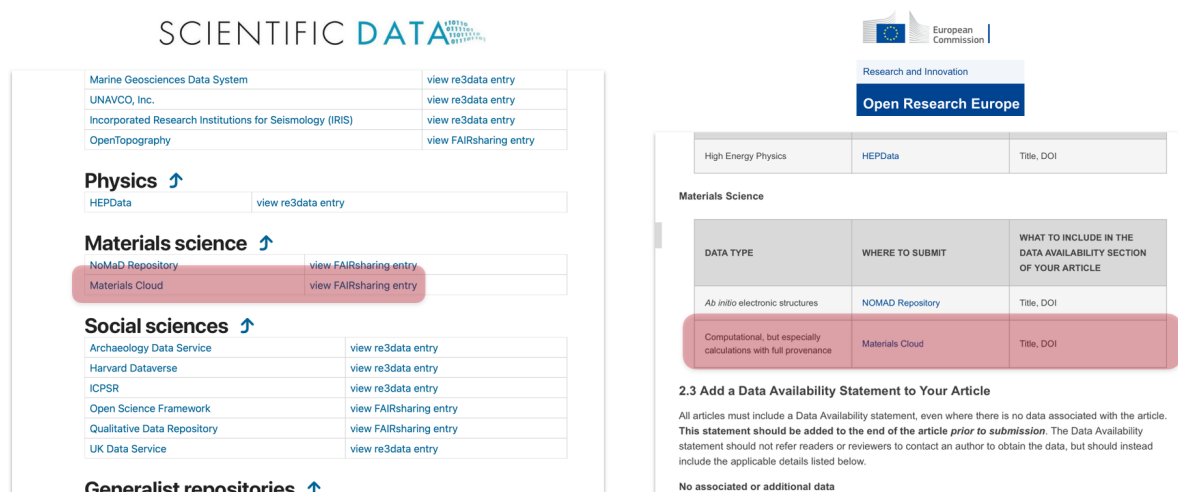
Finally, Materials Cloud is a recommended repository for materials science by two key scientific journals in the field: Nature's journal *Scientific Data* and EC commission's new journal *Open Research Europe* (see Fig. 8).



**Fig. 7:** Materials Cloud Archive is one of the services offered via the EOSC Marketplace portal, at the URL: <https://marketplace.eosc-portal.eu/services/materials-cloud-archive>

## Deliverable D5.6

Report on the data stewardship policies and on the scalability and features of the Materials Cloud Archive



**Fig. 8:** Materials Cloud Archive is recommended as a repository to host data in the field of Materials Science both by Nature's journal Scientific Data and by the EC commission's new journal Open Research Europe. Screenshots from <https://www.nature.com/sdata/policies/repositories#materials> (left) and <https://open-research-europe.ec.europa.eu/for-authors/data-guidelines> (right).

## 4 Conclusions

Materials Cloud Archive is now an established platform for hosting, publishing, sharing and preserving open research data associated with simulations on materials and their properties. It is open to any researcher in the world to host data needed to reproduce simulations, and it is already recommended by two international peer-reviewed journals (Nature's Scientific Data and the EU Commission's Open Research Europe). In addition, it is indexed by the EOSC/EUDAT B2FIND service and by Google Dataset Search, as well as FAIRSHARING.org and re3data.

Based on the robust invenio v3 platform developed at CERN (powering, among others, the Zenodo repository), scalability to millions of entries is guaranteed. Its adoption in the community of materials research is steadily increasing, with a 2-3x increase of submissions of relevant data entries from researchers in the community every year. This demonstrates the need that the community had for such a service now provided by MaX, and that Materials Cloud Archive is an asset in enabling FAIR and reproducible research, open research data and open science.