FAIRness assessment of Materials Cloud

Leonardo Guerreiro Azevedo, Julio Tesolin, Gabriel Banaggia
IBM Research – Brazil
lga@br.ibm.com, {julio.tesolin,gbanaggia}@ibm.com

1 FAIRness of Materials Cloud

This section presents the answers for our proposal of FAIRness questionnaire (version 23.02.2023) considering the responses of the Material Cloud's for the GO FAIR FIP (FAIR Implementation Profile [2]) presented in the Materilas Cloud's FIP [3]. Materials Cloud provides an ecosystem that supports researchers throughout the life cycle of a scientific project, and helps them make their researcher output FAIR and reproducible [4]. It used Invenio framework [1] to build its data repository as presented in some of the answers to the questionnaire. The answers are presented in Table 1, Table 2, Table 3, Table 4, Table 5, Table 6, and Table 7.

The Materials Cloud's FIP used mechanisms to answer question A1.1 that are not protocols. Because of that, we created two questions for A1 to consider that information. However, after analyzing the answers and principles, we understand that Accessibility is more related to the protocol than to retrievability which should be considered in F4. Hence, we moved these two questions to F4 in the next version of the questionnaire, used to assess PubChem repository.

The related questions and answers are:

- Which is the standardized mechanism or service used to provision the metadata?

  Answer: Open Archives Initiative Protocol for Metadata Harvesting, used for DOI harvesting.

- Which is the standardized mechanism or service used to provision the data?

  Answer:

  - Optimade API, which allows for interoperable exchange of atomic structure data between databases.
  - AiiDA REST API, which Allows querying AiiDA provenance graphs stored in AiiDA archives.

Table 1: Questionnaire answers for the context of the FAIRness assessment of Materials Cloud.

Q.ID	Questions
Q1	What is your community? Materials Cloud [4].
Q2	Which digital object will be evaluated in this assessment? This assessment is answered using the Materials Cloud FIP [3] which does not make explicit which digital object is assessed.

References

[1] CERN. Invenio digital library framework. https://invenio.readthedocs.io/en/maint-3.1/general/introduction.html, 2015. Accessed in 2023–02-15.

[2] GO FAIR Initiative. Fair implementation profile (fip) mini-questionnaire. https://bit.ly/yourFIP. Accessed in 2023–02-02.

[3] Valeria Granata. Materials cloud fip. https://archive.materialscloud.org/static/documents/fip/Materials%20Cloud%20V1.1.pdf, 2020.

Table 2: Questionnaire answers to assess F1 considering the Materials Cloud's FIP [3].

Principle	Questions
F1	What is the attribute that identifies the data (*i.e.*, data is understood as any digital object)? Use of Universally Unique Identifier generated by Invenio 3 framework. It is intended to move to Handle System planned for proper GUPRIs (Globally Unique, Persistent and Resolvable Identifiers) for datasets. .
	Is the data identifier (ID) globally unique or is it only unique in the database domain or for a specific context? Currently, it is used a Universally Unique Identifier generated by Invenio 3 framework which is unique in the repository. However, they plan to use a Handle System in the future for handling GUPRIs.
	Is the data ID persistent? It is not possible to answer this question based on the Materials Cloud's FIP.
	Is the data ID resolvable, *e.g.*, to a landing page? It is not possible to answer this question based on the Materials Cloud's FIP.
	Are there metadata describing the data? Yes, there are.
	Do the metadata have a distinct ID from the data? Yes, data and metadata have distinct identifiers. *If the data and metadata have distinct IDs*:
	- What is the identifier of the metadata? Digital Object Identifier (DOI).
	- Is the metadata ID globally unique or is it only unique in the database domain or for a specific context? It is globally unique.
	- Is the metadata ID persistent? Yes, it is persistent.
	- Is the metadata ID resolvable, *e.g.*, to a landing page? Yes, it is resolvable to a landing page.

[4] Leopold Talirz, Snehal Kumbhar, Elsa Passaro, Aliaksandr V. Yakutovich, Valeria Granata, Fernando Gargiulo, Marco Borelli, Martin Uhrin, Sebastiaan P. Huber, Spyros Zoupanos, Carl S. Adorf, Casper Welzel Andersen, Ole Schütt, Carlo A. Pignedoli, Daniele Passerone, Joost VandeVondele, Thomas C. Schulthess, Berend Smit, Giovanni Pizzi, and Nicola Marzari. Materials Cloud, a platform for open computational science. *Scientific Data*, 7(1):299, September 2020. Number: 1 Publisher: Nature Publishing Group.

[5] wikidata. Open access. `https://www.wikidata.org/wiki/Q232932`. Accessed in 2023–02-28.

Table 3: Questionnaire answers to assess F2, F3 and F4 considering the Materials Cloud's FIP [3].

Principle	Questions
F2	Which metadata schemas, if any, are used to describe the data? Dublin Core (http://www.wikidata.org/entity/Q624610), Schema.org (http://www.wikidata.org/entity/Q3475322), DataCite (http://www.wikidata.org/entity/Q821542)
	What kinds of metadata (*e.g.*, descriptive, administrative and structural) are used to describe the data? Not answered considering the FIP.
	Which of these metadata schemas are domain specific and which are domain-agnostic? Not answered considering the FIP.
F3	What is the technology that links metadata to the data (and vice-versa)? Built-in solution of invenio 3 framework. Not yet linked via machine-actionable API from the metadata GUPRI. Plan to introduce link via "index" metadata of Handle System (http://www.wikidata.org/entity/Q3126718).
	How are the metadata and data linked? Materials Cloud's FIP presents that a built-in solution of invenio 3 framework is used, and it does not explains how this solution works.
F4	Which technology is used to make metadata available (or indexed)? The following technologies are used to index metadata: - B2FIND, which is specific for research data. - Google Dataset Search, which offers a great integration with JSON-LD and is easy to setup. - Google, which requires little to no extra effort needed for indexing.
	How is the metadata available or indexed (*e.g.*, in a search engine, as a static web page, in a database, through an API call)? Not answered considering the FIP.
	Which technology is used to make your data available (or indexed)? optimade.science, which allows deep searches of atomic structures from optimade-enabled materials databases.
	How is the data available or indexed (*e.g.*, in a search engine, as a static web page, in a database, through an API call)? Not answered considering the FIP.

Table 4: Questionnaire answers to assess Accessibility considering the Materials Cloud's FIP [3].

Principle	Questions
A1	Which is the standardized mechanism or service used to provision the metadata? Open Archives Initiative Protocol for Metadata Harvesting, used for DOI harvesting.
	Which is the standardized mechanism or service used to provision the data? The following services are used to provision the data: - Optimade API, which allows for interoperable exchange of atomic structure data between databases. - AiiDA REST API, which Allows querying AiiDA provenance graphs stored in AiiDA archives.
A1.1	Which standardized communication protocols are used to access the metadata? HTTPS
	Is the protocol used to access the metadata open, free, and universally implementable? Yes, it is.
	Which standardized communication protocols are used to access the data? HTTPS
	Is the protocol used to access the data open, free, and universally implementable? Yes, it is.
A1.2	What are the security mechanisms used for metadata access, such as ones used for authentication and authorization, and access conditions and access levels? Open access [5]
	What are the security mechanisms used for data access, such as ones used for authentication and authorization, and access conditions and access levels? Open access [5]
	What security information is provided in the metadata that allows one to access the data manually or through a client application? Not answered considering the FIP.
A2	Are data and metadata independently stored? Yes, they are.
	What is the metadata longevity plan? The longevity plan is the Materials Cloud metadata longevity plan. It covers longevity of both metadata and data.
	What is the data longevity plan, if any? The longevity plan is the Materials Cloud metadata longevity plan. It covers longevity of both metadata and data.

Table 5: Questionnaire answers to assess Interoperability I1 considering the Materials Cloud's FIP [3].

Principle	Questions
I1	What is the knowledge representation used for metadata? *E.g.*, Relational, Document, Key Value, Graph, Object, Hierarchical, Network. Not answered considering the FIP.
	Is the knowledge representation used for metadata formal, accessible, shared, and broadly applicable? Not answered considering the FIP.
	In what format the knowledge representation used for metadata is provided? *E.g.*, eXtensible Markup Language (XML), Turtle (TTL), JSON, JSON-LD, CSV, BLOB, CLOB. The Materials Cloud FIP does not presents the knowledge representation language. They state the data format they use for data representation, which are: - JSON-LD - JSON, internal representation used by the invenio 3 framework - eXtensible Markup Language, which is used for representation of dublin-core and oai-pmh metadata.
	Is the format used for knowledge representation of metadata formal, accessible, shared, and broadly applicable? Not answered considering the FIP.
	What is the knowledge representation used for data? *E.g.*, Relational, Document, Key Value, Graph, Object, Hierarchical, Network. Not answered considering the FIP.
	Is the knowledge representation used for data formal, accessible, shared, and broadly applicable? Not answered considering the FIP.
	In what format the knowledge representation used for data is provided? *E.g.*, eXtensible Markup Language (XML), Turtle (TTL), JSON, JSON-LD, CSV, BLOB, CLOB. Not answered considering the FIP.
	Is the format used for knowledge representation of metadata formal, accessible, shared, and broadly applicable? Not answered considering the FIP.

Table 6: Questionnaire answers to assess Interoperability I2 and I3 considering the Materials Cloud's FIP [3].

Principle	Questions
I2	Which structured vocabularies are used to annotate the metadata? The following structured vocabularies are used: schema.org, Dublin Core, and DataCite.
	Are the used vocabularies for metadata FAIR? Not answered considering the FIP.
	Which structured vocabularies are used to encode the data? AiiDA Ontology for provenance graphs which is applied only to AiiDA data records.
	Are the used vocabularies for data FAIR? Not answered considering the FIP.
I3	Which qualified references the data include to other data? Materials Cloud Archive Record Schema v1.0.0, which is derived from invenio's record schema.
	Which qualified references the metadata include to other metadata? AiiDA archive format, which includes extensive provenance tracking for full reproducibility. However it is not available for all datasets.

Table 7: Questionnaire answers to assess Reusability considering the Materials Cloud's FIP [3].

Principle	Questions
R1	What are the relevant metadata attributes? Not answered considering the FIP.
	What are the relevant data attributes? Not answered considering the FIP.
	What is the required accuracy of each attribute, if any? Not answered considering the FIP.
R1.1	Which usage license is used for your metadata? `https://creativecommons.org/licenses/by-sa/4.0/`, Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). Results from data mining of the repository should be distributed under an open license as well.
	Is the metadata usage license clear? Yes, it is.
	Is the metadata usage license accessible? Yes, it is.
	Which usage license is used for your data? `https://spdx.org/licenses/Apache-2.0.html` and several others listed in the FIP.
	Is the data usage license clear? Yes, it is.
	Is the data usage license accessible? Yes, it is.

Table 8: Questionnaire answers to assess Reusability considering the Materials Cloud's FIP [3].

R1.2	Which metadata schemas do you use for describing provenance of the metadata? Materials Cloud Archive Record Schema v1.0.0, a Built-in solution of invenio 3.
	Which metadata schemas do you use for describing provenance of the data? AiiDA archive format, which includes extensive provenance tracking for full reproducibility. It is not available for all datasets.
	What are the attributes used for data provenance? Not answered considering the FIP.
R1.3	What are the domain-relevant community standards for metadata? Not answered considering the FIP.
	Do the metadata under assessment meet these domain-relevant community standards? Not answered considering the FIP.
	What are the domain-relevant community standards for data? Not answered considering the FIP.
	Do the data under assessment meet these domain-relevant community standards? Not answered considering the FIP.