The Journal of Open Source Software

# FAIRLinked: Data FAIRification Tools for Materials Data Science

**Van D. Tran [1,3], Ritika Lamba [2,3], Brandon Lee[2,3], Henry Dirks[3,4], Quynh D. Tran [1,3], Balashanmuga Priyan Rajamohan [2,3], Ozan Dernek [1,3], Laura S. Bruckman [1,3], Yinghui Wu [2,3], Erika I. Barcelos [1,3], and Roger H. French [1,2,3]¶**

**1** Department of Materials Science and Engineering, Case Western Reserve University, Cleveland, OH, 44106, USA **2** Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH, 44106, USA **3** Materials Data Science for Stockpile Stewardship Center of Excellence, Cleveland, OH, 44106, USA **4** Department of Physics, Case Western Reserve University, Cleveland, OH, 44106, USA ¶ Corresponding author

**DOI:** 10.xxxxxx/draft

**Software**

- Review
- Repository
- Archive

**Editor:** Mojtaba Barzegari

**Reviewers:**

- @nsunami
- @exaexa
- @Manishms18

**Submitted:** 22 September 2025
**Published:** unpublished

**License**
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

## Summary

`FAIRLinked` is a software package created to support the FAIRification of materials science data, ensuring proper alignment with FAIR principles: Findable, Accessible, Interoperable, and Reusable (Wilkinson et al., 2016). It is built to be compatible with MDS-Onto, an ontology designed to capture the semantics of various types of materials data, enabling integration and sharing across different research workflows (Rajamohan et al., 2025). The package is subdivided into three subpackages: `InterfaceMDS`, `RDFTableConversion`, and `QBWorkflow`. The first subpackage, `InterfaceMDS` allows users to search for terms using either string search or various filters, explore different domains and subdomains, and add terms to MDS-Onto. `RDFTableConversion` is used for serialization and deserialization of data from CSV into JSON-LDs and vice versa in a way that capture the semantics of the data using MDS-Onto. Lastly, `QBWorkflow` is a serialization and deserialization workflow that incorporates RDF Data Cube vocabulary, useful for working with multidimensional datasets. By offering these packages, `FAIRLinked` lowers the barrier of creating FAIR, machine-actionable data for researchers in the materials science community.

## Statement of Need

Modern materials science research generates data from a wide range of experimental techniques (e.g. synchrotron X-ray diffraction, IV measurements, Suns-Voc, pyrometry, spectroscopy, degradation measurements) spanning multiple application domains like photovoltaics, advanced manufacturing, and electronic components. These experiments produce measurements of various material properties under a multitude of environmental conditions.

The heterogeneity of these data sources introduces the well-known "3V" challenges of big data: volume, velocity, and variety (Laney, 2001). Materials science datasets can also be multimodal, consisting of numerical tables, images, time-series measurements, and other formats. Additionally, different research groups often use inconsistent terminologies, abbreviations, or naming conventions for the same quantities, instruments, or experimental procedures. This inconsistency creates substantial barriers to integrating datasets across laboratories and domains, thereby reducing interoperability and increasing the effort required to reuse data (Dernek et al., 2025; Tran et al., 2025).

To minimize the effort required to process and reuse historical materials data, these datasets must be machine-actionable. The FAIR principles, which stands for Findable, Accessible, Interoperable, and Reusable, offer a widely recognized framework for achieving this objective (Brinson et al., 2024; Hernandez et al., 2024; Huerta et al., 2023; Scheffler et al., 2022). Rather than prescribing specific technical standards, these principles define the qualities a dataset should possess to minimize human intervention and enable automated processing. One widely adopted approach to realize FAIR is through the Resource Description Framework (RDF), which represents knowledge as subject–predicate–object triples within a graph structure (Allenmang et al., 2020). RDF facilitates semantic interoperability by linking data to shared vocabularies and ontologies, enabling better integration and reuse across diverse experimental sources and terminological variations.

There exists a notable lack of dedicated software packages designed specifically to support materials research scientists in FAIRifying their data according to these guidelines. There are multiple tools designed to transform tabular data into linked data like `Virtuoso` (*Virtuoso*, 2025) and `morph-kgc` (Arenas-Guerrero et al., 2024), but they require using R2RML Mapping Language (Das et al., 2012). This mapping language requires a good understanding of both relational databases and RDF data model, making it difficult for many materials researchers to use these packages. `FAIRLinked` is created to be a dedicated simple package that enables both lightweight and RDF Data Cube-based FAIRification in materials data science by providing practical workflows and tools that transform terminologically inconsistent materials data into RDF-based, machine-actionable formats fully compliant with the FAIR principles.

# Materials Data Science Ontology (MDS-Onto)

The Materials Data Science Ontology (MDS-Onto) was developed to support the FAIRification of materials science data by providing consistent vocabularies and abbreviations for a wide range of experimental contexts (Rajamohan et al., 2025). Materials science research produces data from diverse facilities, experimental techniques, and analysis workflows, resulting in highly variable vocabulary and inconsistent terminology. Differences in naming conventions and the omission of critical metadata, such as instrument details, pose challenges for data sharing and reuse. MDS-Onto addresses these issues by providing a standardized semantic framework that improves clarity, ensures contextual completeness of shared datasets, and facilitates interoperability across research groups. This common data model advances the goal of machine-actionable materials data science.

Terms in MDS-Onto are categorized using three attributes: domain, subdomain, and study stage. Domains and subdomains categorize different types of data within materials science, while study stages represent generic procedural steps in a study protocol. By embedding ontology terms with these attributes, MDS-Onto enables targeted term retrieval, allowing users to filter vocabulary based on research needs. For instance, a researcher focusing on photovoltaic cells can easily access only the terms tagged with the "PV-Cell" subdomain. This structured organization improves discoverability and ensures researchers can quickly identify the most relevant vocabulary for their work.

# Key Features

The `FAIRLinked` package comprises of three subpackages: `InterfaceMDS`, `RDFTableConversion`, and `QBWorkflow`, each addressing distinct aspects of FAIRification based on MDS-Onto.

- `InterfaceMDS`: Searching, filtering and adding terms to MDS-Onto.
- `RDFTableConversion`: Create FAIR Linked data (JSON-LD).
- `QBWorkflow`: Create FAIR Linked data compliant with the RDF Data Cube Vocabulary.

DRAFT

## Interfacing with MDS-Onto (InterfaceMDS)

The `InterfaceMDS` subpackage streamlines access to the large MDS-Onto by providing functions for:

- Retrieving the latest version of MDS-Onto,
- Searching ontology terms by string,
- Filtering terms by domain,
- Listing available domains and subdomains, and
- Adding new terms to a local ontology file.

These features make it easier for users to explore and discover relevant vocabulary without manually inspecting the ontology file.

## FAIRLinked Core Workflow (RDFTableConversion)

The `RDFTableConversion` subpackage implements the core FAIRification workflow by guiding users through metadata template preparation, converting tabular datasets into JSON-LD, and enabling deserialization back into CSVs with relevant metadata as shown in Figure 3. Each row of a CSV is transformed into an individual JSON-LD file with unique names created based on the study stages present in the data. Within these JSON-LDs, data are also linked with standardized QUDT units (*QUDT*, 2022) and ontology-backed terminology and definition. The workflow also supports iterative updates, allowing researchers to update JSON-LDs with new data obtained from analysis. Compared to the more complex RDF Data Cube approach, this provides a simpler path to making datasets FAIR and reusable but does not provide as much statistical contexts as `QBWorkflow`.

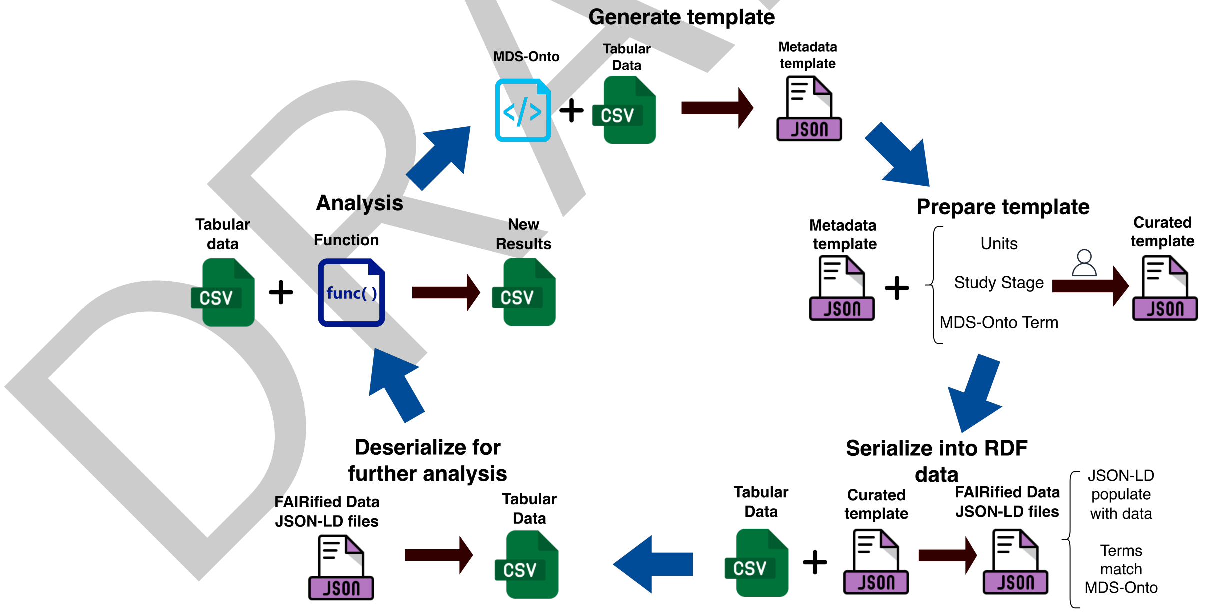


**Figure 1:** FAIRification Workflow for materials science data, which includes four steps: metadata template generation, conversion to ontology-compliant JSON-LD files, deserialization back to CSV, and iterative data analysis and update.

## RDF Data Cube Workflow (QBWorkflow)

For users who wish to add richer metadata to their dataset, FAIRLinked provides the `QBWorkflow` subpackage which utilizes RDF Data Cube vocabulary to capture the structure of multidimensional data into a linked data format (*RDF Data Cube Vocabulary | DCC*, 2024).

The main advantage of `QBWorkflow` over `RDFTableConversion` is allowing users to declare whether a variable is a *dimension* or *measure*, which gives specific statistical contexts defined by Statistical and Metadata Exchange (SDMX) standards (*SDMX User Guides | SDMX – Statistical Data and Metadata eXchange*, 2025). Through interactive guidance, `QBWorkflow` prompts users for the necessary metadata, generates an Excel template as detailed in Table 1 and shown in Figure 4 to help users structure the data to fit the RDF Data Cube vocabulary, and then converts Excel template into JSON-LD files. These files can be turned into CSV, Apache Arrow, or Parquet files for further analysis using `QBWorkflow`. This workflow ensures complex, high-dimensional datasets are properly annotated with semantically interoperable and machine-readable units and statistical contexts using QUDT and RDF Data Cube vocabularies.

**Table 1:** User metadata requirements for the QBWorkflow RDF Data Cube template as shown in Figure 4. Terms have been abbreviated to fit page margins. Here 'unit:UNL' stands for 'unit:UNITLESS'

| **Study Stage** | | Result | Result | Result | Recipe | Recipe | Recipe | Recipe |
|---|---|---|---|---|---|---|---|---|
| **Alt. Label** | ... | ... | ... | ... | ... | ... | ... | ... |
| **Unit (QUDT)** | unit:UNL | unit:UNL | unit:UNL | unit:UNL | ??? | ??? | ??? | ??? |
| **Is Measure?** | NO | NO | NO | NO | YES | YES | YES | YES |
| **Existing URI** | mds:ExpId | ____ | ____ | ____ | ____ | ____ | mds:Eng | ____ |
| **Variables** | **ExpId** | **LogFile** | **ImgSeq** | **StkNum** | **DetLen** | **Wave** | **Energy** | **PixSize** |

# Appendix

| | Sample | chemical_formula | processing_method | sample_dept | sample_id | sample_lengt | sample_width |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | EBM3 | Ti6Al4V | Electron Beam Melting | 0.8 | Ti64-EBM3 | 3.78 | 2.81 |
| | PB1 | Ti6Al4V | Laser Powder Bed Fusion | 0.83 | Ti64-PB1 | 3.81 | 2.82 |
| | CeO2 | Ceria calibrant | | | CeO2 | | |
| | EBM1 | Ti6Al4V | Electron Beam Melting | 0.87 | Ti64-EBM1 | 3.97 | 2.86 |
| | LENS1 | Ti6Al4V | Laser Engineered Net Shaping | 0.8 | Ti64-LENS1 | 3.74 | 2.78 |
| | LENS2 | Ti6Al4V | Laser Engineered Net Shaping | 0.8 | Ti64-LENS2 | 3.76 | 2.79 |
| | PB2 | Ti6Al4V | Laser Powder Bed Fusion | 0.77 | Ti64-PB2 | 3.76 | 2.74 |
| | EBM2 | Ti6Al4V | Electron Beam Melting | 0.87 | Ti64-EBM2 | 3.88 | 2.86 |
| | LENS2 | Ti6Al4V | Laser Engineered Net Shaping | 0.8 | Ti64-LENS2 | 3.76 | 2.79 |

**Figure 2:** Minimal CSV required for `RDFTableConversion`. Users should include 3 blank rows to make room for column metadata.

| __Label__ | Sample | chemical_formula | processing_method | sample_depth | sample_id | sample_length | sample_width | __rowkey__ |
|---|---|---|---|---|---|---|---|---|
| Type | mds:XraySample | mds:ChemicalFormula | mds:ProcessingMethod | cco:ont00000607 | mds:SampleID | cco:ont00000738 | cco:ont00000324 | |
| Units | {'@id': 'unit:AU'} | {'@id': 'unit:IN'} | {'@id': 'unit:'} | {'@id': 'unit:IN'} | {'@id': 'unit:'} | {'@id': 'unit:IN'} | {'@id': 'unit:IN'} | |
| Study Stage | Recipe | Sample | Recipe | Recipe | Sample | Formulation | Modeling | |
| 1 | EBM3 | Ti6Al4V | Electron Beam Melting | 0.8 | Ti64-EBM3 | 3.78 | 2.81 | REC217489_SA383167_FOR798209_MOD652672 |
| 2 | PB1 | Ti6Al4V | Laser Powder Bed Fusion | 0.83 | Ti64-PB1 | 3.81 | 2.82 | REC621755_SA384821_FOR182060_MOD751747 |
| 3 | CeO2 | Ceria calibrant | | | CeO2 | | | REC695244_SA884581_FOR386549_MOD386549 |
| 4 | EBM1 | Ti6Al4V | Electron Beam Melting | 0.87 | Ti64-EBM1 | 3.97 | 2.86 | REC808835_SA805384_FOR159059_MOD223534 |
| 5 | LENS1 | Ti6Al4V | Laser Engineered Net Shaping | 0.8 | Ti64-LENS1 | 3.74 | 2.78 | REC187098_SA754648_FOR720915_MOD162634 |
| 6 | LENS2 | Ti6Al4V | Laser Engineered Net Shaping | 0.8 | Ti64-LENS2 | 3.76 | 2.79 | REC257872_SA789817_FOR696580_MOD716881 |
| 7 | PB2 | Ti6Al4V | Laser Powder Bed Fusion | 0.77 | Ti64-PB2 | 3.76 | 2.74 | REC852382_SA827046_FOR696580_MOD689951 |
| 8 | EBM2 | Ti6Al4V | Electron Beam Melting | 0.87 | Ti64-EBM2 | 3.88 | 2.86 | REC847839_SA130218_FOR190488_MOD223534 |
| 9 | LENS2 | Ti6Al4V | Laser Engineered Net Shaping | 0.8 | Ti64-LENS2 | 3.76 | 2.79 | REC257872_SA789817_FOR696580_MOD716881 |

**Figure 3:** Deserialized CSV with metadata included.

| | If you don't mention an ExperimentId, IDs will be generated for each experiment | Result | | | Recipe | | | |
|---|---|---|---|---|---|---|---|---|
| Alternative Label - write mappings to the columns that you use in your dataset | | | | | | | | |
| Indicate what unit you are using to qualify your measure. Refer to https://www.qudt.org/doc/DOC_VOCAB-UNITS.html | | | | | | | | |
| If Measure indicate YES, else (dimension) indicate NO | | | | | | | | |
| If you already have a URI for this particular column, indicate it like so mds:InstrumentId. Ensure mds is defined in the other sheet | | | | | | | | |
| | ExperimentId | LogFile | ImageSequence | StackNumber | DetectorLength | Wavelength | Energy | PixelSize |
| | 1 | | | | | | | |
| | 2 | | | | | | | |

**Figure 4:** RDF Data Cube Template for FAIRification using RDF Data Cube. Users can fill out the required metadata for correct serialization into RDF Data Cube JSON-LDs.

| Namespace you are using | Base URI |
|---|---|
| rdf | http://www.w3.org/1999/02/22-rdf-syntax-ns# |
| rdfs | http://www.w3.org/2000/01/rdf-schema# |
| owl | http://www.w3.org/2002/07/owl# |
| xsd | http://www.w3.org/2001/XMLSchema# |
| skos | http://www.w3.org/2004/02/skos/core# |
| void | http://rdfs.org/ns/void# |
| dct | http://purl.org/dc/terms/ |
| foaf | http://xmlns.com/foaf/0.1/ |
| org | http://www.w3.org/ns/org# |
| admingeo | http://data.ordnancesurvey.co.uk/ontology/admingeo/ |
| interval | http://reference.data.gov.uk/def/intervals/ |
| qb | http://purl.org/linked-data/cube# |

**Figure 5:** User provides the namespaces used in the Namespace Template

# Code Availability

The source code for `FAIRLinked` can be retrieved from PyPi or in the GitHub repository.

# Acknowledgement

This work was supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Solar Energy Technologies Office (SETO) through Agreement Numbers DE-EE0009353 and DE-EE0009347. Additional support was provided by the Department of Energy (National Nuclear Security Administration) under Award Number DE-NA0004104 and Contract Number B647887 and from the U.S. National Science Foundation under Award Number 2133576.

The authors would like to sincerely thank these organizations for their financial assistance as well as all of the individuals who participated in the project.

DRAFT

## References


Allenmang, D., Hendler, J., & Gandon, F. (2020). *Semantic Web for the Working Ontologist: Effective Modeling for Linked Data, RDFS, and OWL* (3rd ed.). Association for Computing Machinery. ISBN: 978-1-4503-7614-3

Arenas-Guerrero, J., Chaves-Fraga, D., Toledo, J., Pérez, M. S., & Corcho, O. (2024). Morph-KGC: Scalable knowledge graph materialization with mapping partitions. *Semantic Web*, *15*(1), 1–20. https://doi.org/10.3233/SW-223135

Brinson, L. C., Bartolo, L. M., Blaiszik, B., Elbert, D., Foster, I., Strachan, A., & Voorhees, P. W. (2024). Community action on FAIR data will fuel a revolution in materials research. *MRS Bulletin*, *49*(1), 12–16. https://doi.org/10.1557/s43577-023-00498-4

Das, S., Sundara, S., & Cyganiak, R. (2012). *R2RML: RDB to RDF Mapping Language*. https://www.w3.org/TR/r2rml/

Dernek, O., Mehdi, R., Yue, W., Bachman, J. A., Holt, F. R., Ponon, G. O., Tripathi, P. K., Willard, M. A., Ernst, F., & French, R. H. (2025). A Perspective on Synchrotron Data Science. *Interdisciplinary Information Sciences*, *31*(1), 83–88. https://doi.org/10.4036/iis.2025.A.08

Hernandez, K. J., Barcelos, E. I., Jimenez, J. C., Nihar, A., Tripathi, P. K., Giera, B., French, R. H., & Bruckman, L. S. (2024). A data integration framework of additive manufacturing based on FAIR principles. *MRS Advances*, *9*(10), 844–851. https://doi.org/10.1557/s43580-024-00874-5

Huerta, E. A., Blaiszik, B., Brinson, L. C., Bouchard, K. E., Diaz, D., Doglioni, C., Duarte, J. M., Emani, M., Foster, I., Fox, G., Harris, P., Heinrich, L., Jha, S., Katz, D. S., Kindratenko, V., Kirkpatrick, C. R., Lassila-Perini, K., Madduri, R. K., Neubauer, M. S., … Zhu, R. (2023). FAIR for AI: An interdisciplinary and international community building perspective. *Scientific Data*, *10*(1), 487. https://doi.org/10.1038/s41597-023-02298-6

Laney, D. (2001). *3D data management: Controlling data volume, velocity and variety*. META group research note.

*QUDT*. (2022). https://www.qudt.org/

Rajamohan, B. P., Bradley, A. C. H., Tran, V. D., Gordon, J. E., Caldwell, H. W., Mehdi, R., Ponon, G., Tran, Q. D., Dernek, O., Kaltenbaugh, J., Pierce, B. G., Wieser, R., Yue, W., Lin, K., Kambo, J., Lopez, C., Nihar, A., Savage, D. J., Brown, D. W., … French, R. H. (2025). Materials Data Science Ontology(MDS-Onto): Unifying Domain Knowledge in Materials and Applied Data Science. *Scientific Data*, *12*, 628. https://doi.org/10.1038/s41597-025-04938-5

*RDF Data Cube Vocabulary | DCC*. (2024). https://www.dcc.ac.uk/resources/metadata-standards/rdf-data-cube-vocabulary

Scheffler, M., Aeschlimann, M., Albrecht, M., Bereau, T., Bungartz, H.-J., Felser, C., Greiner, M., Groß, A., Koch, C. T., Kremer, K., Nagel, W. E., Scheidgen, M., Wöll, C., & Draxl, C. (2022). FAIR data enabling new horizons for materials research. *Nature*, *604*(7907), 635–642. https://doi.org/10.1038/s41586-022-04501-x

*SDMX User Guides | SDMX – Statistical Data and Metadata eXchange*. (2025). https://sdmx.org/user-guide/

Tran, Q. D., Barcelos, E. I., & Bruckman, L. S. (2025). Designing Data-Centric Study Protocols Guided by FAIR Principles. *Interdisciplinary Information Sciences*, *31*(1), 75–82. https://doi.org/10.4036/iis.2025.A.07

*Virtuoso*. (2025). OpenLink Software. https://vos.openlinksw.com/owiki/wiki/VOS


DRAFT

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1). https://doi.org/10.1038/sdata.2016.18

DRAFT