Harmonised terminologies and schemas for FAIR data in materials science and related domains WG - Case Statement

—----------

Table of Contents

1. Charter

The proposed Working Group on "Harmonised terminologies and schemas for FAIR data in materials science and related domains" recognises the wide range of activities in several countries and world regions aiming to achieve FAIR data, support semantic data management and interoperability in the materials sciences and related domains. While the main focus of the WG is in the material sciences, close interactions with cognate domains, in particular chemistry, are crucial in order to avoid conflicting approaches and also to utilise and integrate with already existing semantic artefacts and resources in these fields. The WG will review the current landscape in the context of good practice and FAIR maturity and elaborate recommendations and guidance for practitioners working in the materials sciences and related domains.

Objectives of the Working Group:

- Collect and review existing vocabularies/terminologies, schemas, etc. for the material sciences and cognate domains such as chemistry;
- Propose and verify best practices for materials science practitioners to improve FAIR data based on terminologies, schema, ontologies, etc.;
- Identify needs for harmonisation of terminologies and schema in the material sciences;
- Demonstrate successful implementation of FAIR terminologies, schema and ontologies in specific use cases.

The WG will focus on increasing uptake of the FAIR Principles in chemical and materials research (in particular in connection with Interoperability and Reusability), supported by improved resources (having improved FAIR maturity), and more widely-shared metadata and ontologies. The WG objectives are also closely aligned with (and support) the EOSC interoperability framework concerning its semantic layer. In particular, it addresses the problems and needs in Section 3.2.1 of the EOSC Interoperability Framework Document, namely, "lack of common explicit definitions about the terms that are used by user communities", and lack of "common semantic artefacts across communities (e.g., general ontologies that can be shared)", etc.[1] The WG will contribute to filling the Semantic Artefact, Ontology, Terminology, Controlled Vocabulary blocks of Figure 7. in the EOSC Interoperability Framework in the chemical and materials sciences.

Subsequently, the objective is to increase FAIRness in the materials sciences. While all metadata-related FAIR maturity indicators are in scope, the WG will specifically address the following:

- Interoperability maturity Indicators, in particular:
    - RDA-I1-01M "Metadata uses knowledge representation expressed in standardised format".
    - RDA-I2-01M "Metadata uses FAIR-compliant vocabularies"
- Reusable maturity Indicators, in particular:
    - RDA-R1-01M "Plurality of accurate and relevant attributes are provided to allow reuse." As above, metadata, where provided, lack rich description to allow re-use, and also lack a common approach across the different materials science communities to do that.

[1] https://data.europa.eu/doi/10.2777/620649

- RDA-R1.3-01M Metadata complies with a community standard

The WG will review the FAIR maturity of existing semantic artefacts, provide recommendations regarding best practices applied to the domains of interest, and elaborate "harmonised" common semantic artefacts across the communities covered by this Working Group.

In particular, the Working Group will provide the following outputs:

A. A review and FAIR maturity assessment of the existing semantic artefacts landscape (terminologies, taxonomies, schema, ontologies) for materials science and related domains; producing a documentation of the process of FAIR maturity review, application and elaboration of semantic resources for the domain.
B. Best practice recommendations for materials science practitioners to improve FAIR data based on terminologies, schema, ontologies documented in A, including how to apply the FAIR principles within the context of publishing semantic artefacts in the domains of this WG.
C. New, harmonised semantic artefacts (in particular terminologies and schema) required to improve FAIR maturity in the domains of this WG.

2. Value Proposition

Materials scientists in academia or industry are increasingly tasked to apply the FAIR Principles to their data. Materials research and industry, including whole value chains (such as that of battery chemicals, materials, batteries and their application in transport) are in need of coherent, harmonised data documentation to support exchange and re-use. There are a multitude of initiatives and projects at national and international level tasked with digital representations of materials to support the above aims and a vision of creating a 'materials commons' based on FAIR data in materials and related domains.

However, while there are general guidelines for FAIR data, there are various, disparate developments of terminologies, metadata, schema, taxonomies and ontologies in all fields including materials sciences. Currently, there is a lack of controlled vocabulary in the materials science field, in contrast to, e.g., the IUPAC Compendium of Chemical Terminology (the "Gold Book") for Chemistry. There are numerous efforts in national and international initiatives to address this, but they lack alignment and collaboration, risking another 'tower of babel' in the field. While a single WG cannot address all needs, the review, recommendations and instantiation outputs described above are designed to demonstrate a way towards a common approach.

Since the WG specifically includes cognate domains to materials sciences, multi-disciplinary cooperation will strongly be impacted, across chemistry, physics and materials science communities, and also into engineering, as a key re-user community of materials data. Working towards a harmonised and standardised documentation of materials will also be essential to achieve objectives such as materials passports for goods, that provide information across and along value chains, and eventually to the users about materials in the products they use.

3. Engagement With Existing Work in the Area

There are a wide range of pertinent activities in the area of the WG in Europe, Asia and the US. A number of these were presented at Plenary 20 in a BoF session[2] organised by this group and most have joined or are planning to join this WG.
Activities in Chemistry and Materials are sometimes overlapping but often distinct, although their integration is desired to provide interoperability, hence there is strong engagement with the RDA IG Chemistry Data and Materials Data communities.

The WG Leaders and Members are engaged in and supported by major initiatives in Asia, Europe, and the US. These include in particular the EMMC ASBL[3] and related EU projects[4], the National Centre for Materials Research Data[5] in Korea; the Materials Genome Initiative (MGI)[6] and Materials Research Data Alliance (MARDA)[7] in the US, the German initiatives Platform MaterialDigital[8], NFDI[9], HMC (Helmholtz Metadata Collaboration)[10], and NIMS[11] in Japan.

In particular, we note the following activities, which will be engaged in this WG:

- EU Horizon projects including OntoCommons (Ontology-driven data documentation for Industry Commons) CSA project[12] and OntoTrans (Ontology driven Open Translation Environment) project[13], as well as the EMMC ASBL (governing body of the EMMO ontology and its related collaborations[14]). These are contributing to
    - Development and application of the EMMO[15] (Elementary Multiperspective Material Ontology) in a number of European projects, with governance by EMMC (European Materials Modelling Council).
    - OntoCommons Ontology Ecosystem (OCES) development and recommendations, including Top, Middle- and Domain level ontologies pertinent to materials and manufacturing in the European OntoCommons CSA project, with successor projects being planned. The project elaborated on domain ontology harmonisation and developed the so-called bridge concept (template and specific materials bridge concepts) to support ontology alignment.

[2] https://www.rd-alliance.org/data-representation-materials-and-chemicals-based-harmonised-domain-ontologies
[3] https://emmc.eu/focus-areas/digitalisation-interoperability/
[4] https://emmc.eu/emmc-related-initiatives/
[5] https://www.ncmrd-symposium.net/home
[6] https://www.mgi.gov/
[7] https://www.marda-alliance.org/
[8] https://material-digital.de/
[9] https://www.nfdi.de/?lang=en
[10] https://helmholtz-metadaten.de/en
[11] https://www.nims.go.jp/eng/
[12] https://ontocommons.eu/
[13] https://ontotrans.eu/
[14] https://github.com/emmo-repo/EMMO
[15] https://github.com/emmo-repo/EMMO

- German national initiatives and projects including Platform Material DIgital (PMD) and NFDI (in particular NFDI4Chem, NFDI4Cat, NFDI4Ing, NFDI Matwerk and FAIRMat) including efforts on improved vocabularies (see e.g. [16]), and ontologies (e.g. [17])
    - NFDI4Chem Ontologies for Chemistry Workshop[18]
    - NFDI4Chem Terminology Service[19]
    - NFDI4Ing Terminology Service[20]
- The Czech Republic is just starting to build a National Data Infrastructure with dedicated research data repository for the Materials Sciences and Engineering domain. It can function as a test-bed of the recommendations, agreed schema and terminologies (including ontologies). Coordination with other communities will be supported via the current RDA Ambassador for this domain, Marek Cebecauer[21].
- Korea launched the National Center for Materials Research Data (NCMRD) in 2020 to build the materials research data platform as the harmonised data repository in materials science and engineering domain (https://kmds.re.kr/en/). Standard schema and vocabulary for materials research data has been developed in the committee for materials R&D data standard. This schema is characterised by using the concept of "materials system", which can be widely adapted to various materials research data. The homogeneous data schema is expected to enhance the interoperability of the materials research data. The Korean community also makes efforts for classification of materials, a prerequisite for increasing the number of materials systems in the standard data schema.
- In Japan the National Institute for Materials Science (NIMS) has started to build a materials data platform in 2017. Here, the materials database suite, MatNavi (https://mits.nims.go.jp/), is one of the key data resources, and ontologies are being introduced to facilitate interoperability between these databases and external resources. Currently, they have constructed ontologies for the polymer database PoLyInfo and the superconductor database SuperCon, both of which have Basic Formal Ontology as their top concept. In particular, PoLyInfo has a large macromolecular concept inherent in it and is expected to play an important role in machine readability of organic chemistry.
- MaRDA, The Materials Research Data Alliance (MaRDA) is a community-led network focused on connecting and integrating U.S. materials research data infrastructure to realise the promise of open, accessible, and interoperable materials data. MaRDA provides a platform that promotes the convergence of ideas, people, data, and tools to accelerate discovery, enable new insights into materials mechanisms, and lay the foundation for both human-centred and artificial intelligence-assisted approaches to materials design. MaRDA may have a number of working groups aligned with this proposed working group and significant effort will be made to coordinate activities and work together.

[16] Moustakas, Nikolaos, Behr, Alexander, Borgelt, Hendrik, Huskova, Nadiia, Khare, Rachit, Talab, Manal, Köbl, Julia, Chandrashekhar, Vishwas, Petrenko, Taras, Dörr, Mark, & Linke, David. (2023). Voc4cat: Vocabulary guidelines for NFDI4Cat. Zenodo.https://doi.org/10.5281/zenodo.7669183
[17] https://materialdigital.github.io/core-ontology/index-de.html
[18] https://docs.google.com/document/d/1EteBVVmyZEsv9Q55uvk2XjbZcPeSr8BB/edit
[19] https://terminology.nfdi4chem.de/ts/
[20] https://terminology.nfdi4ing.de/ts/
[21] https://www.rd-alliance.org/marek-cebecauer

- Harmonisation (and in particular avoidance of conflicting conceptualisations) between Materials Science and Chemistry through engagement with IUPAC and specifically the continued development of the IUPAC Gold Book (Compendium of Chemical Terminology) which republishes machine-accessible definitions of chemical concepts via IUPAC Recommendations
- Interactions with the CODATA/RDA WorldFAIR Initiative[22] to take their recommendations into account and feed into their work. Several work packages have relevant activities: Chemistry, Nanomaterials, Geochemistry, Cross-Domain Interoperability Framework, and possibly others.
- The work ongoing with the PANET ontology (Photon and Neutron Experimental Techniques ontology), including via the RDA IG Research data needs of the Photon and Neutron Science community.

4. UN Sustainable Development Goals (SDGs)

Relevant UN SDGs include:

#7 Ensure access to affordable, reliable, sustainable and modern energy for all
Superconductivity is one of the materials that can dramatically improve the efficiency of power transmission to improve energy efficiency. NIMS in Japan has a superconductivity database, SuperCon, and has completed the construction of its ontology and has begun to utilise it. In particular, we are collecting descriptors through data linkage for the discovery of superconducting materials by machine learning. The ionic radii and electronegativity of the contained elements are collected from wikidata, and chemical properties such as vapour pressure are collected from PubChem and other sources. This initiative will facilitate the construction of a system that contributes to the SDGs.

#9 Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation
Knowledge sharing is key to building more dynamic and inclusive innovation systems that enable a wider range of actors to participate and new business models to emerge. A crucial ingredient for such systems is a shared knowledge conceptualisation and FAIR data, supported by this WG.

#14 Conserve and sustainably use the oceans, seas and marine resources for sustainable development
Through this initiative, we will present concrete solutions, such as biodegradable plastics, to the recent increasing marine pollution caused by plastics. NIMS in Japan owns the polymer database PoLyInfo, has completed the construction of its ontology, and has begun to utilise it. In fact, by creating a federation ontology, NIMS has already integrated data with NCBI (The National Center for Biotechnology Information) and succeeded in a feasibility study on the prediction of biodegradable plastics and bio-degrading microbes. The knowledge integration using the ontologies to be achieved through this initiative will ensure that we will contribute to solving the international problems presented in the SDGs.

[22] https://worldfair-project.eu/

5. Work Plan

As outlined in Section 1, the WG will produce a set of outputs which will be in the **form of reports and semantic artefacts.**

Recommendation (in the form of a report) will include

1. Documentation of the process of FAIR maturity review, application and elaboration of semantic resources for the domain.
2. Best practice recommendations for materials science practitioners
    a. for the development of FAIR semantic artefacts at application and domain level.
    b. on how to apply the above in order to achieve FAIR data in the domains of this WG.
3. Recommendations for new, harmonised semantic artefacts (in particular terminologies and schema) required to improve FAIR maturity in the domains of this WG.

In addition, the WG will create an instantiation (or instantiations) for the existing and new, harmonised semantic artefacts, for example by creation of a FAIRsharing collection to create a "live view" of semantic artefacts reviewed and elaborated. This will support the WG during its active phase and support the adoption phase.

The work will be organised in Streams, where Streams A and B work largely in parallel to carry out a review and establish current best practice, followed by Stream C on Harmonisation, and finally Stream D on Adoption.

T**he timeline and tasks are outlined in a Gantt chart in the Appendix**. It shows a preparation phase which is already ongoing, the WG core 18-month period, and a subsequent Adoption phase.

Stream A: Collection, review and FAIR maturity assessment of the existing semantic artefacts. First, it will be clarified what we mean by vocabularies, terminologies, taxonomies, schema, and ontologies and how they are applied to improve FAIR maturity. The definitions and guidelines for what constitutes the different types of semantic artefacts will be documented based on existing definitions, but with examples, e.g. schema.org for schema. Following that, the semantic artefacts collection will take place. Our streams will collaborate with the FAIRsharing WG in reviewing and assessment of the existing terminology landscape for this domain, creating a FAIRsharing collection as a "live view" of those terminologies reviewed by this WG. All current repositories of semantic artefacts will be taken into account, including MatPortal[23] and IndustryPortal[24]. Addition of missing domain terminologies will be supported, in particular via the FAIRsharing registry which will be pointed to in the output report. As example of other WG utilising FAIRsharing, see the short introduction (https://fairsharing.gitbook.io/fairsharing/record-sections-and-fields/general-information/registry-type#collections) to FAIRsharing collections.

This stream will produce a documentation of the process of FAIR maturity review, application and elaboration of semantic resources for the domain.

[23] https://matportal.org/
[24] https://industryportal.enit.fr/

Stream B: Best practices for materials science practitioners to achieve FAIR data based on terminologies, schema, ontologies

This stream will review best practices, evaluation and implementations of pertinent existing recommendations and templates including FAIR Implementation Profiles[25] in the context of the domains for this WG. It will review work on harmonisation efforts, user-accessible vocabularies and ontologies. Collection of champion use-cases together with the user experience from diverse stakeholders (researchers, students, data stewards, data curators), feedback to developers of standards from a broader research community, terminology annotation recommendations, review of collaborative tools in use for terminology (including capabilities for annotations, machine conversion to rdf, etc.) and output recommendation for a widely shared materials terminology resource for materials terminologies and taxonomies (in the last part feeding into Stream C).

This stream will produce recommendations for materials science practitioners to achieve FAIR data and how to apply the FAIR principles within the context of publishing semantic artefacts in the domains of this WG.

Stream C: Elaborate harmonised terminologies and schema in materials sciences

A harmonised set of metadata including their terminologies (elucidations, relation to other concepts) for materials and chemistry is required. The latter should be based closely on existing (and evolving) terminologies and metadata such as the IUPAC Gold Book and relevant Schema.org metadata. These should include the basic concepts in materials science and chemistry and cover the most frequent and widely required concepts.

This stream will, via workshops, action plan and community engagement, create a plan and instantiation for such new semantic artefacts (in particular schema) in the domains of this WG. The aim is to clarify confusion and multitude of somewhat different uses of these terms and their relationships and identify gaps in defined concepts. The Stream will elaborate plans and recommendations for building on this catalogue of materials terminology alongside and complementary to the IUPAC Gold Book, as a source of unique metadata for FAIR data annotations.

Stream D: Adoption: Demonstrate FAIR terminologies, schema and ontologies

The recommendations will be disseminated via the WG members and related projects and initiatives (see Adoption Plan below). A Strategic Planning workshop will be held at the end of the 18-month period to plan further specific adoption activities.

Intermediate documents and milestones

These include (see also Gantt Chart in Appendix 1):

- Creation of a resource for shared materials science terminology on FAIRsharing (M3)
- Collection of existing semantic artefacts and resources documented in Report (M3)
- Community Engagement Events targeted at RDA Plenaries (M6, M12, M18)
- Stream A and B progress showcases at M6, M12, M18
- Stream B documentation of findings on existing best practices, for public comments at M3
- Harmonisation Action Plan at M9
- Created new semantic artefacts for public comment (M15)

[25] https://www.go-fair.org/how-to-go-fair/fair-implementation-profile/

WG's mode and frequency of operation

The WG will work in Streams/Work Packages, focussing on particular areas as follows below. Streams will have teams that meet monthly online. WG as a whole meets quarterly, including at RDA Plenaries, which are also a key channel for engagement of the community.
The Work Streams are detailed in the Gantt Chart, which also indicates half-yearly community engagement sessions that are planned to coincide with RDA Plenaries.

How the WG plans to develop consensus, address conflicts, stay on track and within scope, and move forward during operation

The WG will develop consensus based on regular communication, open sharing of all information, and periodic seeking of input on findings and approaches, and through collaboration, inclusiveness and ensuring issues are heard and discussed without prejudice. WG Co-Chairs actively support and all participants commit to upholding the above principles. In particular, participants will be asked to work collaboratively in the spirit of a consensus formation. Where necessary, remaining issues will be documented in a kind of backlog, and discussed, reviewed and hopefully resolved separately, so as not block the operation and progress of the WG.

Approach to broader community engagement and participation

Community engagement will take place as follows:

- Engagement of RDA community, in particular those connected to relevant IGs (Materials, Chemistry Research Data IG, Research Data Management in Engineering IG, Vocabulary Services IG) by organising WG meetings at each of the RDA Plenaries for the duration of the WG. Providing information about how to join and maintaining good documentation and online resources to lower the barrier for people joining and contributing.
- Working with RDA Ambassador for Materials and RDA TIGER programme on dissemination, including webinars reporting on objectives and progress.
- Each of the partners, in particular the main contacts in the different countries and geographic areas will support dissemination and engagement in their networks, as also described in the Adoption plan.

6. Adoption Plan

Plans for adoption/implementation in organizations represented by WG members are as follows:

- EMMC and related projects: EMMC coordinates collaboration across a number of Horizon 2020 and Horizon Europe projects that both develop semantic artefacts and use semantic artefacts for FAIR data. EMMC will organise collaborative workshops (online) and sessions at its biennial International Workshop to support adoption in these communities.
- NIST: A project within NIST has performed some initial exploratory work extending schema.org in much the same way the life sciences community has formed the bioschemas.org project. NIST intends to update all work following the output of this WG.
- NIMS has a large polymer database, PoLyInfo, which has a polymer ontology (https://dice.nims.go.jp/ontology/PoLyInfo-ont/Schema#) and instance-level RDF. We will

redefine this resource in RDF to be consistent with the IUPAC Gold Book as well as the IUPAC Purple Book that defines nomenclature for polymers[26].

- We will investigate the existing schema expression ShEx (Shape expression, http://shex.io/shex-semantics/) and validate the use of ShEx for chemical schema sharing. The related documents will be published in the Materials Data Repository (MDR, https://mdr.nims.go.jp/?locale=en), a repository that aims to follow the FAIR principles. Note that ShEx has been adopted by wikidata and others (https://www.wikidata.org/wiki/Wikidata:WikiProject_Schemas).
  If necessary, the vocabulary will be extracted from the Polymer Dictionary (https://polymer.nims.go.jp/PoLyInfo/guide/en/term_polymer.html#chap03) and shared according to the WG's policy.
- NCMRD of Korea has been developing a harmonised schema and terminology for material research data. The output of this WG will be immediately adapted and implemented in the standard data schema and vocabulary. NCMRD will also collaborate with other organisations/institutions for building global scale materials research data resources by mapping schema and terminologies based on the recommendation of this WG. One example would be the database of perovskite solar cell research resulting from the collaboration between NCMRD (https://kmds.re.kr/en/), Perovskite Solar Cell Data Platform (https://solar.chemdx.org), and Perovskite Database (https://www.perovskitedatabase.com).
- NFDI MatWerk consortium has a designated Task Area (TA) “Ontologies for Material Science” in which semantic artefacts for materials science data are developed in so called “Infrastructure Use Cases” (IUCs). Developments are harmonised via participation of TA-members in the IUCs and regular exchange between those, including common top-level alignment (e.g. with MatWerk ontology[27]) – the semantic backbone of the MSE KG[28]. These are the interfaces in which the RDA WG output will be used for adoption within NFDI MatWerk.
- Helmholtz Metadata Collaboration (HMC) supports a number of Materials Science and Engineering related ontology developments within the Helmholtz Research Fields “Information”, “Energy” and “Matter”. Here representatives of the RDA WG are involved in the development of an overarching strategy for semantic harmonisation. For the relevant areas the RDA-WG output will be taken into account and used for interaction with application-level use cases and adoption with these developments.
- Czech EOSC WG for Materials Sciences and Engineering is building a regional thematic repository which will function as a testbed for outputs produced by this RDA WG. The repository will integrate recommended schemas, vocabularies and ontologies with other services such as FAIRification tools or specific PIDs for physical samples (IGSN) and for instrument instances (PIDINST). User feedback will be collected at each step of the recommendations’ implementation. Diverse stakeholders (e.g., researchers, students, data stewards, data curators, repository administrators) will be providing feedback.

[26] https://iupac.org/what-we-do/books/purplebook/
[27] https://git.rwth-aachen.de/nfdi-matwerk/ta-oms/mwo
[28] https://demo.fiz-karlsruhe.de/matwerk/

Plans for adoption/implementation more broadly in the materials sciences community:

- Via RDA IGs: the WG members have a strong track record of RDA in the above-mentioned IGs and collaboration in the WG on International Materials Resource Registries. The WG also includes the RDA's Domain Ambassadors for Materials Sciences and Engineering and Chemistry-Catalysis fields (Marek Cebecauer and Pedro Mendes, respectively) and coordinates closely with IGs in Chemistry, Materials and Vocabulary to improve coordination and dissemination of efforts (i.e., aligned with RDA Strategic Plan).
- Via EMMC, EMCC communities, European Advanced Materials Initiative (AMI2030) and new European Coordination and Support Action DigiPass (starting April 2024).
- Via Platform Material Digital PMD (Germany), interaction with PMD projects, PMD annual meeting, participation in biennial MSE Congress (2024, 2026) (where PMD projects typically meet).
- Via NIMS (Japan): NIMS has a Repository based on the FAIR Principles and a lexicon sharing dictionary (e.g. https://matvoc.nims.go.jp/explore/ja/results/Q713). These will be used as the sharing platform by NIMS in this project to spread the resource widely. We will also utilise the NIMS public endpoints (https://materials-open-rdf.nims.go.jp/sparql) and expand them to trial visualisation of chemical knowledge (https://materiage.org/).
- Via MaRDA (US): The WG will work closely with MaRDA groups during the operation of the group. After finalisation, the WG will seek opportunities for longer term efforts within MaRDA.
- Via National Centre for Materials Research Data NCMRD (Korea): NCMRD is operating an open materials research data platform (https://www.kmds.re.kr) based on the FAIR principles.
- Via EOSC-Association Task Forces 'Upskilling countries (and domains) to engage in EOSC' and 'Researcher engagement & adoption'.

The WG will explore the possibility of making an application for an RDA TIGER third-party grant in order to support one or more of the above suggested implementations. This will be explored in the second half of the WG lifecycle in coordination with the RDA TIGER Facilitator.

7. Initial Membership

Initial leadership of the Working Group (from Europe, Asia and the United States):
Gerhard Goldbeck (GCL, UK)
Kwang-Ryeol Lee (KIST, Korea)
Zachary Trautt (NIST, US)
Leah McEwen (IUPAC, US)
Masashi Ishii (NIMS, Japan)
Iseult Lynch (University of Birmingham, UK)

Initial Members include the following:
https://www.rd-alliance.org/node/78564/members

Appendix: Working Group Gantt Chart