DOI: 10.5281/zenodo.17949060

Metadata for cataloguing datasets from materials science and related domains

Otello Roscioni and Gerhard Goldbeck, Goldbeck Consulting Ltd, Cambridge, UK

Contents

DOI: 10.5281/zenodo.17949060

Introduction

The work presented has been carried out in relation to the RDA Working Group "Harmonised terminologies and schemas for FAIR data in materials science and related domains"[1].

The HarmonisedMatChem RDA Working Group aims to identify a set of strategies to improve the FAIR maturity[2] of materials data, for example, regarding indicator RDA-F2-01M[3]: "Rich metadata is provided to allow discovery". There is a need to adopt high-level metadata/schemas, agree on Application Profile(s) based on domain requirements and establish standardised vocabularies to support harmonised applications.

This report reviews the state of the art, with a particular focus on the current use of DCAT[4] and Schema.org[5], and the use of metadata for materials data cataloguing in different regions. Material for this report was gathered by means of interviews with stakeholders, desk research, input from Working Group co-chairs and a workshop (see Appendix).

General background

Here we briefly discuss the most widely used general metadata schema and their application in scientific disciplines, namely DCAT and Schema.org. Further context is provided by CDIF[6], the Cross-Domain Interoperability Framework.

While Schema.org is geared to the annotation of resources published on the web, thus improving their discoverability, DCAT-AP is built to drive interconnected data spaces. Choosing the best approach for the adoption of semantic technologies in materials science depends on the intended goals and associated complexity for achieving them.

For general background on metadata, their evolution, standards and application in the field of materials science, see a recent article by Greenberg et al[7].

[1] https://www.rd-alliance.org/groups/harmonised-terminologies-and-schemas-fair-data-materials-science-and-related-domains-wg/
[2] https://datascience.codata.org/articles/1241/files/submission/proof/1241-1-8492-2-10-20201027.pdf
[3] https://www.go-fair.org/fair-principles/f2-data-described-rich-metadata/
[4] https://www.w3.org/TR/vocab-dcat-3/
[5] http://schema.org/
[6] https://cross-domain-interoperability-framework.github.io/cdifbook/introduction.html
[7] Jane Greenberg et al, Towards MatCore: A Unified Metadata Standard for Materials Science, https://arxiv.org/abs/2502.07106

DOI: 10.5281/zenodo.17949060

DCAT

The European Commission adopted a European strategy for data[8] in 2020 which aims at creating a single market for data that will ensure Europe's global competitiveness and data sovereignty. Common European data spaces will ensure that more data becomes available for use in the economy, society and research, while keeping the companies and individuals who generate the data in control. The EU Commission supports the development of common European data spaces[9] in strategic economic sectors and domains of public interest. Some existing EU data spaces include the European Health Data Space and the European Mobility Data Space, based on the mobility DCAT-AP.

The DCAT Application Profile for data portals in Europe (DCAT-AP) is a specification based on the Data Catalogue Vocabulary (DCAT)[4] developed by W3C. It is a universal metadata scheme based on RDF, ready to be further profiled for specific domain needs.

A DCAT application profile (DCAT-AP) is a specification that reuses terms from one or more base standards, adding more specificity by identifying mandatory, recommended and optional elements to be used for a particular application, as well as recommendations for controlled vocabularies to be used. The basic use case for DCAT-AP is to enable cross-data portal search for data sets and make public sector data better searchable across borders and sectors. This can be achieved by the exchange of descriptions of datasets among data portals. Application profiles are specifications for metadata records to meet the specific application needs of data portals in Europe. They provide semantic interoperability with other applications on the basis of reuse of established controlled vocabularies (e.g. EuroVoc[10]) and mappings to existing metadata vocabularies (e.g. Dublin Core[11], SDMX[12], INSPIRE metadata[13], etc.). From the start, the DCAT Application Profile had the purpose of adapting DCAT to facilitate the reuse of data. Specifically:

- It proposes mandatory, recommended and optional classes and properties to be used for a particular application;
- It identifies requirements to control vocabularies for this particular application;
- It gathers other elements to be considered as priorities or requirements for an application such as a conformance statement.

DCAT Application Profiles

A notable use of DCAT-AP to managing the information requirements of the statistical and geospatial domains was developed in the ISA² Programme[14] of the European Commission, which has since become Interoperable Europe[15]. The two relevant extensions to DCAT application profiles are:

[8] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066
[9] https://interoperable-europe.ec.europa.eu/collection/semic-support-centre/data-spaces
[10] http://eurovoc.europa.eu/
[11] https://www.dublincore.org/
[12] https://sdmx.org/
[13] https://knowledge-base.inspire.ec.europa.eu/index_en
[14] https://ec.europa.eu/isa2/home_en/
[15] https://interoperable-europe.ec.europa.eu/interoperable-europe

DOI: 10.5281/zenodo.17949060

- StatDCAT-AP[16], developed in close collaboration with Eurostat, brings the statistical and open data communities closer by enhancing the visibility and facilitating the inclusion of statistical data sets in open government data portals; and
- GeoDCAT-AP[17] describes geospatial datasets, dataset series, and services. It provides an RDF vocabulary and the corresponding RDF syntax binding for the metadata elements of the core profile of ISO 19115:2003 , those defined in the framework of the INSPIRE Directive[18] of the European Union, and those defined in the Technical Guidance for the implementation of INSPIRE dataset and service metadata[19] based on ISO/TS 19139:2007. Its basic use case is to make spatial datasets, data series, and services searchable on general data portals, thereby making geospatial information better searchable across borders and sectors.

For further details on GeoDCAT, see the presentation by Jakub Klimek in the workshop.

Further Application Profile developments of relevance to materials sciences include:

- DCAT-eos-AP, discussed below. The development seems discontinued.
- DCAT-CHEM-AP[20], an Application Profile intended to be used by the NFDI4Chem & NFDI4Cat initiatives, to provide chemistry-specific metadata for a dataset.

Schema.org

Schema.org is a metadata standard primarily intended for annotating data web sites; however, it is applied more widely in linked data contexts (e.g. with JSON-LD). There are plenty of resources that explain the use of schema.org, e.g. https://schema.org/docs/gs.html

Bioschemas

Bioschemas[21] is an open community initiative aiming to improve the findability of web resources related to life sciences by using markup tags based on schema.org. The project extends and adapts schema.org types to better represent biological entities, datasets, and related resources, thereby facilitating data integration and discovery across heterogeneous sources. They provide a tutorial on what Bioschemas is, what the added value to schema.org.is and what the main elements in Bioschemas are.

Bioschemas provides a customised version of 'types', called 'profiles' which are used to define the semantics of a particular property, the valid value(s) and ranges that may be attributed to that property, and the cardinality with which that property may appear.

Bioschemas develops specifications for describing a wide range of life science objects, with some level of applicability also to materials sciences. They include metadata (properties) that

- exist in the core of Schema.org
- are proposed by Bioschemas to Schema.org

[16] https://interoperable-europe.ec.europa.eu/collection/semic-support-centre/solution/statdcat-application-profile-data-portals-europe
[17] https://semiceu.github.io/uri.semic.eu-generated/GeoDCAT-AP/releases/3.0.0/
[18] https://semiceu.github.io/uri.semic.eu-generated/GeoDCAT-AP/releases/3.0.0/#bib-inspire-dir
[19] https://inspire-mif.github.io/technical-guidelines/metadata/
[20] https://nfdi-de.github.io/chem-dcat-ap/
[21] https://bioschemas.org/

- exist in the pending area of Schema.org
- are reused from external vocabularies/ontologies

Note that for any "new"/added properties, i.e. not part of Schema.org or other established metadata schema, the Schema.org type "DefinedTerm" is used.

Most relevant ones include

- ChemicalSubstance[22]: includes a proposed "chemicalComposition", "chemicalRole[23]" etc.
- ComputationalTool: includes "applicationCategory[24]" and "featureList" etc
- ComputationalWorkflow[25]: includes proposed properties "input" and "output"
- Dataset: includes measurementTechnique and variableMeasured
- Sample[26]: which is a 0.2 release and lacks development of recommended properties.

These specifications are created through a collaborative process involving domain experts, data providers, and software engineers, ensuring that the resulting profiles are both scientifically accurate and technically sound. By providing guidelines and examples for annotating web content, Bioschemas enables resource providers to expose structured metadata that can be harvested by search engines and other automated agents, enhancing the visibility and reuse of biological data on the web. The initiative also maintains a registry of profiles and encourages its adoption through outreach, documentation, and community engagement.

Materials Science application

A materials science perspective on schema.org was given in a presentation by Zachary Trautt (NIST) in the workshop. In particular we note the use of schema.org in the European MaterialsCloud platform. Here we note the summary of opportunities to elaborate guidance for the use of schema.org metadata for materials science datasets, and that of the use of the DefinedTerm metadata that can be combined with existing terminologies to improve consistency. Both are relatively low-hanging fruit, short of launching a materials science equivalent to bioschemas.

Cross-Domain Interoperability Framework (CDIF)

In order to gain an overarching perspective on the state of the art of establishing data resources in a large diverse and interdisciplinary field such as Materials Science and its related domains, excellent recommendations and resources are provided by CDIF. As stated there, "CDIF is designed to address the question of what standards should be used, and how they should be implemented, which arises almost immediately in the implementation of systems in accordance with the FAIR Data Principles. The FAIR principles require that data and metadata

[22] https://bioschemas.org/profiles/ChemicalSubstance/0.4-RELEASE
[23] https://bioschemas.org/properties/chemicalRole
[24] https://schema.org/applicationSubCategory
[25] https://bioschemas.org/profiles/ComputationalWorkflow/1.0-RELEASE
[26] https://bioschemas.org/profiles/Sample/0.2-RELEASE-2018_11_10

be described according to common standards and made available through common protocols and mechanisms, but as principles, they do not specify which standards and protocols."

"CDIF attempts to provide recommendations to help resolve this issue with a set of non-domain-specific standards and their implementation for scenarios where FAIR exchanges are taking place across domain and infrastructure boundaries."

Here we note just some pertinent points from CDIF as a basis for our elaboration and recommendations.

CDIF is concerned with:

- Discovery (patterns for metadata content, serialization and publication)
- Data access (documentation of access conditions and permitted use)
- Controlled vocabularies (practices for the publication of controlled vocabularies and semantic artefacts)
- Data integration (documentation of the structural and semantic aspects of data to make it integration-ready)
- Universals (description of 'universal' elements – time, geography, and units of measurement).

For Discovery, the metadata recommendations of CDIF[27] fall into three groups:

- A Content model that specifies the information expected to be included in any metadata record, with required, recommended and optional content items.
- A JSON-LD serialization for that content using the **Schema.org** vocabulary to define the fields in a metadata record, and an implementation using the **DCAT** rdf vocabulary
- Workflows to publish CDIF metadata so that is can be found and indexed by search providers using standard web technology

CDIF specifies a metadata content model, utilising schema.org, DCAT and other standards.

It also emphasises the importance of Controlled Vocabularies: "Terminology-based semantic resources as a key element in information systems, establishing the binding between the symbols (strings) manipulated by computers and human-intelligible meaning of properties, types, values, or any other element in a volume of data. These are critical component in scenarios involving (but not limited to) data integration and harmonisation."

[27] https://cross-domain-interoperability-framework.github.io/cdifbook/metadata/recommendations.html

DOI: 10.5281/zenodo.17949060

Insights from initiatives in the US, Europe, Japan and Korea

United States

Material Research Data Alliance (MaRDA)

The Material Research Data Alliance (MaRDA[28]) is a community network focused on developing the open, accessible, and interoperable materials data that fuels the Materials Genome Initiative (MGI). The MaRDA FAIR materials microscopy and LIMS data working groups' community recently published a paper providing recommendations for managing, processing, and sharing research data and experimental context produced on modern scientific instrumentation, focusing on the microscopy domain, and laboratory information management systems (LIMS) for material science. The goal of this initiative is to plan, prepare, and submit research data in order to assemble significant amounts of FAIR data and enable breakthroughs in materials science research.

The Microscopy Metadata Group recommends:

1. instruments should capture comprehensive metadata about: a) operators, b) specimens/samples, c) instrument conditions, d) data formation;
2. microscopy data and metadata should use controlled vocabularies and standardized terminologies.

Examples of recommended vocabularies and standardised terms are:

Field	**Standard** (PIDs)
Bibliographic information: *who*, *what*, *where*, and *when*.	Dublin Core
Microscopy conditions.	HMSA, ISO/DIS 5820
Materials and electron microscopy instrument conditions.	NeXus and NXem
People.	ORCiD
Instruments.	PIDInst
Samples.	IGSN
Organisations.	ROR

[28] https://www.marda-alliance.org/

Other elements (e.g. dataset identifiers).	UUID
Published data.	DOI

The group evaluated existing schemas and concluded that “the **DCAT-eOS-AP**[29] data model "most closely adheres to the previously mentioned recommendations. It splits the data storage model into two primary groups, see also Figure from the paper:

- “core” elements that pertain to any type of data such as
    - Project, creator, description, publisher, date etc
- “granular” elements that contain domain-specific metadata.

A UML diagram and example of DCAT-eOS-AP[30] is available. The model is extensible, as other domains can plug into the model with customized schemas at the granular level.

MatCore

The MatCore project, already mentioned above, aims to “define required and optional metadata to accompany datasets generated through computational materials science techniques that will allow researchers to understand, use, and, if desired, reproduce the data. Similar to the hierarchical structure of DCAT-eOS-AP, it consists of two tiers.

However, in MatCore the “minimal” level includes both the general metadata such as creator, title and description as well as those pertaining to the domain, including “material”, “calculation-type” and “model-type”, while the more detailed metadata relate to specific types of modelling such as DFT.

Europe

Zenodo

Zenodo [31]is a widely used repository for open science and open data in Europe. For further information, see https://www.openaire.eu/zenodo-guide. Here we only briefly mention the

Metadata items available for representing entities uploaded in Zenodo, which are listed here: https://developers.zenodo.org/#representation[32]. They include a number of items with a Controlled Vocabulary such as “upload_type”, with Controlled vocabulary:

* publication: Publication

* poster: Poster

* presentation: Presentation

* dataset: Dataset

[29] https://www.osti.gov/servlets/purl/1777073/
[30] https://www.osti.gov/servlets/purl/1658673
[31] https://zenodo.org/
[32] https://developers.zenodo.org/

* image: Image

* video: Video/Audio

* software: Software

* lesson: Lesson

* physicalobject: Physical object

* other: Other

In addition, there are free text items such as “method” for the research method used, and mixed free/controlled vocabulary items such as

“subjects”: Specify subjects from a taxonomy or controlled vocabulary. Each term must be uniquely identified (e.g. a URL). For free form text, use the keywords field. Each array element is an object with the attributes:

* term: Term from taxonomy or controlled vocabulary.

* identifier: Unique identifier for term.

* scheme: Persistent identifier scheme for id (automatically detected).

Available controlled vocabularies include EuroVoc[33], MeSH[34] and GEMET

Some domain specific metadata are available including “Material sample ID”: An identifier for the material sample from which the occurrence was derived. However, the latter is allocated to the Biodiversity domain.

Zenodo entries can be exported in different metadata formats, and in particular DCAT-AP is supported.

In conclusion, we note that Zenodo is compatible with a DCAT-AP approach and collects a hierarchy of metadata including potentially domain-specific. However, as e.g. in schema.org, the materials science field is not well catered for, and can at best lean on biomedical metadata.

European Data Portal

An overview of the European Data Portal[35] is provided in the transcript of the presentation during the workshop below. Here we just reiterate that the whole portal is based on DCAT-AP, which ensures data linking via the metadata. In addition, Controlled Vocabularies are used, in particular EU vocabularies that cover e.g. publisher name, country names, file types, etc.

ESRF/ILL

European photon and neutron research facilities adopted the European open science policy through the PaNOSC[36] [2018-2022] and ExPaNDS[37] [2018-2023] projects with the aim to provide scientists with access to data, software and services from many scientific data sources in Europe and making them FAIR across facilities and scientific domains. For example, the

[33] https://op.europa.eu/en/web/eu-vocabularies
[34] https://www.nlm.nih.gov/mesh/meshhome.html
[35] https://data.europa.eu/en
[36] https://www.panosc.eu/panosc-project/
[37] https://www.panosc.eu/expands-project/

ESRF Data Policy for Public Access[38] for data taken at the ESRF beamlines, adopted in December 2023, includes recommendations from the PaNOSC FAIR Research Data Policy Framework[39], with the exception of maintaining the ESRF as the custodian of raw data and metadata. The metadata are stored in the ICAT metadata catalogue[40] and datasets can be accessed online[41] to browse and download (meta)data. The PaNOSC data portal[42] collects datasets from European photon and neutron facilities and uses a search engine compatible with OpenAIRE[43].

ILL has partnered with DataCite[44] as the registration agency for persistent identifiers. DataCite's mission is to be a world-leading provider of persistent identifier services to help make research outputs and resources findable, citable, connected and reused globally. The organisation has a dedicated working group developing and maintaining the DataCite Metadata Schema[45], a list of core metadata properties for accurate and consistent identification of a wide range of research outputs for citation and retrieval purposes. Regarding materials science relevant metadata, the ILL catalogue has the field `Formula` in the `Sample Parameters` metadata for specifying the chemical composition of a sample.

A more detailed overview of the ESRF data portal is given in a workshop presentation.

PSDI

An overview of the approach pursued by the UK Physical Science Data Infrastructure (PSDI[46]) initiative is provided in the transcript of the presentation during the workshop. Here we note in particular their use of DCAT and the top-down, hierarchical approach to metadata[47], ranging from metadata about resources catalogues, metadata about properties within these resources, metadata about data within the resources to metadata about provenance.

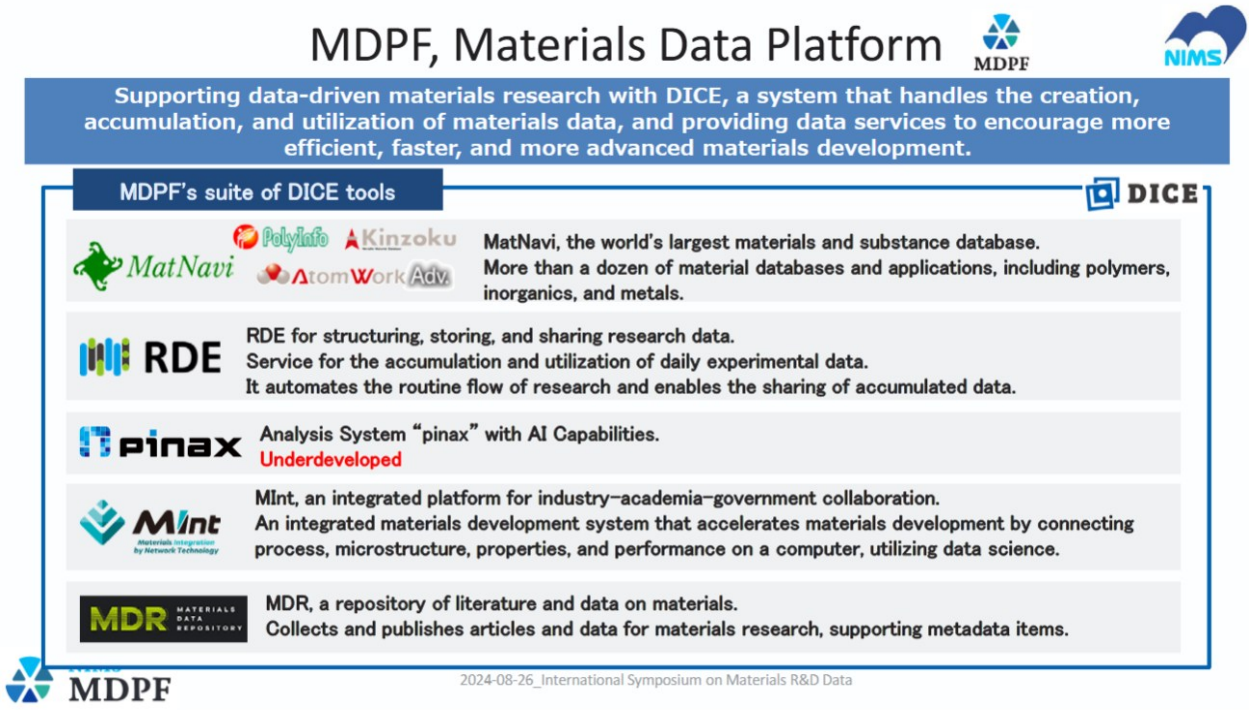
Japan

The Materials Data Platform, MDPF, includes a suite of data services tools, called DICE (Infrastructure of data offering service). An overview is shown in the figure below, from a presentation by Minamoto, Satoshi and Kadohira, Takuya at the 3rd International Symposium on Materials R&D Data[48] (26-27 August 2024 in Jeju, Korea).

Regarding metadata and data cataloguing, we note in particular the MDR, which is a repository for materials literature and data, supported by metadata items. More details about MDR can be found on the MDR services pages[49]. There are links to a Glossary and dataset metadata[50].

[38] https://www.esrf.fr/files/live/sites/www/files/Infrastructure/Computing/ESRF-data-policy_20240101.pdf
[39] https://doi.org/10.5281/zenodo.3862701
[40] https://repo.icatproject.org/site/icat/server/4.10.0/schema.html
[41] https://data.esrf.fr/
[42] https://data.panosc.eu/
[43] https://www.openaire.eu/
[44] http://www.datacite.org/
[45] https://schema.datacite.org/meta/kernel-4.6/
[46] https://www.psdi.ac.uk/
[47] https://metadata.psdi.ac.uk/
[48] https://www.ncmrd-symposium.net/previous-symposia/the-3rd-symposium-2024
[49] dice.nims.go.jp/services/MDR/
[50] https://dice.nims.go.jp/services/MDR/manual/html/metadata-data.html

DOI: 10.5281/zenodo.17949060



We note that the metadata are structured and some have defined lists/controlled vocabularies for possible entries:

- Generic metadata that are not materials/science specific
    - Managing organisation (responsible for the work)
    - First published URL
    - Title
    - Resources type, e.g. article, dataset, presentation, poster (defined list)
    - Keyword
    - Rights statement (defined list)
    - Creator
    - etc
- Generic metadata that related to materials science
    - Data origin (defined list): experiment, simulation, theory, etc
    - Material/specimen name
    - Material type
- Method related metadata with defined list as below
    - Characterization methods
    - Computational methods
    - Properties addressed

    - Synthesis and processing
    - Feature for machine learning (e.g. Vocabulary identifier for feature category.)
- Instruments metadata, including
    - Name of instrument
    - Data used
    - Instrument function
    - Manufacturer
    - Model number
    - Operator
    - Managing organisation
- Specimen details metadata
    - Chemical composition
    - Name/title
    - Crystallographic structure
    - Etc

The MDR schema is also maintained on github and serialised in yaml and json.

We also note the controlled vocabularies that have been developed and are available via the MatVoc Explorer[51]. MatVoc constructs dictionaries per domain and adds synonym definitions using semantic descriptions (RDF) for vocabularies that can be shared. While respecting the vocabulary within each domain, it supports FAIR operations through the strategy that fosters connections with other domains. In fact, MatVoc's vocabularies are used in MDR, ensuring that vocabulary management in MatVoc can be reflected in repository operations. Furthermore, integration with external databases has been achieved, e.g. Materials dictionary for the XAFS database project compiled by NIMS. For example, by linking vocabularies with databases such as the Crystallography Open Database, Materials Project, and MatNavi, successful data integration has been accomplished. IXDB[52] performs cross-database searches across more than 10 databases based on this vocabulary linkage. Statistics regarding this linkage can be found at https://ixdb.jxafs.org/statistics. MatVoc provides an external SPARQL endpoint, ensuring that the recorded vocabularies are always available to anyone in its latest version. IXDB is one concrete application of this.

Korea

The Korean Materials R&D Data Standard Expert Committee has developed a schema and vocabulary of materials R&D data, as shared on github[53]. The github site also includes overview presentations.

[51] https://matvoc.nims.go.jp/explore/en/home
[52] https://ixdb.jxafs.org/
[53] https://github.com/NCMRD/Standard-Expert-Committee

DOI: 10.5281/zenodo.17949060

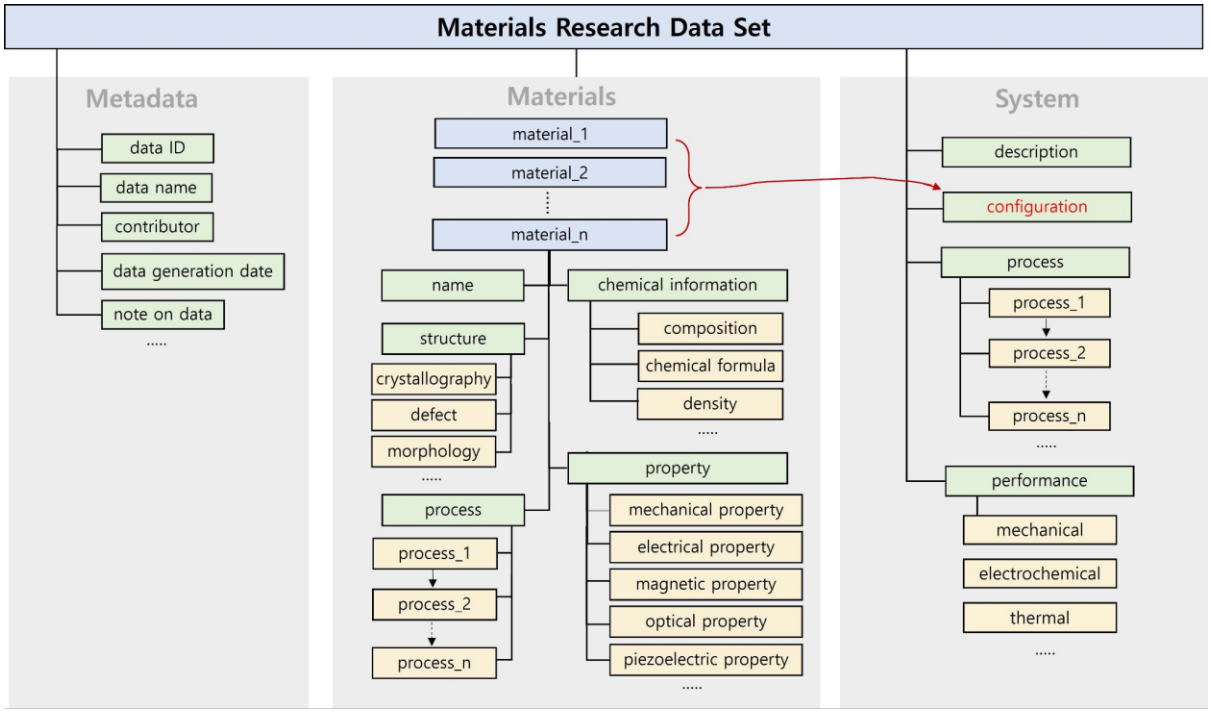
The context and application of the schema development is K-MDS, a materials research data collection, sharing, and utilization platform established and operated at the national level with the aim of creating a materials data-based research and innovation ecosystem in Korea.

The aim of the Korean data structure is to build a harmonized schema that can cover the wide range of disciplines of materials research. The data structure design is inspired by the workflow of materials research, from synthesis to performance characterization.

Basically the schema is hierarchical, including sections on:

- Generic dataset metadata including ID, name contributor, date
- Materials metadata for each material in the so-called Materials System
    - Name
    - Chemical information
    - Process (by which made)
    - Structure
    - Model
    - Property
- Materials System metadata
    - Description
    - Configuration
    - Process (in which the system operates/for which it is design, e.g. some catalysis)
    - Performance

The schema is also explained and illustrated below as well as on the K-MDS site[54].



[54] https://kmds.re.kr/web/guest/standardization

DOI: 10.5281/zenodo.17949060

The concept “Materials System” was introduced to describe the materials and their performance in the context of applications. It encompasses materials composed of single materials and multiple (composite) materials, developed to maintain a consistent data organization system in various application fields.

The separation of Materials metadata and Materials System metadata enables the description of materials with common (or same) Materials data structure, i.e. Chemical information, Process, Structure, Properties, but variety in Performance in different applications (Materials Systems).

For example, in the case of nanoparticle data, the Materials System data structure differs between nanoparticle application as chemical catalyst or as biomedical markers. Catalytic performance varies depending on the support materials and promoters. So, when documenting the performance of materials, we need to define all used materials in the system, their configurations, and process for building the system. This is the background of Materials System concept.

The schema and related controlled vocabularies are published in json format. The most recent published version is 2025-1 .

Recommendations

Materials science metadata profiles and harmonisation

To avoid the proliferation of new solutions, adding complexity to an already fragmented semantic landscape, it is advised to utilise and build on existing general-purpose semantic assets, such as Schema.org and DCAT-AP.

Two complementary approaches, DCAT-APs and Schema.org, have been already used for data annotation in scientific domains and should therefore be prioritised, along with using other existing general purpose resources such as DataCite.

1. Adopt a widely supported metadata standard as the basis for interoperability. The preferred solution seems to use `DCAT-AP` as a foundational metadata schema, which is a W3C standard, supports Linked Open Data, and is the adopted format in the European data portal.
2. Leverage Schema.org for web integration. Embed `Schema.org` metadata into dataset landing pages to improve discoverability via search engines. This schema too should be tailored to materials science data, following the example of bioschema.org. Use `JSON-LD` format for structured metadata that can be parsed by web crawlers and reused across platforms.
3. Elaborate recommendations and profiles for the use of existing metadata with materials science datasets. This should include (a) recommendations for the use of DCAT and Schema.org metadata and collaborating on a Material-DCAT-AP tailored to materials science, similar to GeoDCAT-AP for geospatial data.
4. Normalize terminology with controlled vocabularies:
    - Avoid open-text fields for critical metadata such as sample composition.

    - Use enumerated lists containing links to persistent identifiers or controlled terms.
    - Use existing resources like EuroSciVoc[55].

5. Enable metadata harvesting and federation. Leverage existing data portals, which may require developing specific tools for parsing and serialising the data in a specific structured format.
6. Design reusable metadata templates for common experiment and computational simulation types, with clearly defined properties. Allow flexibility but encourage publishing templates to shared registries for broader adoption and reuse.

We recommend adopting a hierarchical approach, consistent with several studies and implementations in the US, Europe, Japan and Korea, discussed in this report. We recommend three tiers:

1. A first level of "generic" metadata, for example following the CDIF recommendations.
2. A second level of physical sciences/materials science metadata. Materials-specific metadata should provide information on material, material composition or chemical structure, synthesis and processing methods, provenance and experimental conditions, theoretical models and approximations used. These should be widely agreed and accompanied by Controlled Vocabularies to ensure harmonisation, as recommended by CDIF.
3. A third level which is the responsibility of domains. A number of domain-specific initiatives already exist and their schema can be utilised.

Methodology

Improving data cataloguing in materials science and making the above recommendations a reality requires motivating the community.

It is useful to consider first of all potential scenarios for publishing data. In one case, decentralised institutions and individuals alike publish their data independently, for example on the web. In the other case, more 'centralised' data portals are created (or existing ones adapted) and maintained by a network of established institutions which agree on protocols for data submission, metadata harvesting and ideally interoperability between the repositories.

In the decentralised scenario, the contributing data are aggregated by specialised search engines (e.g. Google's Dataset Search) or data portals (e.g. Europe's Data Portal[35]). The advantages are that potentially every stakeholder in material science could become a data provider, allowing a plurality of semantic solutions to be accommodated, given that quality assurance and mapping to industry standards would need to be carried out by a curating body. While providing a lot of freedom, there may also be barriers to adoption due to a lack of support for the potential contributors, and a lack of clarity about what happens to the data. Without a clear benefit for the contributing user (or, alternatively, a normative framework e.g. requiring publicly funded research to be accompanied by corresponding datasets), it can be expected the growth of FAIR datasets for material science to be modest, at best.

[55] https://op.europa.eu/en/web/eu-vocabularies/euroscivoc

DOI: 10.5281/zenodo.17949060

In the centralised scenario, the burden of choosing a specific semantic approach (e.g. selecting standard vocabularies and choosing the relevant metadata for different disciplines) falls on a central authority, as do the tasks of receiving, validating, storing, and distributing the contributed datasets. Whilst this approach has a built-in mechanism for ensuring conformity to a single or a handful of standards, its success will heavily rely on the creation of information management systems in research institutions, which will ultimately be responsible for providing a platform through which individuals can upload their contributions. There is also a need for agreement on and financial support for the authoritative bodies that e.g. are responsible for metadata catalogues, controlled vocabularies, implementation of portals etc.

In any case, the following points should be considered to encourage involvement of the community:

1. Engage with domain scientists, data stewards, and software developers in defining metadata needs. Use application profiles as a collaborative tool to formalize shared practices and expectations.
2. Provide immediate feedback to researchers on how metadata improves searchability and reuse (e.g., filtering by material class or property). Use successful use cases also from other fields (e.g., Paleo project[56]) where metadata led to tangible benefits.
3. Integrate metadata collection into electronic lab notebooks (ELNs) and workflow orchestrators capturing context as experiments are performed. Reduce manual burden by automating metadata extraction where possible.
4. Encourage national-level templates (e.g., Czech Republic's DCAT-AP-CZ) and cross-domain alignment. Identify controlled vocabularies and metadata profiles to ensure compatibility across disciplines and countries.
5. Develop searchable registries of metadata templates and schemas, allowing researchers to find and adopt existing models. An open repository could be used to collect and refine schemas, using community feedback and adoption rate as a democratic approach to standardization.

[56] https://paleo.esrf.fr/

DOI: 10.5281/zenodo.17949060

Acknowledgement

The authors would like to acknowledge the contributions of the following co-chairs of the RDA WG “Harmonised terminologies and schemas for FAIR data in materials science and related domains”:

- Kwang-Ryeol Lee (Korea Institute of Science and Technology, Korea)
- Masashi Ishii (National Institute for Materials Science, Japan)
- Marek Cebecauer (J. Heyrovsky Institute of Physical Chemistry, Czechia)
- Zachary Trautt (National Institute of Standards and Technology, US).

Also, we are grateful to the presenters at the workshop and their input:

- Jakub Klímek (Charles University, Prague, Czechia)
- Simon Steuer (Publications Office of the European Union)
- Aileen Day (University of Southampton and PSDI, UK)
- Guillaume Gaisné (ESRF, France)

This work received funding from the RDA via its RDA Tiger support scheme and also from UK Research and Innovation (UKRI) under the UK government’s Horizon Europe funding guarantee (grant number 10091190) related to the European Union’s Horizon Europe Research and Innovation Programme, GA No. 101137725 (BatCAT).

DOI: 10.5281/zenodo.17949060

Appendix: Workshop on Data Cataloguing for Materials Science and related domains

Background

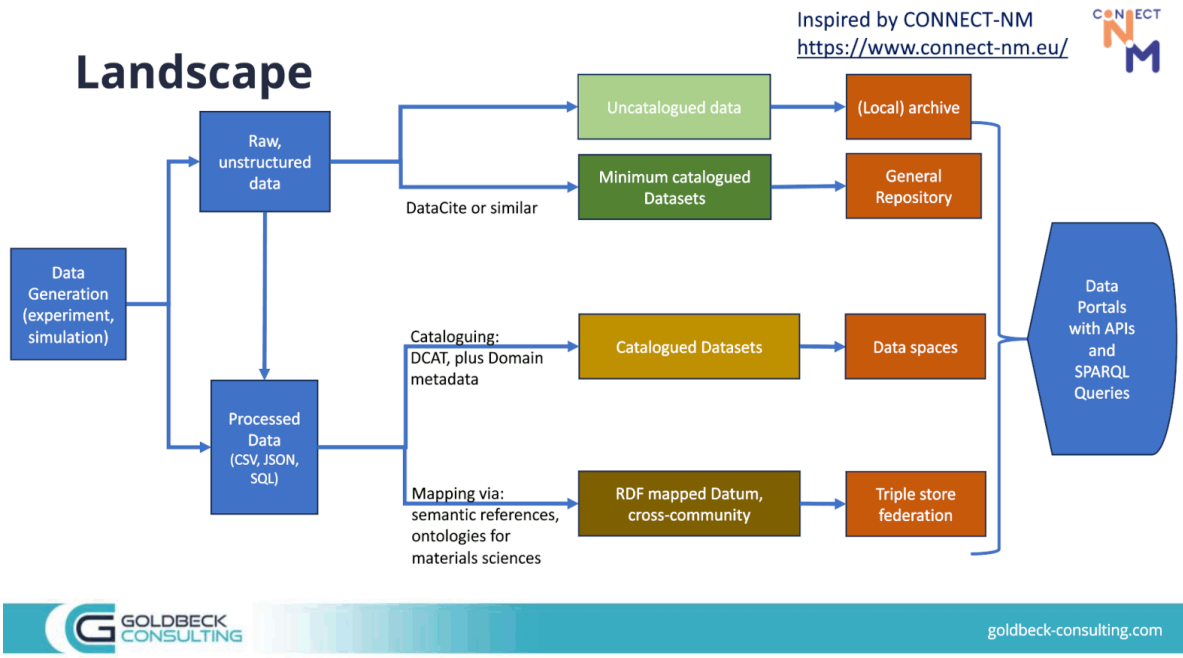
The workshop was held online on 24th September 2025 to provide an overview of some of the key, mostly European, initiatives relevant to topic of data cataloguing. The question to be addressed was: what can we learn from these initiatives and other initiatives, and what are the barriers to overcome?

Workshop contributions

- Introduction (Otello Roscioni and Gerhard Goldbeck, Goldbeck Consulting Ltd)
- GeoDCAT-AP (Jakub Klímek, Charles University, Prague)
- Materials and Schema.org (Zachary Trautt, NIST)
- Data-Europa (Dr. Simon Steuer, Deputy Head of Unit, Publications Office of the European Union)
- Data integration in the UK Physical Sciences Data Infrastructure (Aileen Day, University of Southampton and PSDI)
- ESRF data portal / AddMorePower project (Guillaume Gaisné, ESRF)

Transcript of the workshop

Introduction (Gerhard Goldbeck)



This landscape, from the left to right starts with a data generator which could come from either experiments and computational simulations, resulting in unstructured raw data, which are then

DOI: 10.5281/zenodo.17949060

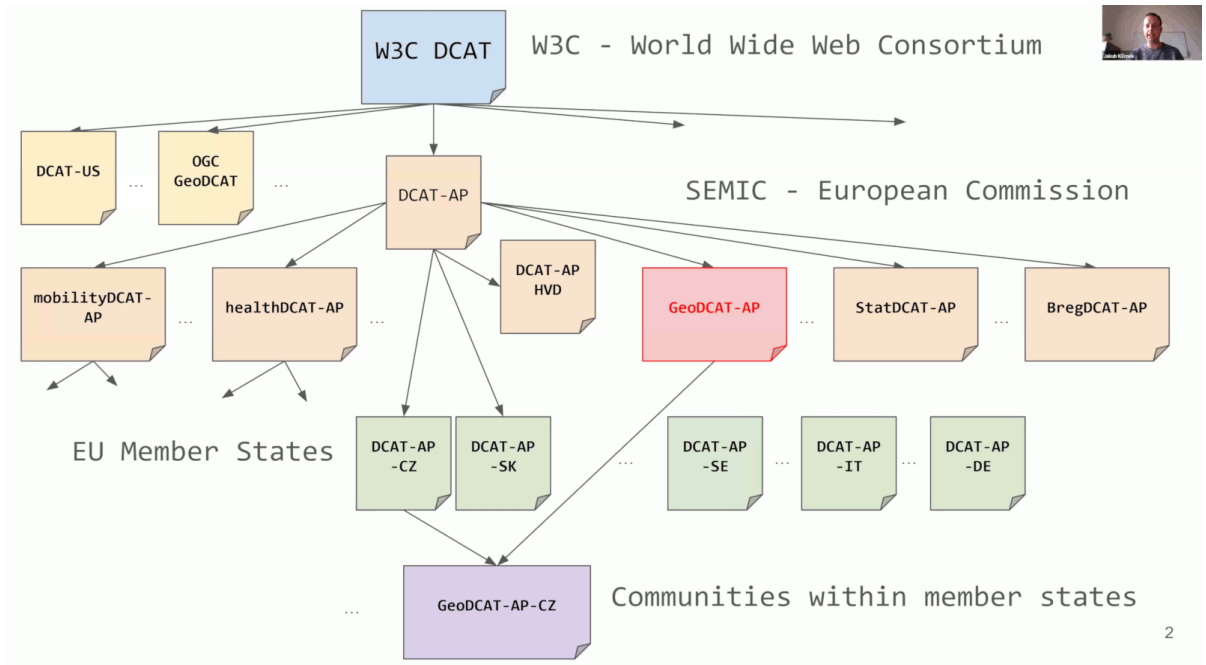
processed in some way. They can exist as e.g. CSV files or, better, as JSON files, or even better in some sort of database schema, but they're still kind of local. Going back to the top, you could have either uncatalogued data in some local archive or, with a data site or similar, they are added to a general repository. More detailed cataloguing would use, for example, DCAT to be included in some sort of data spaces with more rich metadata.

Eventually, but not today, we could be talking about RDF and semantic reference integration all the way to what is called data. So really all of the data details that are actually inside the dataset to be mapped out. But the question is, what level of detail about the data is in an application profile?

GeoDCAT-AP for metadata of geospatial datasets (Jakub Klímek)

I am (also) the editor of Geo-DCAT-AP, which is an application profile of DCAT-AP for metadata geospatial datasets. We have seen two terms already in the previous presentation, DCAT, by the W3C and DCAT-AP, but the ecosystem of DCAT and its application profiles is much wider than that nowadays. So I'll provide a little bit of an overview.

In the route on the top of the hierarchy, there is the W3C DCAT defining the main terms in



metadata or in data catalogues worldwide, such as what is a dataset, what is a distribution, what is a data catalogue. But since it is a W3C recommendation and is used worldwide, it is also a little bit vague on purpose so that it fits as many use cases as possible worldwide.

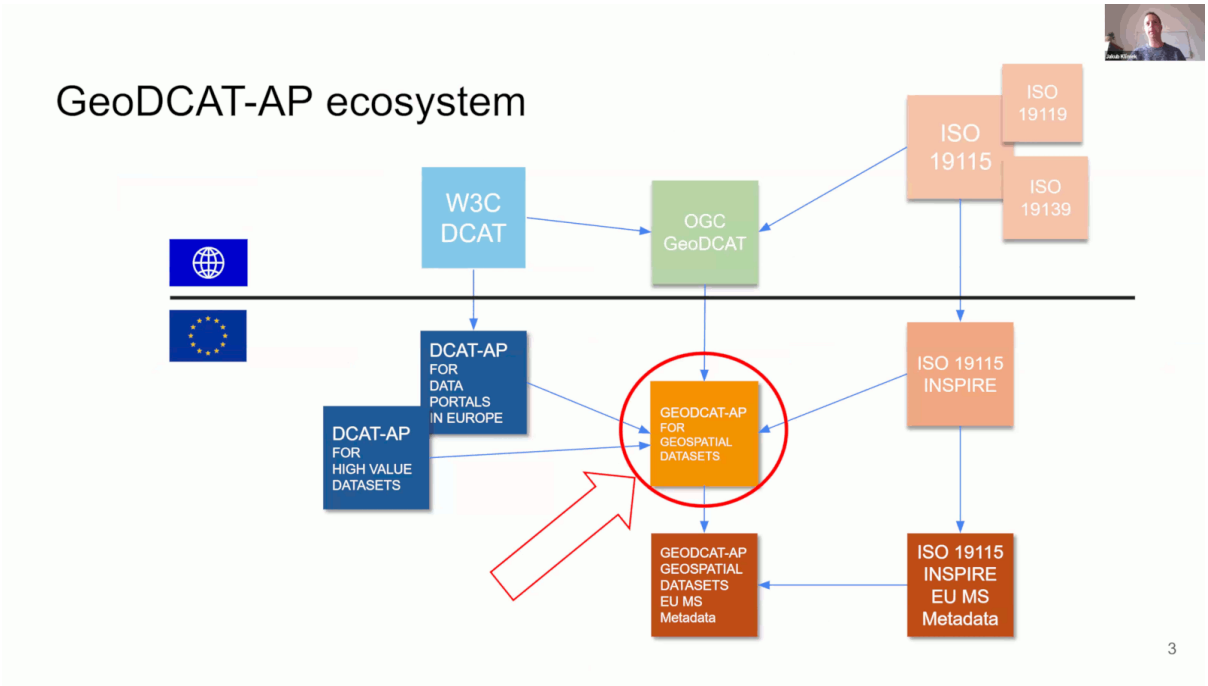
And baked into the W3C DCAT is the support for application profiling, meaning that the authors of DCAT provide the definitions for the basic terms, but how to use them in a specific context is up to you. You should create an application profile of DCAT with the specifications that are relevant to your context. So there is DCAT-AP, which is maybe a little bit confusing because it says DCAT application profile as if there was only one, but DCAT-AP is the European application profile with the European specifics implemented, for instance, in the European data portal. But there are also other application profiles of DCAT, DCAT-US, or the OGC-GeoDCAT, which is a new initiative to create a worldwide application profile for geo data. But let's focus on the European context. So we have DCAT-AP with its extensions for high-

DOI: 10.5281/zenodo.17949060

level datasets. But then we have specific needs in specific domains. So there are application profiles for each domain based on DCAT-AP further specifying or addressing the needs of those communities. So there is mobility DCAT-AP, health DCAT-AP, GeoDCAT-AP , which I'll focus a little bit more on now, StatDCAT-AP for statistical datasets, and so on. So those are all different European application profiles of DCAT-AP. But those are domain specific. And in addition to those, there are also national extensions or national application profiles to DCAT-AP, such as DCAT-AP-CZ, the Czech one, the Slovak one, Swedish, Italian, German, and so on. Basically, every member state has its own DCAT-AP profile. To make things a little bit more complicated, when you take a community of Czech people who are interested in geospatial datasets, they now need to adhere to a Geo-DCAT-AP for geospatial data and DCAT-AP-CZ for the Czech datasets together. And now the key question here is are those profiles compatible? Can they be used together to describe the single dataset? Are there any conflicts, and so on? Because each application profile listed here is managed or maintained by a different community. So it is quite crucial that some rules on how those application profiles are created, and how they look like, are followed, so that the application profiles are combinable downstream in those communities such as the geospatial community in the Czech Republic and so on.

Now let's zoom in on GeoDCAT-AP, which is one of all those application profiles. Now, this specification has multiple inputs. On the DCAT site, it's quite naturally, the W3C DCAT, DCAT-AP, and DCAT-AP for hybrid datasets, because many geospatial datasets are high-value datasets. But on the geospatial side, there already are standards on how to describe metadata of geospatial datasets, and those are ISO standards. And then there is the INSPIRE directive, the European directive that says how metadata of geospatial datasets should look like, and those are XML-based standards. And as we all know now, GeoDCAT-AP and DCAT-AP are all RDF-based. So to address the needs of the geospatial community, there needs to be a mapping of those XML standardized metadata into the world of DCAT-AP, which is RDF-based. Now, to give you just a glimpse of how big the Geo-DCAT-AP profile is, it has the same 13 main entities as DCAT-AP. No changes there. There are 29 supportive entities versus 21 in DCAT-AP.
What's new is specific support for addresses, emails, entities metrics for quality annotations, and so on. Some of the supportive entities in DCAT-AP are unused. Then for the main entities, there are additional properties. A big set of additional properties are agent roles, which are more specific to the geospatial domain, such as custodian, distributor, originator, and so on. And then, some geospatial-specific properties, such as a reference system, spatial resolution, topic category based on the INSPIRE directive, and so on. And then there are 11 controlled vocabularies included in GeoDCAT-AP based on the INSPIRE directive. And only two of those

DOI: 10.5281/zenodo.17949060



are mapped to ones that are already present in DCAT-AP. The rest is used as is, basically, in the GeoDCAT-AP profile.

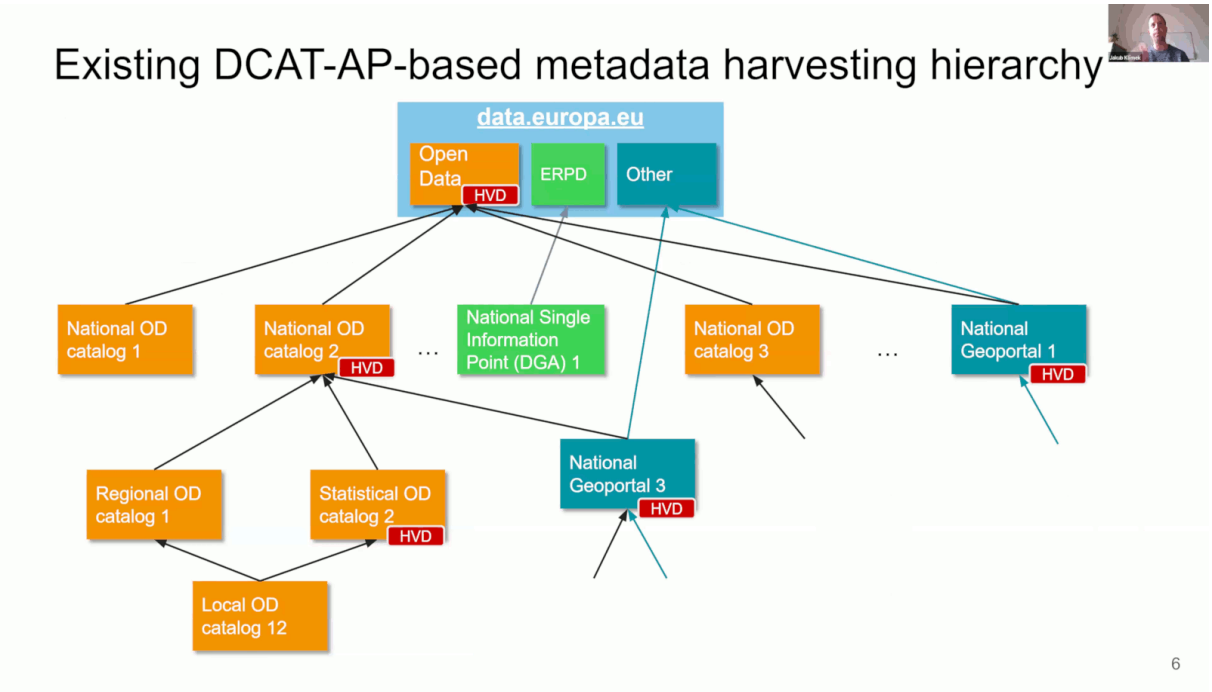
Now, where is Geo-DCAT-AP used? Well, as I said, it is mainly used in the geoportals that fall under the INSPIRE directive. So here you can see quite a large list of all the topics which are covered by the INSPIRE directive. And in each topic, you'll see a set of datasets in each member state that is described by inspired metadata and if the publishers want to make the metadata accessible to the wider open data community, and the DCAT community, it is also described by Geo-DCAT-AP, or at least mapped to Geo-DCAT-AP.

Now, in Europe, we have this DCAT AP based metadata harvesting hierarchy. On the top, there is the European Data Portal (data.europa.eu), and that harvests national data catalogues or other data catalogues, the national data catalogues can harvest metadata from regional or domain-specific catalogues and so on. So it's a whole hierarchy. What's interesting for us now is these geoportals, which are also harvested either directly to data.europa.eu, or through a national open data catalogue and from there to data.europa.eu. Now, you can see that there are multiple paths that metadata can take from the original portal to data.europa.eu. So that's why a unique identification of datasets is very important so that the possible duplicates can be reconciled when they happen in this harvesting hierarchy. Now, as I mentioned, the INSPIRE metadata is XML based. You can see a short example on the left-hand side and DCAT which is RDF based on the right-hand side. And clearly, there is a need for a mapping. So there is a reference XSLT-based mapping from the XML metadata to Geo-DCAT-AP-compatible metadata, available in GitHub and maintained as part of the Geo-DCAT-AP application profile. Now, where this XSLT transformation is used is another question. Some geoportals or most of geoportals inside have metadata represented according to the INSPIRE-XML. And there are two options. Either the geoportal itself implements this transformation and exposes a second endpoint which is Geo-DCAT-AP-based and can be harvested by DCAT-AP-enabled portals. Or this transformation is used on the other side of the harvesting relation. So for

DOI: 10.5281/zenodo.17949060

instance, in data.europa.eu, the XSLT transformation is implemented and therefore this portal accesses the INSPIRE-based API exposed by the geoportals.

Now, how the feature looks like, it seems that in quite a near future Geo-DCAT-AP will become an alternative to the INSPIRE-based metadata. So geoportal maintainers will be free to choose whether they use Geo-DCAT-AP natively for the geoportals, which we see demands for, or whether they will still keep INSPIRE metadata inside and work with that. But as I said, there



is a demand for support in the community for implementing Geo-DCAT-AP native geoportals. And that's it from my short presentation.

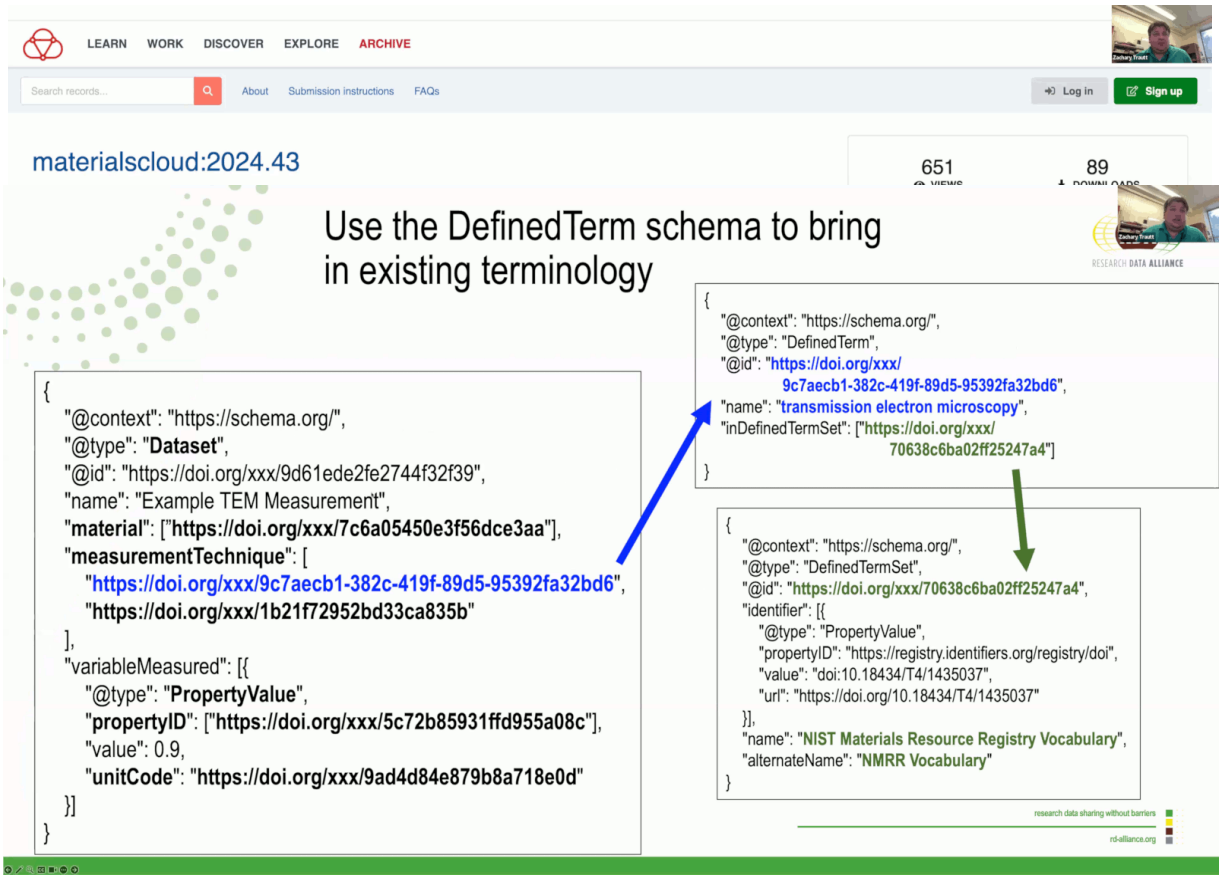
Schema.org and Materials Science and Engineering (Zachary Trautt)

I was asked to share a perspective on Schema.org from the material science and engineering community. So I just wanted to start by showing what Schema.org is. And this is just a screenshot of the homepage and it's a collaborative effort for metadata being embedded into web pages and landing pages for datasets. And it's used. **It has quite widespread use**. So I think it's **something that we shouldn't overlook**. And just to give one example of some of the things that it does, let's consider a schema for a recipe, like for pancakes. And if you go to any major search engine and perform a search for a particular recipe, you'll see that the results at the top have structured output. In other words, if I do a search on Google, it shows me a star rating, how long it takes, and the ingredients. Likewise, if I do a search on DuckDuckGo, it's showing the ingredients. And in Bing it shows the cooking time, the calories, the servings, and a photo from the parent web page. So something interesting is going on behind the scenes to give this structured information at the level of a search engine. And if I go and look at one of these recipes, behold and embedded within the web page is JSON-LD with this highly

DOI: 10.5281/zenodo.17949060

structured metadata and data. And so that's just sort of an interesting introduction of how structured metadata influences our day-to-day lives in a personal way.

There is a dataset schema that's sort of relevant to these discussions. And if I go to a material-specific repository that supports Schema.org, you can see that they've embedded a tremendous amount of metadata using that dataset schema. And I just wanted to sort of share what I think are the opportunities here. And I've sort of ranked them from easiest to more difficult. And one of the easiest things we can do is if we're running a repository or we're trying to publish data and we need to select a repository, we could choose one that supports metadata embedded within dataset lining pages including Schema.org. The next level up with that from that, we could think about, as a community, leveraging this schema with specialized guidance. And I'll talk about each of these. We could also think about pairing the defined term schema with the dataset schema as a normalization layer to reuse existing terminology. Or we could even launch a materials equivalent of bioschemas. So I'm going to skip over that first one, but using a dataset schema with specialized guidance to sort of boil that down and simplify what that would look like. We would basically sort of suggest which properties we



should use. If you go and look at these schemas on Schema.org, they're quite vast. So if everybody's sort of in isolation looks at them, which properties get selected might not be the uniform across the community. So guidance could help with that. And with each property, there could be guidance on how to use that property. Looking at schema.org, it's quite open-ended. So the values that people put in these properties, in addition to selecting very different properties, you could put different values. And then we're back in the same boat of non-uniform metadata. And providing this specialized guidance for an existing schema is what's known as a **profile** in the bioschemas regime.

DOI: 10.5281/zenodo.17949060

In my sort of difficulty chart, I also talked about using the defined term schema in conjunction with a dataset schema. And this is kind of a really overwhelming example, but I'll walk through it one line at a time. So here I'm showing a demo dataset using the schema.org dataset schema. Let's start from this measurement technique property, which has a URL pointer to somewhere else. Before I follow that link, I'll notice that I named my dataset "Example TEM measurement", where TEM is a domain-specific acronym, and it's an open-ended text box: so one person might type in "TEM", while another person might type in "Transmission Electron Microscopy", which is kind of that point of the example that I'm showing here, where using this measurement technique property, there could be a URL in there that points to where that term is further defined. And here I'm showing that this URL points to a defined term as a normalization layer over where that term might exist in varied formats. But what this link points to is that term being transmission electron microscopy. And since it's reusing an existing term, then we have a controlled vocabulary. So if one person is doing "TEM" and another person is doing "Transmission Electron Microscopy" in terms of what they might type into an open-text field, this brings in some level of control over the vocabulary. And here I'm also showing that this term, transmission electron microscopy, lives within a defined term set. And here's another URL pointing to a defined term set for the material's resource registry vocabulary, something that was an output from a past RDA working group, which produced a schema, a demo, and a controlled vocabulary. So I just like to show this because out of the box with the dataset schema and the defined term schema, I think there's a really low hanging fruit of using those as a normalization layer between our datasets and existing vocabulary ontology that are in a lot of different formats.

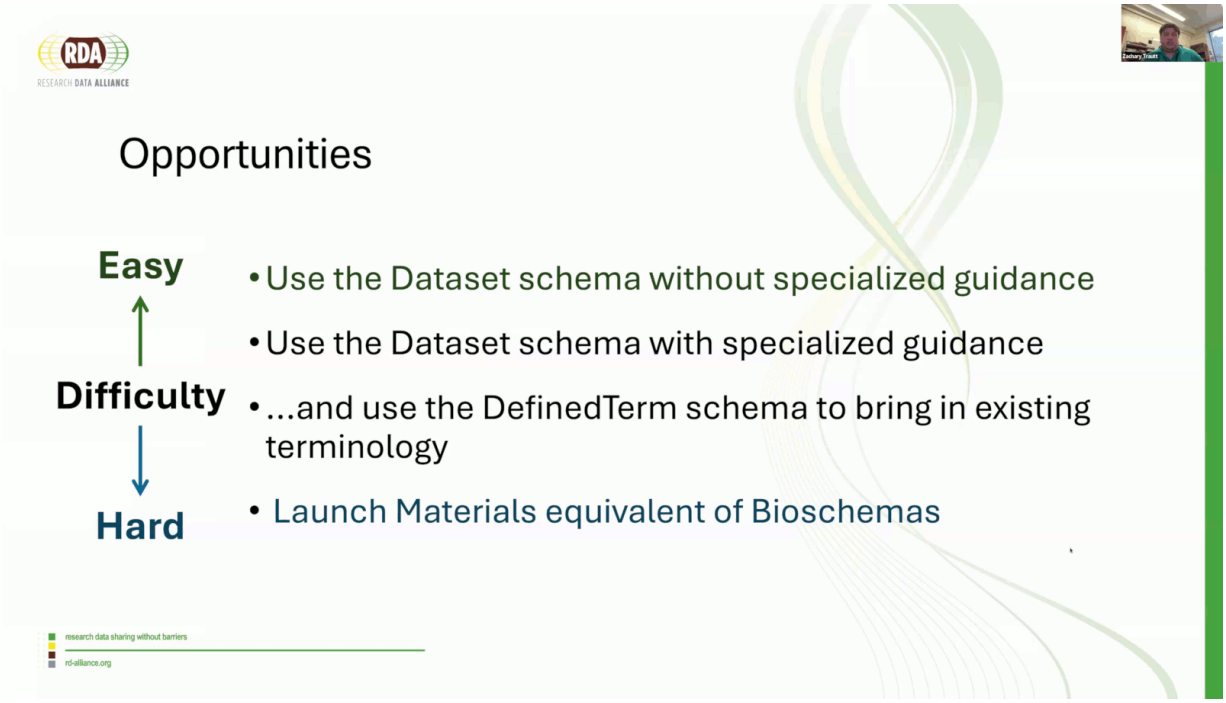
And lastly, bioschemas is something else to look at. It's a little bit more difficult in the sense that this thing is supported by multiple working groups to develop profiles as well as new types and new properties in the sense that, in the schema.org vocabulary, a recipe is a type and the time that it takes to prepare that meal is a property. And schema.org is rather weak in terms of processing history of a material. So if we were to get serious about schema.org and being able to sort of fully support process, structure, property relationships, **we'd have to develop probably new types and new properties on the side of capturing the processing history of experimentally created material.** And to conclude I want to restating what I personally

DOI: 10.5281/zenodo.17949060

see as the opportunities. Others may see other or additional opportunities and different levels of difficulty.

The European Data Portal (Simon Steuer)

I work for the European Commission and I want to present you the European data portal. So I want to show you a bit how the European data path could be interesting for the research domain. But first, what is it? It's basically a catalogue of metadata. So we provide the data about the data so that people can use it, search for it and everything. And we have a legal



basis: it's a directive of open data. So every member stage should publish their open data. And also there's a decision inside the European Commission that all the documents which are available, statistics, files, and so on should be made available for reuse. The same goes for the European Council. And we publish this under Creative Commons, which we think is the best format for this, as people really know they can reuse it. And it's also machine readable. What is also interesting, we translate all the metadata in 24 languages. So you can search across countries, which is sometimes interesting because most statistics don't stop at country borders, for example data about air quality.

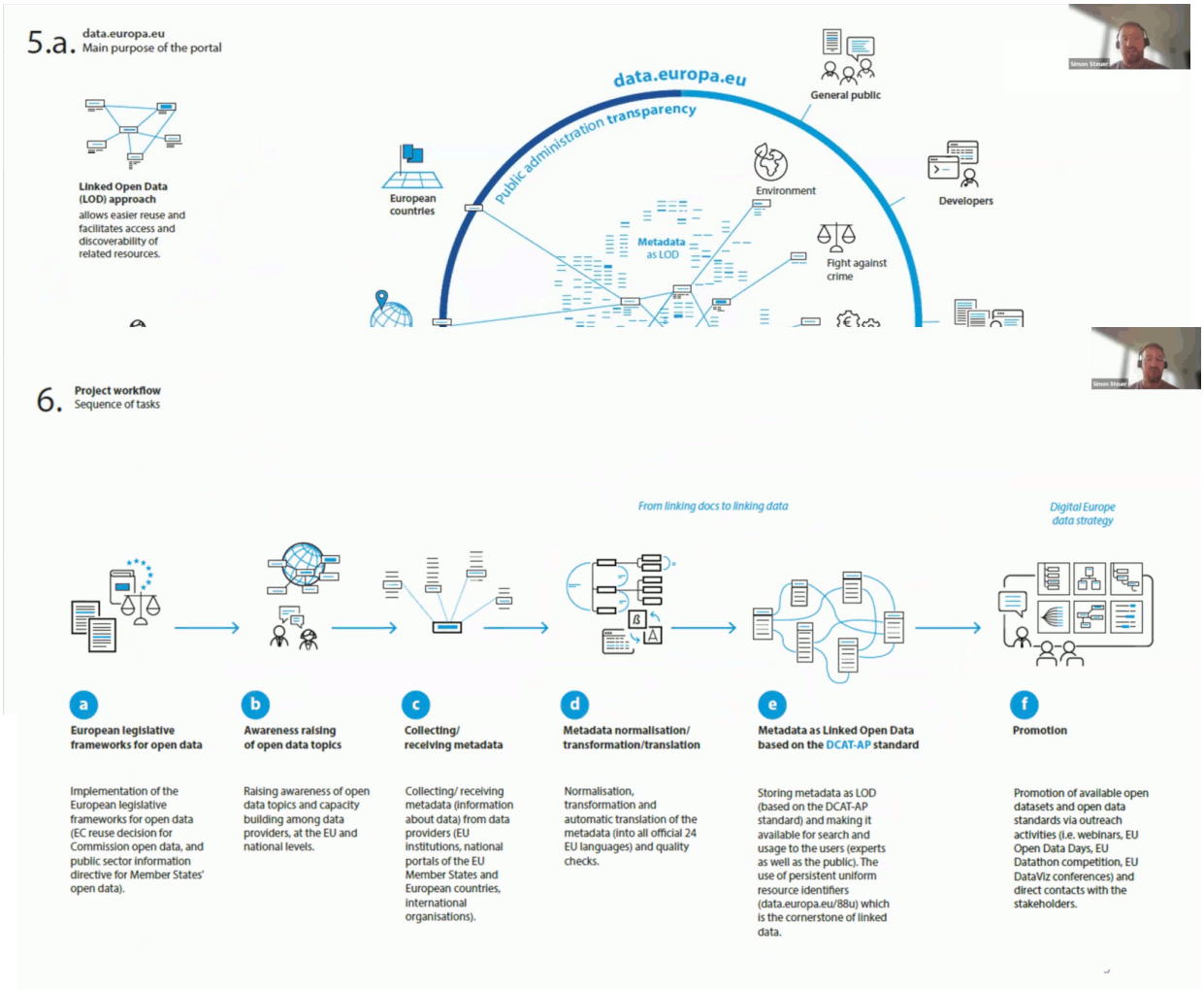
Our main data providers are European countries. It goes beyond the European Union. So in total, we have over 30 countries. And we have all the EU geo data portals. So you can find basically all INSPIRE metadata here. And we have all EU institutions and agencies for EU specific data. And we also have more and more EU research projects. So most of them are required if they get funding from the European Commission, for example, to publish their data. Most of the time they create their own website, but it's not really sustainable. So it's much simpler if they upload the metadata and the data to our portal. And then it will be translated and available for everyone. We offer services on top. So it's not only the data access and metadata quality checks. We check that and give data providers feedback on what can be

DOI: 10.5281/zenodo.17949060

improved. We foster the reuse of data, for example we organize competitions and conferences with quite some outreach. We're also doing data literacy training courses and so on.

In general, the whole portal is based on DCAT-AP, which follows a linked open data approach. So all the datasets are linked to each other via the metadata. So if they come from the same catalogue, they have a link. If they use the same keywords, they have a link. And we're using EU vocabularies, which we think is a really nice solution for tagging datasets because it's a controlled vocabulary already including translations. There are many controlled vocabularies inside EU vocabularies. But for example, for publisher name, for country names, for file types, those are highly recommendable.

So in general, our workflow goes like this. So we have a legislative framework. We raise awareness about those topics. And then we collect the metadata. This is the trickiest part



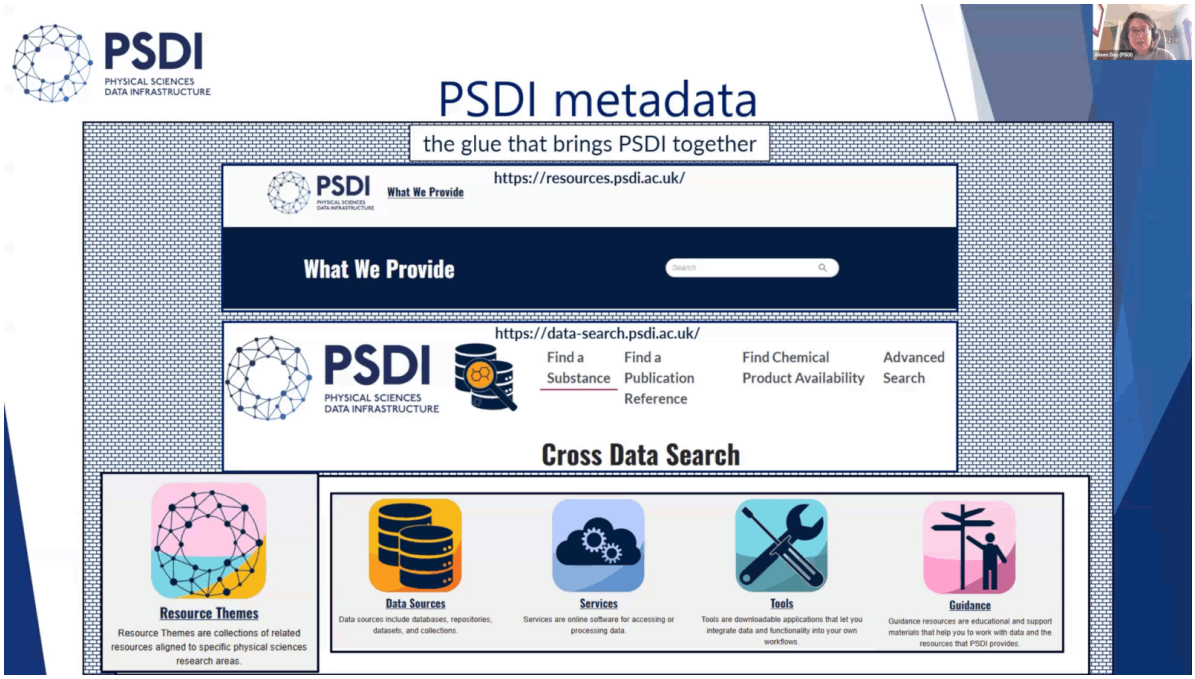
because not all EU countries use the same standard. Many use different standards, have different profiles on their own, different translations and so on. That means we have to do a lot of normalization, transformation, and so on. And we have different harvesting profiles for each country, for each catalogue. And as always, the more people use a standard, the easier it gets, and also the easier the reuse for everyone gets. So this standard is DCAT-AP, which is a worldwide standard. It's really flexible. It's useful. It's highly adaptable. We can only recommend everyone to use it. Also, for your internal databases, it's a really nice way to do everything.

DOI: 10.5281/zenodo.17949060

And then I want to show you a bit of the data hub. So you have all these datasets. We harmonize them, and then we translate them. And you can basically find all the datasets you want. And you can also filter them by metadata quality, for example. You can search throughout the languages. And in total, we have 1.9 million datasets from 202 data providers. You can filter for them. And we also have many features on top. For example, if there's a CSV file, you can automatically transform it into different file types. If there are CSV or geospatial datasets, you can get very quick visualizations without downloading the data. You can get feedback on the metadata quality. You can cite a certain dataset with our citation tool. And if you're a data provider, you can also upload your data and then link to it, which is also a handy tool. And this is more the data part, and now a bit about the outreach.

So we're just offering the portal that is not enough. So we have news where we show a bit about open data. We have events where we show what is going on in the data world. We are doing data stories where we show others what could be done with data and specific datasets. We have an academy where we do training courses for others. We run a yearly open data maturity exercise where we check how far advanced European countries are, and we see a very positive trend for all the countries. So everyone is getting better all the time. And we're also doing some studies for interesting topics. And overall, we just want to promote the open data world and make as much data in as good format as possible available.

Data cataloguing in PSDI (Aileen Day)



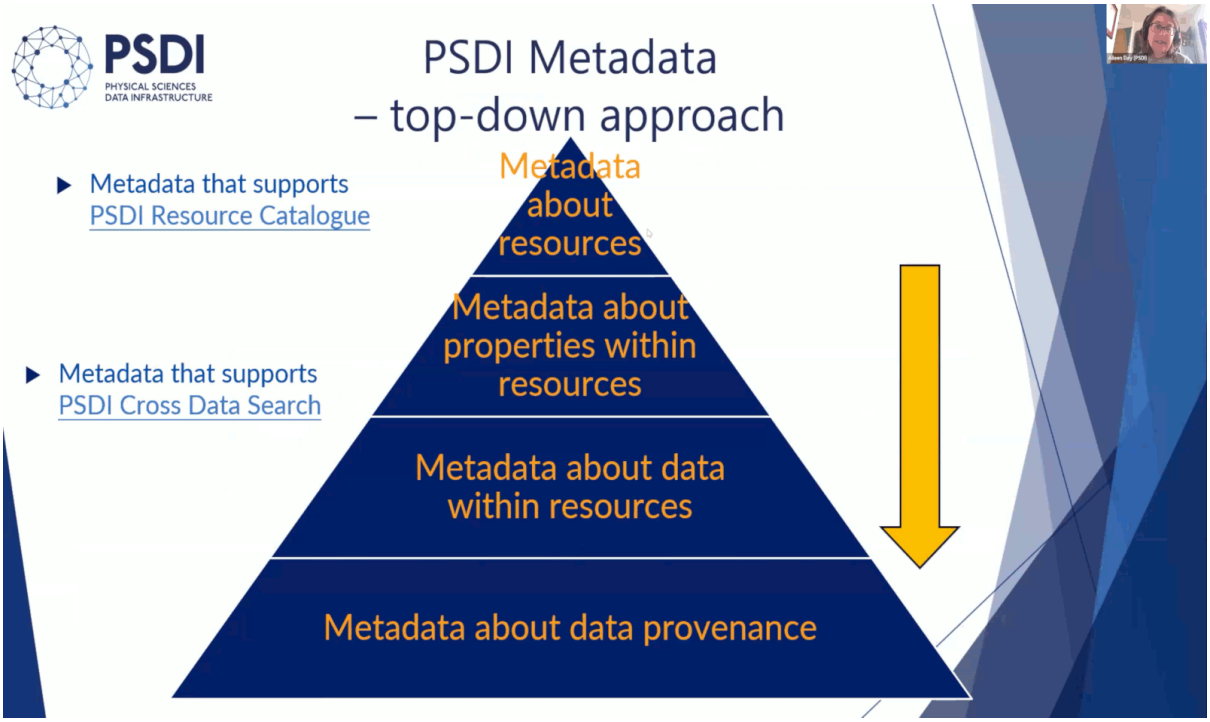
I am based at Southampton University and I work for PSDI, where I am the main metadata lead. PSDI is basically a data infrastructure, the UK physical sciences research data. It's fairly new. I've been working on the project for one and a half years. And the first version of it was released in March this year (2025). And that was very much seen as a Version 1 minimum viable product that we've put out and is now available for people to see. The PSDI metadata gives us the terminology to structure PSDI and to describe data. The PSDI metadata are structured by gathering everything into resource themes, which are about particular areas of the physical sciences sorted by a certain group. The resource themes can contain data, but also services and tools, which are both forms of software services that are run on the website. So you don't need to download it, whereas tools you download onto your own computer and

DOI: 10.5281/zenodo.17949060

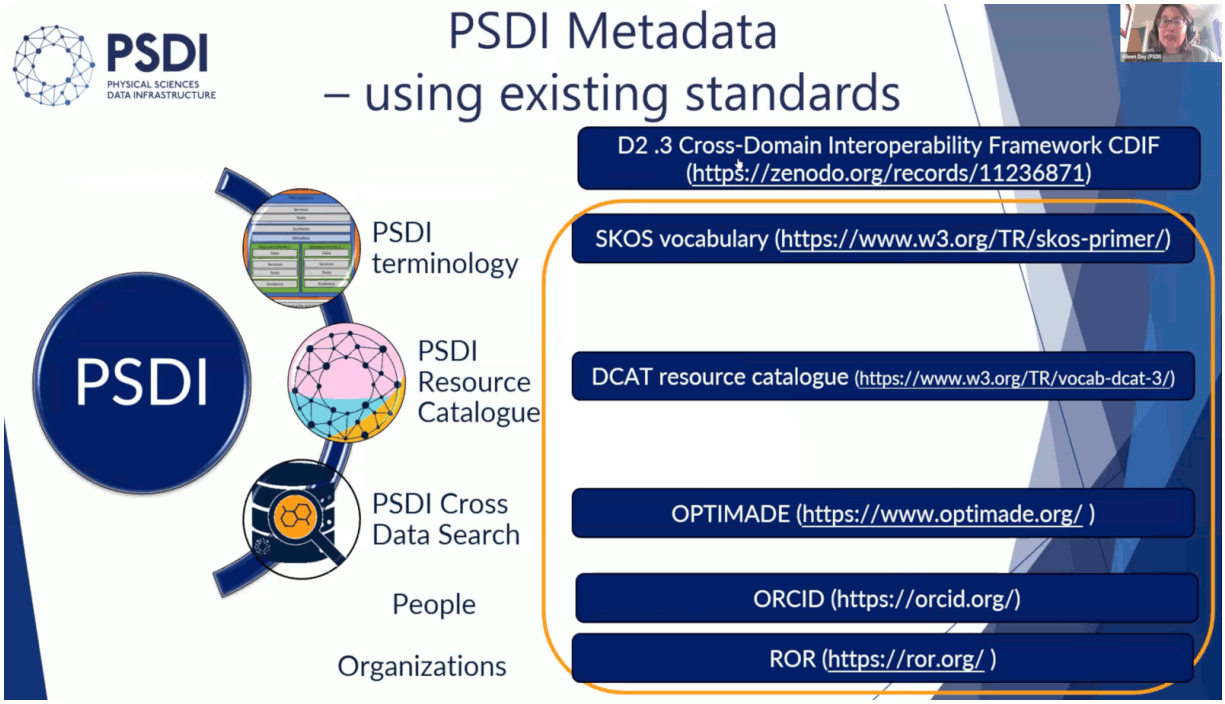
run it on your own infrastructure. And we have guidance for everything as well. The metadata supplies the information behind the “What We Provide” page, showing everything that is available on the PSDI website. And it also provides the information behind the “Cross Data Search” page, which allows searching across all of those data sources by filtering on certain properties. And so we see the PSDI metadata as being the glue that brings PSDI together. So all of this data, lots of different data types from lots of different domains, all of it can be talked to in the same language.

We're taking a top-down approach for the metadata, even though it's fairly early days. So we're starting at the top and we'll aim to go down. We've been focusing on the metadata about the resources, which are describing each donation of data or services or tools to PSDI on a top level. We've got some basic information about the properties within those resources, just enough to put the current cross data search, but we want to expand that. And in the future, we also want to have metadata about on the data level, about each data point, and metadata about the provenance. And so we've not got down to this level yet, but that's where we're going to go.

So we are trying to use existing standards where possible. So the Cross-Domain Interoperability Framework (CDIF) report came out last year and was very timely for us because it basically tells you how to structure the metadata for across the main repository, which is exactly what we're trying to do. So on the back of this, we've got a SKOS vocabulary, a very basic one that describes the top level terminology in PSDI. We've got a DCAT resource catalogue to describe what's in those resources on the top level. We went with DCAT because **CDIF said either use Schema.org or DCAT**. And we work with the Science and Technology Facilities Council in the UK STFC and they've had a lot of experience with DCAT in the past and Alejandro Gonzales Beltran, who was one of the developers of it, was working there and



DOI: 10.5281/zenodo.17949060



was involved with PSDI at the time. So the decision was already made to use DCAT when I started, and I implemented it. We're using Optimade as the basis for the cross-data search, which doesn't agree with CDIF, but is a standard for materials, crystallography data. And where possible we're capturing people with ORCIDs and organisations using ROR, which I think is very standard as well.

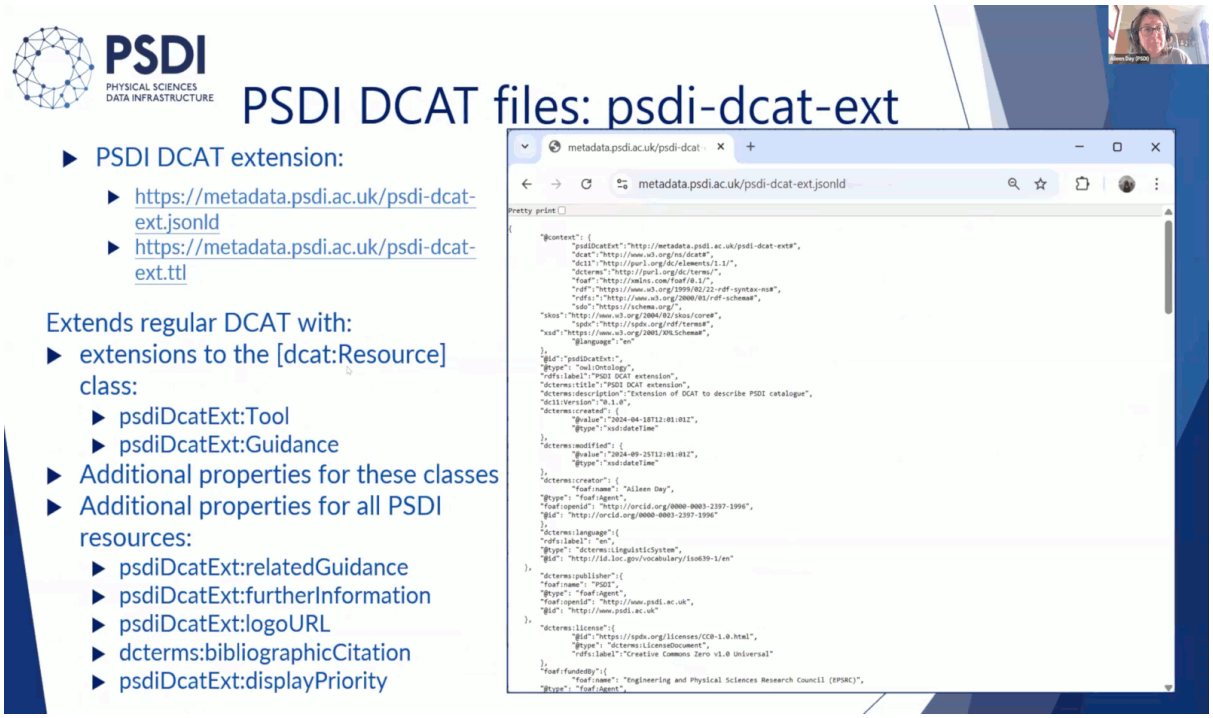
So there's two main aims of the DCAT metadata in PSDI. The first one is for discoverability and interoperability. So the machine-readable description of catalogues is out there. But it's also used to power our user interface for the pages we provide and the resource catalogue. So everything that you see in this resource catalogue is generated on the fly from the JSON-LD behind it. So we've got things gathered into those terms that I mentioned before, the resource themes and data services tools. If you click on any of those, you can see what's inside them. All of this that you see now, including the filters, is entirely based on the metadata in the DCAT behind it.

So we have the metadata published at the URL https://metadata.psdi.ac.uk/[57], which lists everything that's available in the metadata so far, but there will be more in time. And the metadata has guidance linked to it. So there's knowledge-based pages that describe it, how to, what the different fields mean and how it's structured. The bit that's relevant for this DCAT discussion here are the files that I highlight here. And I'll describe those in a bit more detail.

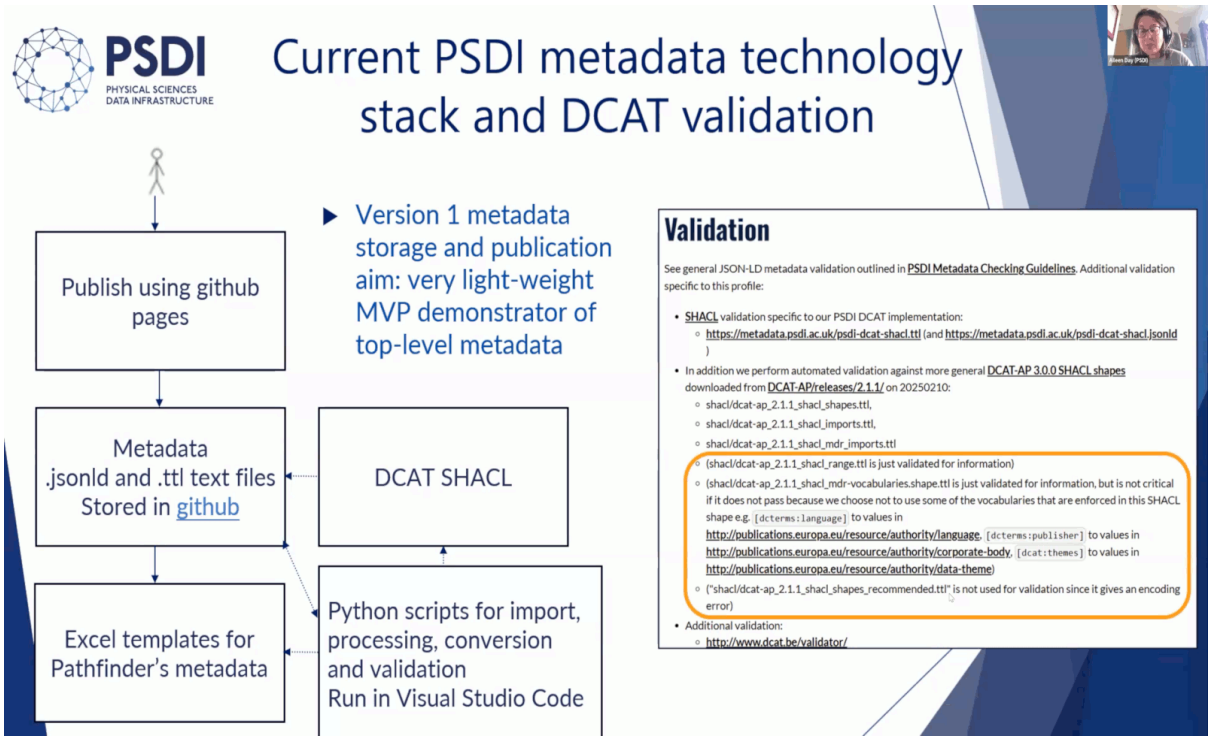
We've got a guidance document which is fairly detailed and describes how the DCAT goes structured and goes into details about particular fields in there as well The main DCAT file is available in both JSON-LD and Turtle formats. It's structured as a top-level DCAT catalogue, which also contains nested DCAT catalogues. These nested catalogues group together PSDI and resource themes, with each corresponding to a specific resource theme. Most data services, tools, and guidance are nested within these theme-specific catalogues. However, some PSDI-wide services, tools, and guidance belong to the top-level catalogue.

[57] https://metadata.psdi.ac.uk/

DOI: 10.5281/zenodo.17949060



While data and services are defined in DCAT, tools and guidance are not. To address this, we've extended the DCAT definition with the file shown here, introducing "tool" and "guidance" as two new resource types. Unlike DCAT's definition of "service" as primarily an online offering, a "tool" is something that can be downloaded, and its properties reflect this distinction. "Guidance" was also added to reflect the existing top-level structure. These new classes include extra properties, and we introduced additional properties to ensure that the entire resource catalogue could be generated from this single file, centralizing all definitions.

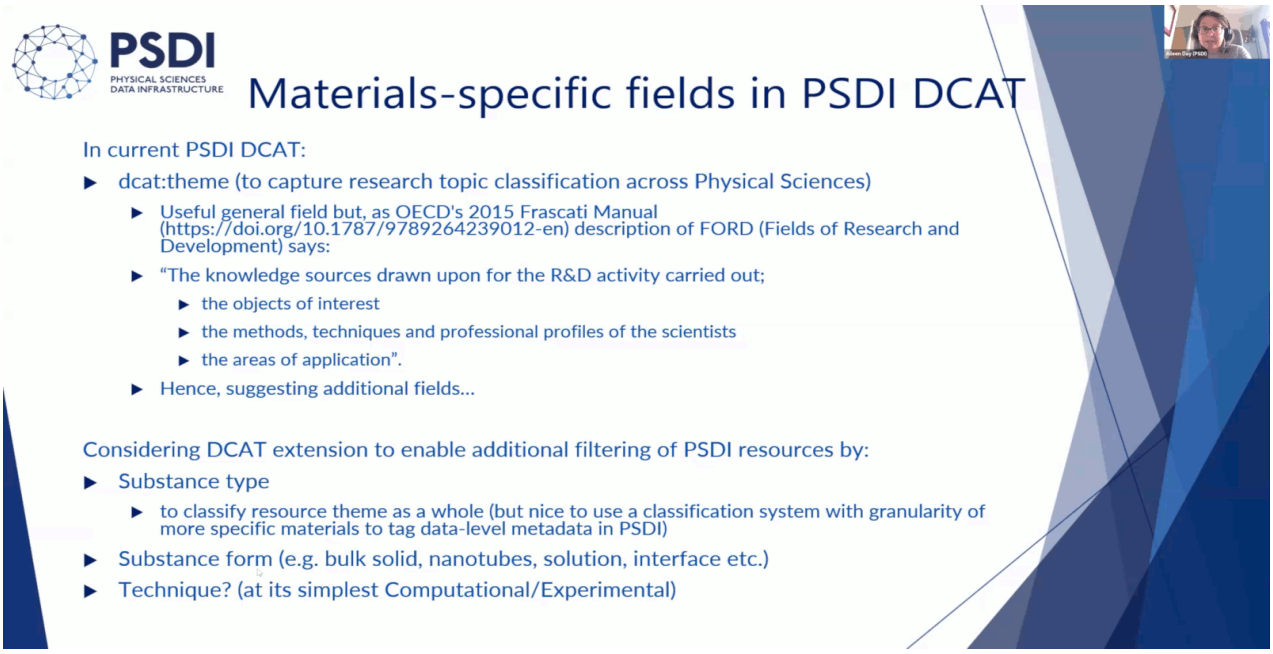


We incorporated additional properties to accommodate specific requirements and developed custom SHACL shapes for validation, given our extensions to DCAT. These shapes are

DOI: 10.5281/zenodo.17949060

currently used for validation, with future plans to leverage them as an input and editing mechanism for JSON-LD. The initial version of PSDI metadata was designed for quick deployment, relying on text files managed by Python scripts. Resource owners provide data in Excel, which is then processed into these text files, stored on GitHub, and published via GitHub Pages. We intend to enhance this system for greater extensibility as the volume of metadata grows. While we currently perform validation using our custom SHACL shapes and some generic ones, not all generic shapes were applicable to our specific DCAT implementation, which may warrant further discussion.

So what in our DCAT is specific to materials? We have the DCAT theme field in there to capture the research topic and all of the scientific discipline. We needed something that captures the whole of physical sciences because that's our scope. This is a useful field, but the difficult aspect of this is that the research theme for physical sciences is about the objects of interest such as the kind of material that you're looking at, but other times it's the method or technique, and it's sometimes the area of application. So it kind of captures three different things and each of these aspects isn't captured consistently across the datasets. That is how people focus their research interest though so it's useful to have this but we also will need to add some extra fields because of that. So the fields that we are considering for extending our DCAT is by substance type. Now, this has to be quite a general kind to classify a resource theme as a whole rather than a particular data point in it, but we would like to be able to use the same classification system when we go down to that data level metadata if we can.



Another way that people might want to find data is by substance form and so we've got few examples here and also technique. Now, this one could be quite complicated: at its simplest, we're just thinking of lacking it as computational or experimental but for each of these we're evaluating all the possible controlled vocabularies and ontology classification systems that standardize how we capture this. We are currently using EuroSciVoc and this is what the filters look like, with EuroSciVoc's most granular terms shown directly as filter categories. However, EuroSciVoc is hierarchical but that hierarchy isn't reflected in the current filtering system. This will be changed as we go, because there are too many resources to show on a single list. So we're looking into either reflecting its hierarchy in the filter list, or looking at alternatives for this.

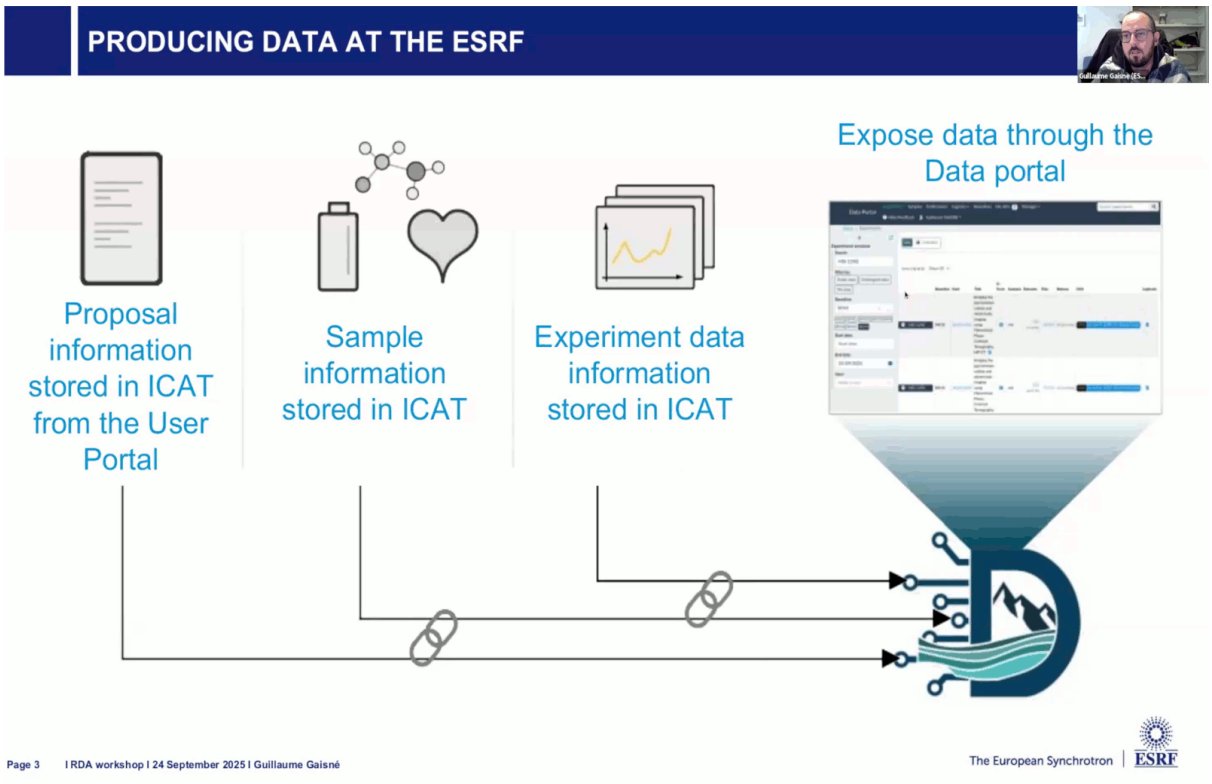
DOI: 10.5281/zenodo.17949060

Here is our list of requirements for how we are going to capture the dcat:theme. The first one is that it needs to be intuitive to allow resource owners and users to tag their data and to find other data. So it needs to be easy for people who don't know about ontologies and philosophy to get to what they need. It's got to cover all the physical sciences. It’s got to be maintained, as research constantly evolves. Nice to have requirements are permanent identifiers. If they don't have permanent identifiers, we have our SKOS vocabulary and so we can always put the terms into that SKOS vocabulary to assign them. Definitions are useful but I don’t think EuroSciVoc currently provides them. Finally, it must be hierarchical. I have been going through all of the different options considered and I'm getting very close to publishing them, along with our evaluation. To summarize our experience with DCAT in PSDI: the initial use of DCAT was led by the STFC history and as I said that's why we picked it. It was confirmed by CDIF recommendation. We found DCAT to be very well documented and supported, there's a lot of validation tools out there. I feel that we might have stretched its definition some more in PSDI. We found that making extensions were straightforward to implement and they were primarily driven by a support for the user experience and we've had to write our own SHACL to validate that. I feel that we've made it work within our environment but last week talked about compatibility with other data providers. I'm not sure whether some of our stretching of the DCAT definitions means that it's still 100% correct DCAT implementation and so I'm just going to highlight that it works for us now and I'm very welcome to take feedback about improvements that we might make.

The future development aims to refine the DCAT theme and add more of these descriptors for filtering resource themes and being able to find them. We're thinking of adding a field capture funding , which doesn't seem to be in DCAT currently. We've got the technology stack replacement and we're very happy to review this and to work collaboratively with anyone that wants to. We're talking to CDIF at the moment about ways to put this out as an example of their implementation and we're very happy to work with this group as well.

DOI: 10.5281/zenodo.17949060

Data management at the ESRF (Guillaume Gaisné)



I'm here to discuss the data management at the ESRF and to explain how we are doing it. I'm Guillaume Gaisné, I'm a data manager at the ESRF and I will make a quick tour of what I'm doing in the ESRF. The data policy document was endorsed in 2015 by the Council, giving guidance on how we deal with data and metadata at ESRF. It was upgraded in 2023 to include processed data andt this is its DOI: https://doi.esrf.fr/10.15151/ESRF-DC-1534175008[58].

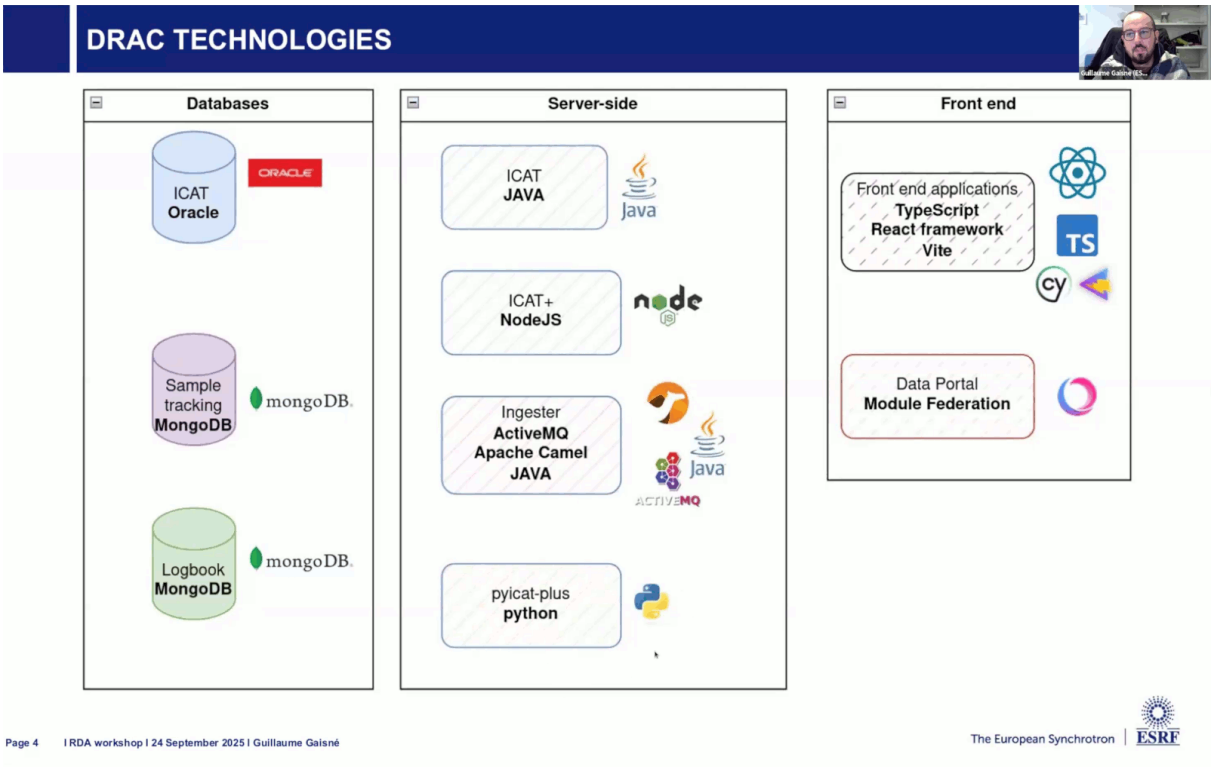
Since the endorsement of the data policy, we're developing the DRAC, the Data Repository for Advancing open sCience, that is dealing with the whole ecosystems of data and metadata. DRAC is gathering all the metadata at different steps in the data life cycle and then it shows them on the data portal. First, when a user wants to do an experiment, he goes on the user portal of the ESRF and will give some information about the proposal he wants to issue, and some information about the sample he wants to study. Then, all this information will be synchronized and stored in ICAT and the proposal information and the sample information will be linked with the experiment session ID. Once the data are scanned, the metadata and these different datasets will be also linked to the proposal information and the sample information, to be finally exposed on the internet through a front-end which allows users to find their data.

Behind DRAC there are several technologies. One of the main blocks is the ICAT database that will gather all the metadata of our datasets, plus two other small datasets: one for sample tracking and one for the log book information at the experiment level. The software layer used to discuss with the ICAT database on the server side is made using several technologies: Java will be used to discuss with ICATs but we will issue some message when a user is doing some experiment. At the end of the experimental scan, a message will be sent to what we call an ingester through different systems comprising Apache Camo, ActiveMQ and Java, and then the ingester will store all the metadata in ICAT. To retrieve the metadata from ICAT, we will

[58] https://doi.esrf.fr/10.15151/ESRF-DC-1534175008

DOI: 10.5281/zenodo.17949060

use a home-made API called ICAT+ which is made using Node.js and we also use a Python package to ingest data in ICAT. To retrieve data on the user and community side, we use the data portal using some module federation and some dedicated domain portals using different applications.



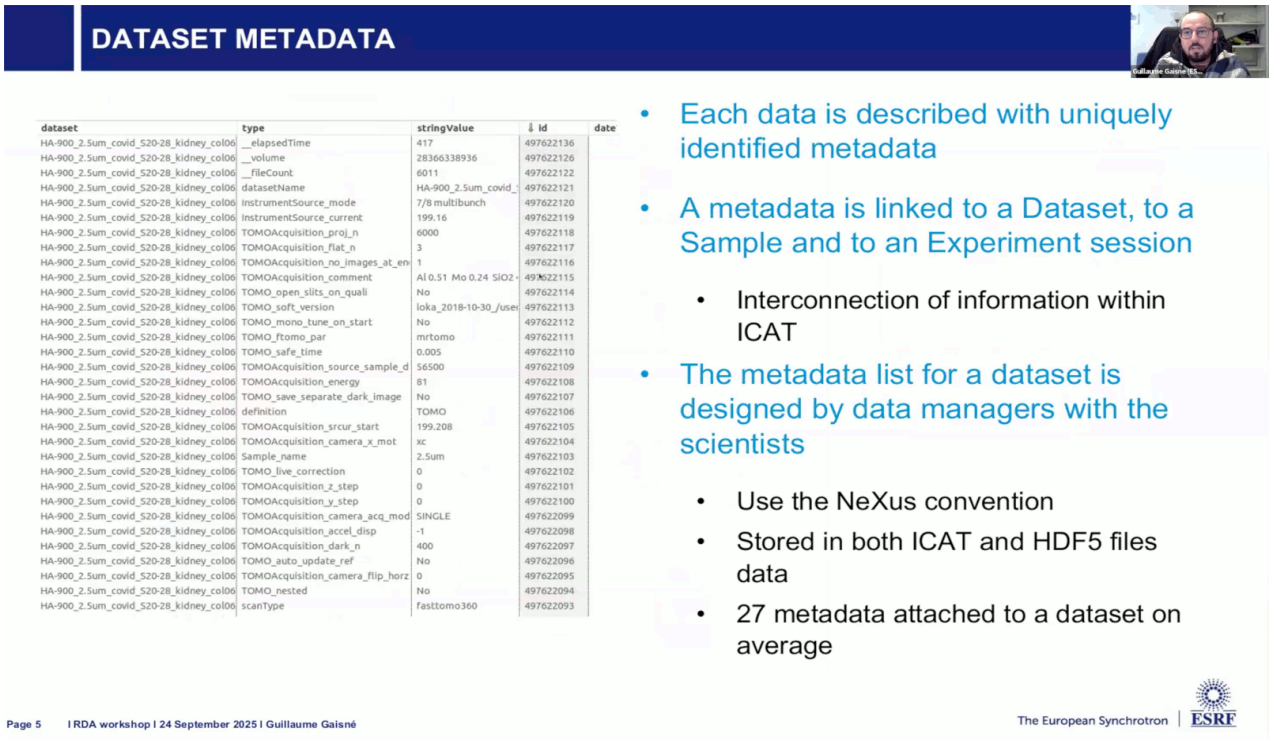
Behind every datasets we have a list of metadata that are stored in ICAT. Here you have an example of a dataset with a list of different metadata and their associated values. Each metadata here is uniquely identified with an ID and each metadata is linked to a dataset that is linked to both the sample and the experimental session through the proposal information.

The relevant metadata are selected through discussions between the data managers and the scientific team behind the experiment. We mostly use the Nexus convention to build our files. The metadata are both stored at the file level in the HDF5 format and also in ICAT. So an HDF5 file will contain the scientific data and the metadata, while ICAT will contain only the metadata related to this data. On the average ESRF dataset there are about 27 metadata, but right now as the time goes we include more and more metadata. So this average is rising slowly for new metadata.

To retrieve our data we use the data portal. All the metadata associated with the datasets will be used to search the data. Here for instance I have a screenshot where I'm searching data for the proposal MD1290 at the beamline BM18. These metadata are used to ease the search for data. We use authentication for users, so that a user can only see his data or the publicly available data. If you access the data portal without being authenticated, you will only see public data. We also have integrated the possibility to mint some DOIs on the datasets using datacite and up to now we have approximately 6 million (private) datasets available on the data portal, approximately 28 Petabytes. So this is exposing all the raw and process data of the ESRF. On top of that, we have domain portals mostly related to specific projects such as the Human Organ Atlas (HOA) and the Paleo databases. These portals are exposing the same information as the data portal, but are tailored towards the specific project, with each portal

DOI: 10.5281/zenodo.17949060

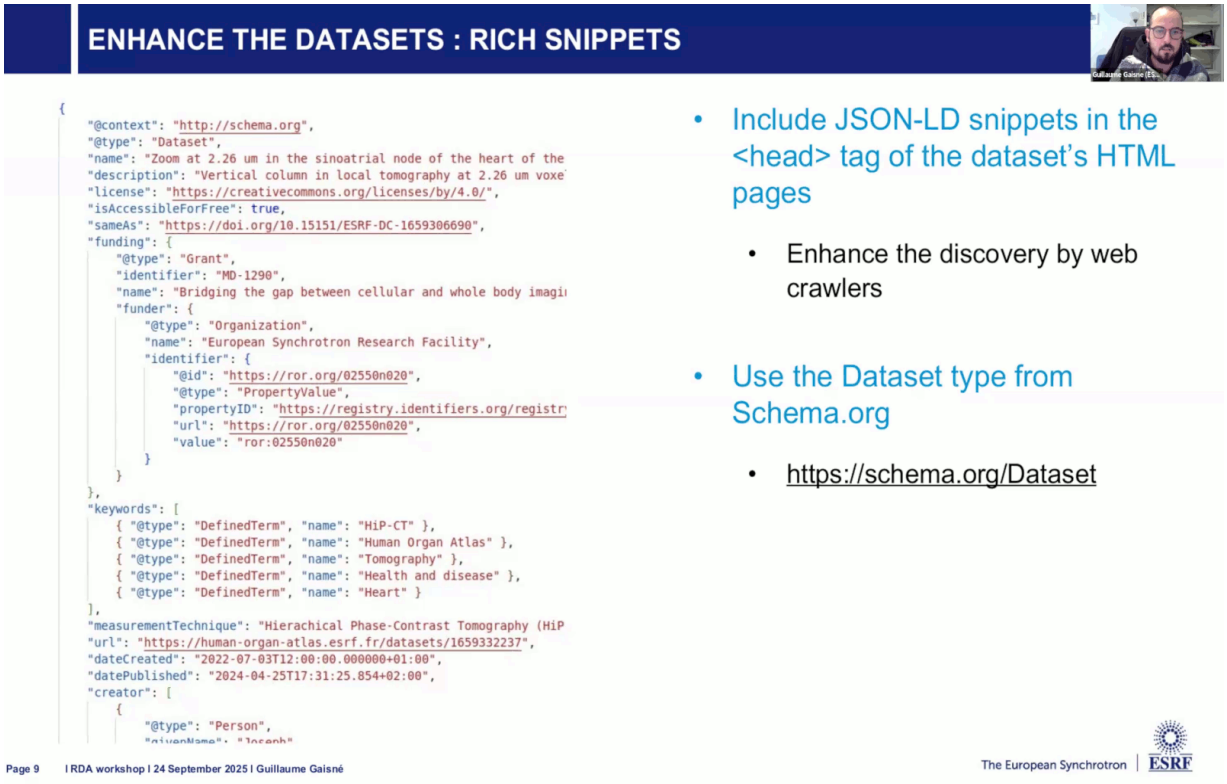
having a dedicated frontend and look. These portals provide access to the dataset DOIs, the list of users identified by their ORCID, and allow downloading the files through the HTTP protocol or the Globus software for bigger files. Also, the metadata for all datasets are grouped as JSON files or as text files, depending on the project. For the HOA, it will be JSON files. In these projects, the datasets are much more documented or have domain-specific metadata like for instance the HOA will have some metadata about the donor's medical history which is not relevant for material science for instance. So these portals allow you to see, browse, and download data from specific projects.



To enrich the metadata used at the ESRF, we are also developing an ontology covering the different experimental techniques used at the ESRF, using so-called building blocks that are like basic physical processes, or some detectors, or other properties, to semantically build what a technique is and this allows us to differentiate all the techniques that are used at the ESRF, and to include this information in the metadata to help distinguish the data even more and have a semantic meaning behind the techniques. This ontology is also linked to the PANET ontology for photon and neutron experimental techniques. However, the logic behind the PANET ontology is a bit different from the ESRF one.

The HOA and the Paleo portals use JSON-LD snippets to embed the metadata directly in the header of the HTML page of the datasets using the Schema.org notation and especially the dataset type. The metadata stored in ICAT is used to construct the JSON-LD information that is embedded in the HTML file to help web crawlers identify what the datasets are about.

DOI: 10.5281/zenodo.17949060



I am also working on the AddMorePower project[59], a European project dedicated to semiconductor characterization which aims to create FAIR data. The project leverages the NOMAD repository, which is focused on simulation and experimental data for material science.

In order to publish on NOMAD we are using the MODAs and CHADAs files, which are descriptive files that explain in detail the processes behind every experiment or every simulation carried out in the project. These documents help create a basis of metadata schema in order to publish everything in NOMAD. However, using the NOMAD software comes with some work to do since the NOMAD portal is designed to be a user-friendly interface to publish data. On our side, we had to develop new tools because initially NOMAD was dedicated to simulation data. However, the development team behind NOMAD is including more and more experimental data mostly with the Nexus standard. As a consequence, the AddMorePower project built new tools to parse our files. The insights from my work is that going for open science might be difficult, since NOMAD was originally designed for simulation data, meaning its development hasn't fully accommodated experimental data, which often has different metadata requirements. Although we utilize the Nexus standard, the specific techniques employed in our projects and the varied application definitions mean that certain metadata architecture is not currently available within the Nexus ecosystem. Consequently, we are developing our own internal metadata architecture within the files.

A significant hurdle in adopting Open Science and FAIR principles is the considerable cultural shift required from scientists. The concept of Open Science can be daunting, and its objectives are often not fully grasped. Therefore, alongside implementing FAIR principles in material science projects, we must also focus on convincing scientists of their value. This presents a challenging yet exciting endeavour.

[59] https://addmorepower.eu/

DOI: 10.5281/zenodo.17949060

Discussions

Gerhard Goldbeck: The big question is which sort of middle-to-high level metadata can describe materials science datasets. For example, Aileen started with the description for the repositories and themes in PSDI. At some point, you get more to the actual data that you have the the search I think cross repositories about substance; in comparison, the NOMAD or ESRF data portals have a huge amount of data and metadata about them, but still it is hard for example to identify all the simulations on semiconductors. When I want to explore a specific type of material or substance, I start by searching for relevant datasets. Once I find them—especially through specialized portals—they often come with rich metadata about the samples, which is very helpful. However, the process between identifying what I need and actually locating the right dataset tends to be quite challenging.

Aileen Day: I think you're referring to the cross data search, which operates directly on the data itself, whereas the resource catalogue provides a higher-level description. The challenge arises when we tag datasets in the resource catalogue—for example, labelling one as containing MOFs. It's unclear whether that tagging effectively filters down to the actual data. So when someone uses cross-data search, it becomes difficult to search at that higher level because the search dives straight into the details. We're trying to figure out how to make those higher-level tags propagate down in a way that still supports meaningful lower-level searches.

Gerhard Goldbeck: That’s why structured schemas work well on recipe websites. For example, if I only have a few ingredients, I can search for recipes that use them without needing detailed provenance data. That becomes my entry point. It’s similar to what you mentioned about MOFs. If I’m interested in MOFs or something related and want to find all relevant data, it’s currently quite difficult to do so at that level. But if we could make it work, it would be incredibly powerful.

Simon Stier: I agree this might be one of the biggest challenges within the materials science domain. It's important to recognize that schema.org was created by search engine providers primarily to support rendering structured content, not necessarily to enforce rigorous normalization of terms like ingredient names. Their focus is on just enough structure to display results properly. While Google may apply statistical techniques, such as identifying similar words or even using vector embeddings from LLMs, the underlying issue remains: we need precise identification of concepts, not just loosely matched strings from a keyword list. Our focus must be on building linked data. But one of the major hurdles is replacing raw string labels with references to well-defined entities, that is, using controlled vocabularies. Even with a perfect metadata schema like DCAT-AP, if you're relying on arbitrary keywords, it becomes difficult to search efficiently for specific material classes. For example, identifying all datasets related to semiconductors would require compiling an exhaustive list of possible string variations, which is a massive task. Ideally, we should be pointing to clearly defined concepts, and doing so through shared agreements. But if there's no consistent pointer—just another keyword or term—it becomes hard to establish meaning. That’s why one of the key responsibilities for data providers is to link their metadata to controlled vocabularies. Unfortunately, many of these vocabularies are still missing or are incomplete.

It’s clear that the European data strategy is quite diverse and includes competing approaches shaped by political and funding mechanisms. For those trying to implement or interpret data, it can be challenging to keep track of the various large-scale initiatives and the digital directions they want to pursue.Perhaps today isn’t the right time to dive into that.

DOI: 10.5281/zenodo.17949060

Federica Bazzocchi: I have a couple of questions for the ESRF speaker, as I was genuinely impressed by their research infrastructure and particularly how they manage metadata and data across multiple layers. They mentioned integrating ontologies for various experimental techniques and producing data in the NEXUS format. I'm curious whether they're proposing a new NEXUS application definition, and whether they're in contact with the NOMAD and FAIRmat groups. Do they use the NOMAD portal directly, or do they upload data to the NOMAD cloud? Additionally, I'd like to understand the level of engagement from scientists in this process. I am working on a materials science research infrastructure project in Italy and in my own experience it has been extremely difficult to engage with the scientists who generate data and do the measurements. Often, the feedback we receive is minimal, and they're reluctant to collect or share data using the tools we've developed. So I'm wondering how this works within a large and complex organization like ESRF.

Guillaume Gaisné: In our team, we have a member who is on the NEXUS advisory committee, and we're actively working on new NEXUS application definitions. For example, we use NXtress and NXtomo for tomography data. However, during my two years at ESRF, I haven't seen many new application definitions fully integrated. The process is quite slow, as each definition must be validated by the council. So while we can work on new definitions now, I doubt we'll have a new NEXUS application definition fully implemented by the end of a four-year project like AddMorePower.

At ESRF, we produce raw data in HDF5 format, and the internal structure of these files adheres to NEXUS classes. Even if the architecture doesn't yet conform to a complete application definition, we have the foundational elements in place. This means we can later adjust the files to align with a formal definition or tag them accordingly once a new definition is established.

Regarding our collaboration with NOMAD and FAIRmat, yes, we are in contact—primarily through my role in the AddMorePower project. We're actively using NOMAD, publishing data there, and working to install NOMAD Oasis instances at the project level. I'm the main person handling this. We also have connections with the SENDER Brookowser (?) and OSCARS projects, aiming to improve the NEXUS ontology and its links to PaNET and the ESRF experimental technique ontology. These integrations are progressing slowly, and since I'm not directly involved in the ontology work, I can't provide specific timelines or details.

As for scientist engagement, that's a major challenge on my side as well. When publishing data on NOMAD, it works smoothly only if the format is already recognized by the software. If not, NOMAD can't process the files, and we need to develop new tools to support those formats. This involves creating custom application definitions or metadata structures, which requires close collaboration with scientists to understand what metadata they need and how they want it organized.

In my experience, metadata doesn't naturally engage scientists. They're confident in their understanding of their own data when it is stored Ion their laptops, and often don't see the immediate value in metadata. But six months later, when they revisit old data or switch projects, it becomes difficult to recall the context. We've all been there. It's frustrating to see them disengage when metadata is discussed, even though it's crucial. If a data manager pushes for metadata use, there's often resistance. However, if the idea comes from the scientists themselves—like wanting to sort or organize their data—they're much more receptive.

To truly engage scientists, we need strong, practical examples that show how FAIR data principles can improve their daily workflows. But that's hard to achieve. I don't have a clear solution. I face the same issues—sending emails about metadata and getting no response, giving presentations and receiving little feedback. Some individuals show interest, but overall, it's a tough crowd. I'm not sure how to improve engagement, especially among those who aren't already invested in FAIR principles.

Simon Stier: NOMAD is a nice exception as it allows users to define custom processing workflows. However, this flexibility comes with the requirement that users must understand and define those processes themselves. NOMAD operates with a domain-specific schema—essentially its own language—which isn't yet widely adopted across the materials science community. That said, with the next phase of funding, there's a dedicated subtask focused on federating a catalog. The idea is to make NOMAD and its Oasis satellite instances more explorable by exposing metadata through standards like DCAT-AP. While a standardized schema hasn't been fully implemented yet, this direction points toward a more open and interoperable ecosystem. Still, this doesn't solve the core challenge: getting people genuinely interested in providing and using metadata. I mentioned this as a concern because, ultimately, we need to move away from the "black box" model—where users upload data and expect metadata to be magically extracted, enabling advanced filtering and scientific queries. That's the ideal scenario many envision, and perhaps LLMs will help us get there. But encouraging people to engage with that process—to step inside the black box—is extremely difficult. I've had similar experiences, and it remains very, very hard.

Marek Cebecauer: I just wanted to add to Simon's point: researchers need support when it comes to data management. They simply don't have the time to navigate complex systems on their own. If the tools can offer guidance or assistance, such as helping them fill in metadata correctly, then the process becomes manageable and effective. Without that kind of support, it's unlikely to succeed.

Guillaume Gaisné: One of the users in the Paleo project was quite happy to spend time documenting his data because he could easily search and retrieve it through the portal. This shows that when researchers receive meaningful assistance, which is not doing the work for them, but guiding them, they get immediate feedback and see the value of their efforts. Even if the process takes time, they recognize its usefulness. It was encouraging to witness this kind of early engagement and development.

Kwang-Ryeol Lee: I'd like to ask Guillaume and Aileen a question. From what I understand, your platform primarily manages measurement data for specific materials. However, from a materials science perspective, especially when aiming to support material design through machine learning, we need a much richer context. That includes not only measurement data, but also process data, structural data, composition data, and property data. All of these elements need to be integrated to form a coherent narrative around each material.

With that in mind, what is your strategy for extending your current data infrastructure to include process and structural information? Do you have any plans or approaches in place to integrate these different types of data into a unified framework?

Guillaume Gaisné: In the AddMorePower project, our focus is on publishing process data. This means we need to carefully document the process itself and establish links to the raw data. For the datasets I've managed to publish on NOMAD with some scientists, I've

emphasized capturing the experimental settings and the software used. However, describing the sample itself has proven to be particularly challenging.

The concept of a “sample” is surprisingly complex. In project meetings, when we raised the question of what constitutes a sample, people initially responded with confidence, but their definitions varied significantly. This highlights the need for community-level work to define and standardize what a sample is and how it should be documented.

In my day-to-day work, sample descriptions tend to be limited. Within the NOMAD environment, we typically include fewer than ten metadata fields for a sample. I’ve tried to push for including elemental composition, but this becomes difficult for e.g. semiconductors that have multiple layers. So, for now, the honest answer is: it’s hard. Coming from an astrophysics background, I’m used to dealing with samples like stars, which don’t change much over time. But since joining ESRF two years ago, I’ve realized that defining a sample in materials science is far less straightforward. It’s a discipline-specific challenge that deserves dedicated attention.

Aileen Day: I think the part we haven’t yet explored in detail is the data-level metadata. Within PSDI, we’ve already introduced some foundational elements through demonstrator projects, which touch on various aspects of what you’re describing. One key work stream focuses on electronic lab notebooks (ELNs), which I believe are essential. Researchers aren’t likely to manually populate extensive metadata fields, so capturing metadata automatically during the experiment is crucial. Another important area is workflow documentation. We’re using a pathfinder version of Galaxy to provide a service where users can capture and execute their workflows. It works primarily for computational workflows, where the system automatically records all relevant metadata, including what was done, start and end data, and other details.

Capturing workflows is vital because it ensures continuity and traceability throughout the entire research cycle: from experiment to processing to publication. While we haven’t fully integrated these components yet, it’s something we plan to do. One of PSDI’s core goals is to make data accessible and usable for AI, and linking metadata across the full lifecycle is key to achieving that.

Kwang-Ryeol Lee: In our vision, the ultimate goal of data collection is to enable material design through machine learning technologies. This makes data integration critically important in materials science. Unlike in astrophysics, where objects are relatively uniform and can be treated as similar entities, materials science deals with complex variables such as defects, grain size, and other structural factors that significantly influence material properties. To support meaningful machine learning applications, all of these aspects need to be captured within a single, comprehensive dataset. Only then can the data truly become valuable for future predictive and design tasks.

Gerhard Goldbeck: It would be helpful if you could also share your own perspectives and challenges. We've already heard some important points, for example that researchers often aren’t fully aware of which metadata they need in order to reuse their data in the future. This presents both technical and human-level challenges. One way to support better metadata practices is to identify key metadata elements from a community perspective, which can make the process more intuitive and manageable.

Cataloging is another critical area, where infrastructure providers play a central role. That brings up the question of repositories: we often hear about NOMAD, but what are the actual options for storing and sharing FAIR data in our field? The typical project response is, “We

have a data management plan and we publish on Zenodo," but that's not a complete solution. Finding and reusing data from Zenodo can be difficult. So if not Zenodo, then what? Besides NOMAD, what other repositories exist globally, and what do they actually implement in terms of metadata?

Metadata also shapes the user experience. For instance, when I visit NOMAD, I might not be interested in a specific VASP calculation with a particular basis set, but I may be interested in data about certain material class. The same applies to ESRF data: there may be vast amounts available, but I might be approaching it from a completely different angle. That's why we need to understand: what users are actually searching for and how to present that information effectively. Metadata must support these diverse search perspectives, and the first step is to gather insights into what people truly want to find.

Simon Stier: I fully agree—what counts as metadata versus data depends on the user's perspective and domain. For example, a dataset may include basic fields like name and license, which are useful for general discovery. But a chemist might want to filter by chemical composition, while a materials scientist may prioritize mechanical properties like elastic modulus. Each pushes different information into the metadata layer based on their needs.

The solution is to store data interoperably and build domain-specific indexes. From this, we can create catalogs—chemical, materials engineering, or general—each surfacing different metadata views. The boundary between data and metadata is fluid, so defining flexible profiles that adapt to different use cases is key. These profiles can apply tailored indexing to the same dataset, enabling diverse and meaningful access.

Marek Cebecauer: We had some discussions in the past involving LLMs, and I think the core issue is the need for templates, but if everyone uses a different template, it becomes problematic.

Gerhard Goldbeck: This isn't about LLMs. In practice, it's the data stewards who prepare the templates. Marek points out that in the Czech Republic people use a core metadata model already integrated into all templates. This model was developed collaboratively and is shared across the community, which ensures consistency. What Marek has done is essentially to define a national template, not just something prepared by individual data stewards. That's a great example of harmonization. It's similar to developing an application profile: you reach a shared agreement on the structure and type of information to be captured.

One of the key challenges we face is that, while there may be agreement on how things should be done, that agreement isn't widespread. It's great to hear that in the Czech Republic this level of alignment was achieved and that the templates are actively used along with LLMs. But we don't have that kind of harmonization across Europe, certainly not globally, and not within the materials science community.

That's why I found the GeoDCAT-AP example so interesting, because it shows that the application profile wasn't something you had to build from scratch. You already had the agreement; it just existed in a different format. Meanwhile, many of us are still trying to reach even a basic consensus on standards. In my view, chemistry has made more progress in this area than other disciplines. Jakub, can you say more from your community, especially regarding the control vocabularies? How does the harmonization process actually work?

Jakub Klímek: There are multiple layers to harmonization: one involves aligning terms within the application profile, and another focuses on harmonizing controlled vocabularies. In both

cases, it's about bringing the community together to reach consensus. If a term is missing, someone might propose adding it, while others may point out an existing equivalent that would suffice. The key is identifying stakeholders, initiating a proposal, and starting the discussion. Of course, resource constraints are always a challenge, but that's true in all communities.

Simon Stier: I think the bio domain has already made some progress. Fundamentally, everyone needs to expose their schemas in a programmatic way. Many applications, like eLabFTW[60], allow users to tweak schemas and define additional properties, but they lack mechanisms to publish and share these schemas for use by other research software. That's a major technical gap.

Before we can reach any agreement, we need to share our definitions. For example, if I create a tensile test template and push it to a repository, others can discover it, agree or disagree, and that sparks progress. Ideally, a domain scientist could say, "I want to document this in my lab," and a smart search tool would suggest existing templates, perhaps one used in 10,000 publications, that match their needs. They wouldn't need to start from scratch.

This is how things should work: not through global enforcement, but through open repositories and search tools that help surface well-established schemas. Over time, popular schemas would gain traction through community adoption: a democratic approach to standardization.

Gerhard Goldbeck: I'd like to thank you all for participating, and take a moment to share a few words about what's currently underway and what lies ahead. Otello has been supporting this effort through a grant from the RDA, specifically working with the RDA Working Group focused on the topics we've discussed today. The scope of that group has been slightly broader, covering harmonization and FAIR semantic resources, including the user journeys I mentioned earlier.

The Working Group is scheduled to conclude this year, with a short extension granted for the final report, but it will definitely wrap up within the year. As part of our recommendations, we're currently discussing the need for a follow-up focused on developing an agreed schema approach and harmonization strategy, which is still missing in the materials science domain. This next phase may involve different application examples and will be part of our final recommendations.

The upcoming RDA Plenary will take place in Brisbane, but it is also possible to attend online. There will be a dedicated session from this Working Group, where we'll reflect on today's discussion and other related topics. Please stay connected with the work we're doing, and we hope you'll join us at the next plenary session.

[60] https://www.elabftw.net/