Joint Lab "Model and Data-driven Materials Characterization" (JL-MDMC)
"Nanoscience Foundries and Fine Analysis" (NFFA)-Europe Pilot (NEP)

The MDMC-NEP Glossary of Terms

Rossella Aversa[1,*], Alexey Boubnov[2], Dario De Angelis[6], Catriona Eschke[3], Sarah Irvine[4], Reetu Elza Joseph[1], Maximilian Kabbe[5,6], Neil MacKinnon[5], Irene Modolo[7], Mirco Panighel[7], Richard Thelen[5], Davide Valentinis[8,9]

* Corresponding author: rossella.aversa@kit.edu
[1] Karlsruhe Institute of Technology, Scientific Computing Center (SCC), Kaiserstraße 12, 76131 Karlsruhe, Germany
[2] Karlsruhe Institute of Technology, Institute of Nanotechnology (INT), Kaiserstraße 12, 76131 Karlsruhe, Germany
[3] Helmholtz-Zentrum Hereon, Institute of Metallic Biomaterials, Max-Planck-Straße 1, 21502 Geesthacht, Germany
[4] Helmholtz-Zentrum Hereon, Institute of Materials Physics, Max-Planck-Straße 1, 21502 Geesthacht, Germany
[5] Karlsruhe Institute of Technology, Institute of Microstructure Technology (IMT), Kaiserstraße 12, 76131 Karlsruhe, Germany
[6] Karlsruhe Institute of Technology, Karlsruhe Nano Micro Facility (KNMFi), Kaiserstraße 12, 76131 Karlsruhe, Germany
[7] CNR-IOM - Istituto Officina dei Materiali, National Research Council of Italy, Strada Statale 14, km 163.5, 34149 Basovizza (Trieste), Italy
[8] Karlsruhe Institute of Technology, Institute for Quantum Materials and Technologies (IQMT), Wolfgang-Gaede-Str. 1, 76131 Karlsruhe
[9] Karlsruhe Institute of Technology, Institute for Theory of Condensed Matter (TKM), Kaiserstraße 12, 76131 Karlsruhe, Germany

Introduction

The Glossary of Terms (Glossary hereafter) is a result of the Metadata Working Group [1], which involves members of the "Nanoscience Foundries and Fine Analysis" (NFFA) EUROPE Pilot (NEP) [2] and of the Joint Lab "Integrated Model and Data Driven Materials Characterization" (JL-MDMC) of the Helmholtz Association [3].
This Glossary aims at describing at a high level both experimental (Figure 2) and computational (Figure 3) workflows, framed in the existing or planned management infrastructure of the involved projects, and at representing the provenance information.
The included terms are intended to reflect the lifecycle of entities and data collected in nanoscience and materials science research studies, from the fabrication of a material to the scientific publication (as overviewed in Figure 1), and then archived for the purposes of further data discovery and data sharing (Figure 6).
The workflows (Figure 3, Figure 4) and lifecycles (e.g., Figure 4) are necessarily idealized and simplified, but still give a contextually sufficient overview to frame the key concepts. As a next step, it is planned to extend the Glossary at a more fine-grained level, in order to

provide an extensive middle-layer model to describe each of the processes and entities involved: as an example, the description of an experimental sample is currently under development. The creation of an ontology is envisaged, too: preliminary tests have been already performed.
As an application use case, the terms of the Glossary will be soon published as RDF data, using the Simple Knowledge Organization System (SKOS) model, in Skosmos [4], a web-based service which allows users to search, browse, and assign vocabularies with unique identifiers (URIs). This way, the terms can be individually integrated into metadata schemas or used in other applications.
The terms in this Glossary can be adopted by other initiatives focused on materials science. This will have the huge advantage of having a basic common description of entities and processes in the materials science domain, offering a set of metadata which, in turn, will increase the interoperability and the reuse of research data by different communities.

The terms

The Glossary is intended as a living document, which can be constantly updated. In the current version, it contains 45 terms considered relevant by the members of the Metadata Working Group. Whenever possible, the assigned terminologies and agreed definitions were adopted from existing ones or adapted to fit the needs of the involved communities.
In addition to the well-known high-level schemas, such as crossref [5], Dublin Core [6], DataCite [7] and schema.org [8], we found particularly relevant the NFFA glossary [9], which in turn uses some terms from the Core Scientific Metadata Model (CSMD) [10, 11], the CASRAI Research Data Management terminology (no longer available, see [12]), the vocabulary of terms used in NeXus [13] files, the National Institute of Standards and Technology (NIST) Materials Resource Registry Vocabulary (NMRRVOCAB) [14, 15], the Elementary Multiperspective Material Ontology (EMMO) [16], and the ontology Metadata4Ing [17]. For some terms, especially related to the computational workflow, we consulted Wikipedia [18], the Materials Design Ontology (MDO) [19] and the Physics Subject Headings (PhySH) [20].
The definitions, reported in this section, reference and use in a consistent manner other Glossary terms, which are written in **bold** with Capital Initial Letters.

Analyzed Data
Specific type of **Research Data**, primary output of any kind of **Data Analysis** performed on **Research Data**, typically on **Processed Data**.

Calculation
Computational **Data Acquisition** performed on a **Model** to process its input **Settings** into output calculated properties using a specific computational and/or theoretical **Technique** based on a theory accepted by the community (e.g., Density Functional Theory, Conformal Field Theory).

Conclusions
Primary output of **Data Interpretation** performed on **Research Data**, typically **Analyzed Data**. **Conclusions** are any kind of insight that support the answer to some specific research question, such as the significance and implications of the research findings of a **Study**, possibly in comparison with **Reference Data**, along with recommendations which may support decision-making about the next processes of a **Study** or about future work. **Conclusions** form an important part of a **Study** debrief and are usually reported in **Scientific Publications**.

Consumable
Auxiliary entity used during **Fabrication**, **Sample Preparation** or **Measurement** which has a limited time capacity or is limited in its number of uses before it is disposed of, is necessary to the process itself and normally acquired from third party manufacturers. Examples are: gloves, syringes, wipes, etching solutions, glass slides, spatulas, weighing paper, two-sided tape.

Correlative Characterization
Action of characterizing and connecting the different types of information from co-referenced (in time or space) multimodal **Research Data** obtained using different **Techniques**. This may include the output of multiple **Data Acquisitions** and/or of any of the processes included in the **Data Analysis Lifecycle** to obtain complementary insights on a region of interest, as well as to put into relation features of different **Systems** across multiple length scales over time.

Data Acquisition
Set of actions carried out by one or more **Research Users**, performed on a **System** or a set of them to generate a single self-consistent unit of **Raw Data** using a **Technique**, an **Instrument** and other **Equipment** under constant or varying controlled conditions described by **Settings**, depending on the particular research context. **Data Acquisition** may be an experimental (**Measurement**) or a computational (**Calculation**, **Simulation**) process. **Data Acquisition** is specific to **Technique**: an investigation on the same **System** conducted using a different **Technique** implies a different **Data Acquisition**. The output of **Data Acquisition** is **Raw Data**.

Data Analysis
Set of actions included in the **Data Analysis Lifecycle** and performed by one or more **Research Users** on **Research Data**, typically **Processed Data**, to extract insights that support the answer to some scientific research question (i.e., **Conclusions**). **Data Analysis** may include: linear combination fitting, least-squares curve fitting, data modelling, pattern extraction and/or segmentation. The output of **Data Analysis** is **Analyzed Data**.

Data Analysis Lifecycle
Set of processes carried out by one or more **Research Users**, performed on **Research Data** using one or more **Techniques** and/or **Research Software** in order to produce synthesized knowledge (e.g., to detect patterns, determine relationships, develop explanations, test hypotheses and/or prove theories) and to eventually suggest the **Conclusions** of a **Study**. **Data Analysis Lifecycle** includes (but is not limited to): **Data Processing**, **Data Analysis** and **Data Interpretation**. These processes may be iterative and may be combined in chains or workflows.

Data Collaboration Platform
Operational information system which allows **Research Users** to keep their **Research Data**, **Datasets** and related documents (e.g., drafts of **Scientific Publications**) synchronized and up-to-date, and to exchange them with other **Research Users**, who are typically members of the same **Project**. The system is intended for the long-tail and still volatile data, which can change and are still subject to active research. Therefore, a **Data Collaboration Platform** offers versioning of all ingested files but does not usually assign **Persistent Identifiers** to them.

Data Interpretation
Set of actions, included in the **Data Analysis Lifecycle**, performed by one or more **Research Users** on **Research Data**, typically **Analyzed Data**, to determine the **Conclusions** of the **Study**, possibly in comparison with **Reference Data**. **Data Interpretation** supports decision-making about the next processes of the **Study** or about future work.

Data Processing
Set of actions, included in the **Data Analysis Lifecycle** and performed by one or more **Research Users** on **Research Data**, typically **Reference Data** or **Raw Data**, to prepare it for one or more further processes, e.g., **Model Preparation**, **Data Acquisition** (in case of **Calculations** or **Simulations**), **Data Analysis** and/or **Data Interpretation**. **Data Processing** usually consists of routine actions. It may include: filtering, denoising, transformation, fusion or compression of **Reference Data**, as well as calibration, normalisation, statistical data reduction, background subtraction and/or correction of artefacts. The output of **Data Processing** is **Processed Data**.

Data Repository
Information system used to store, manage and provide access to digital resources, following a set of rules that define storage and access norms. A **Data Repository** is particularly suitable for **Research Data** (especially **Datasets** and/or **Publication Data**) which are not likely to be altered again. Many **Data Repositories** automatically assign globally unique **Persistent Identifiers** to deposited resources. **Data Repositories** may be associated with an **Institution** or a group of them, with an **Instrument** or a group of them, or with a **Technique** or a group of them, or may be run by a third party. **Data Repositories** may or may not be directly used by **Research Users**.

Dataset
Collection of scientifically related (depending on the research context) **Research Data**, along with their respective descriptive **Metadata**, typically stored in a **Data Collaboration Platform** and/or in a **Data Repository**. A **Dataset** may consist of other **Datasets**. The components of a **Dataset** remain individually identifiable.

Equipment
Any kind of physical or virtual item, device, machine or other tools used to perform one or more **Fabrication**(s), **Sample Preparation**(s), **Model Preparation**(s), **Data Acquisition**(s) and/or any of the processes included in the **Data Analysis Lifecycle**. Usually, the **Equipment** is located in a **Laboratory** hosted by an **Institution** and/or can be virtually or remotely accessed. **Equipment** is usually an investment. According to this definition, an **Instrument** is a particular type of **Equipment**.

Fabrication
Set of actions (physical changes or chemical reactions) carried out by a commercial enterprise, one or more **Research Users** or a third party, and performed on one or more **Inputs** to produce one or more **Precursors** under controlled conditions described by **Settings**. **Fabrication** may require the use of **Equipment**, **Consumable**(s) and **Instrument**(s). A **Data Acquisition** may be performed during the **Fabrication**, e.g., to characterize the intermediate stages and/or the final resulting **Precursor**(s). The output of **Fabrication** is one or more **Precursors**.

Input
Physical **System** (typically a piece of material) which undergoes a **Fabrication**.

Institution
Hierarchical entity which hosts one or more **Laboratories**.

Instrument
Physical or virtual identifiable piece of **Equipment** used to perform a **Data Acquisition** and to generate **Raw Data**. The **Instrument** is located in a **Laboratory** hosted by an **Institution** and/or can be virtually or remotely accessed. A virtual **Instrument** may be any computational resource or HPC infrastructure (cloud infrastructure or supercomputer) needed to perform **Calculations** or **Simulations**.

Laboratory
Physical or virtual place hosted by an **Institution**, where one or more **Instruments**, as well as the **Equipment**, are located and/or can be virtually or remotely accessed, and the **Data Acquisition** may be performed.

Measurement
Experimental **Data Acquisition**, typically performed on a **Sample** using an experimental **Technique**. It may also be performed during **Fabrication** or **Sample Preparation**, e.g., to characterize the intermediate stages and/or the final resulting **Precursor**(s) or **Sample**(s), respectively. A **Measurement** may require the use of **Consumables**.

Metadata
Any descriptive data intended to contextualize or otherwise qualify **Research Data** and/or **Datasets** and/or **Publication Data** and their management through time. Depending on the mode of use, **Metadata** contains information pertaining to any aspect of the **Study**, including (but not limited to) processes, outputs, and **Research Users** involved in the **Project**. **Metadata** may include descriptions of how files are named, structured and stored. **Metadata** may be registered in a **Metadata Repository**.

Metadata Repository
Information system used to store, manage and provide access to **Metadata**, following a policy or a set of rules that define storage and access norms. **Metadata Repositories** may be associated with an **Institution** or a group of them, or may be run by a third party. **Metadata Repositories** may or may not be directly used by **Research Users**.

Model
Digital representation of a **System**, primary output of any kind of **Model Preparation**, aimed to be used in **Calculation**(s) or in **Simulation**(s) for its description or for predictions of its behaviour. A **Model** represents the **System** by direct similitude (e.g. small-scale replica) or by capturing in a logical framework the relations between its properties (e.g. mathematical **Model**). A **Model** typically consists of **Settings** which may be stored in a file.

Model Preparation
Set of actions carried out by one or more **Research Users** and performed on **Research Data** (including collection and **Data Processing** of **Reference Data**) to define and/or formulate a **Model**. **Model Preparation** may require the use of **Equipment** and **Instrument**(s). The output of **Model Preparation** is **Model**.

Persistent Identifier
Long-lasting reference to a digital resource which provides the information required to reliably identify, verify and locate **Research Data** (typically **Datasets** or **Publication Data**) or **Scientific Publications**.

Precursor
Physical **System** (typically a piece of material) which is formed or manufactured during the **Fabrication** and is used during the **Sample Preparation** to produce a **Sample**. It may include one or more substrates, layers, masks, evaporation materials, coatings and/or molecules. A single **Precursor** might itself become the only **Sample Component** of a **Sample** in case it undergoes a **Measurement**.

Processed Data
Specific type of **Research Data**, primary output of any kind of **Data Processing** performed on **Research Data**, typically **Raw Data** or **Reference Data**. **Processed Data** is usually an intermediate result, to be used as input of one or more further processes, e.g., **Model Preparation**, **Data Acquisition** (in case of **Calculations** or **Simulations**), **Data Analysis** or **Data Interpretation**.

Project
Enterprise (potentially individual but typically collaborative) of one or more **Research Users**, planned to perform one or more **Studies**.

Publication Data
Dataset(s) generated in the course of a **Study**, that has undergone quality assessment and can be referred to as citations (i.e., a **Persistent Identifier** is assigned to it), e.g., to validate the results and/or the **Conclusions** presented in a **Scientific Publication** or appearing in it. **Publication Data** may include any kind of **Research Data**, as well as the relevant **Metadata** about the actions performed. **Publication Data** may be attributed to some or to all the **Research Users** who are members of the **Project**.

Raw Data
Specific type of **Research Data**, primary output of a **Data Acquisition** performed on a **System**, before any subsequent **Data Processing**.

Reference Data
Any **Research Data** not produced during the current **Study**, which is reused during the **Study** (e.g., during the **Model Preparation**) or is used as reference to compare and/or to validate the outputs of the **Study**, typically during the **Data Analysis Lifecycle**.

Research Data
Data collected, created or examined by one or more **Research Users** to be analyzed or considered as a basis for reasoning, discussion or calculation in a research context, with the purpose of generating, verifying and validating original scientific claims that support the answer to some specific research question (i.e., **Conclusions**). Examples of **Research Data** include files containing the **Settings** of a **Model**, as well as any digital resource input or output of **Data Acquisition**, **Data Processing** or **Data Analysis**. According to this definition, **Raw Data**, **Processed Data**, **Analyzed Data** and **Reference Data** are particular types of **Research Data**. **Research Data** is typically in the form of a data file, but it may potentially be a data stream or any other form of data which is relevant in a particular data management

context. **Research Data** may be described by **Metadata** and may be stored in a **Data Collaboration Platform** and/or in a **Data Repository**. **Research Data** may be part of a **Dataset**.

Research Software
Any software used to process, analyze or visualize **Research Data** (including data rendering and/or plotting). Depending on the research context, **Research Software** can be used during **Model Preparation**, **Data Processing**, **Data Analysis** or **Data Interpretation**. Any software used during **Fabrication**, **Sample Preparation** or **Data Acquisition** is considered part of the **Instrument** and should be described as such.

Research User
Person, usually member of a **Project**, who conducts any part of the **Study**, in order to collect and/or analyze **Research Data**, or is interested in reusing **Research Data** by a third party (e.g., **Reference Data**) with the final aim to extract insights that support the answer to some specific research question (i.e., **Conclusions**). **Research Users** may be assigned with a role (data curator, instrument scientist, team leader, team member).

Sample
Physical **System** (typically a piece of material) composed of one or more **Sample Components**, exposed to the **Instrument** during a **Measurement**, typically after a **Sample Preparation**. **Sample** may be held by a **Sample Holder** and/or carried by a **Sample Carrier** during the **Measurement**.

Sample Component
Physical **System** (typically a piece of material) which constitutes a part of a **Sample**. It may include, e.g., one or more substrates, layers, masks, embedding or filler or evaporation materials, coatings, conducting powders and/or molecules.

Sample Carrier
Piece of **Equipment** used for carrying one or more **Samples** and/or one or more **Sample Holders** which is helpful, e.g., for referencing, handling or height adjustment. **Sample Carrier** may be, e.g., a naked wafer, a glass slide or an individually designed metal frame.

Sample Holder
Piece of **Equipment** that makes one or more **Samples** accessible for a **Measurement**, or holds them in place in the pre-defined position to be mounted inside the **Instrument** (e.g., glass slide, TEM grid or tilting support). **Sample Holder**(s) may be carried by a **Sample Carrier**.

Sample Preparation
Set of actions (physical changes or chemical reactions) carried out by one or more **Research Users**, performed on (or between) one or more **Precursor**(s) or **Sample**(s) to produce one or more **Samples** and/or to make the **Sample**(s) fit to perform a **Measurement** under controlled conditions described by **Settings**. **Sample Preparation** may require the use of **Equipment**, **Consumable**(s) and **Instrument**(s). A **Measurement** may also be performed during the **Sample Preparation**, e.g., to characterize the intermediate stages and/or the final resulting **Sample**(s). The output of **Sample Preparation** is one or more **Samples**.

Settings
Set of configuration parameters which may be involved, for example, in a **Data Acquisition** (e.g., **Settings** of the **Instrument**), in any of the processes included in the **Data Analysis Lifecycle** (e.g., **Settings** of the **Research Software**), or to describe a **Model** (e.g., by specifying the type of solver used).

Scientific Publication
Any of the following contributions, peer-reviewed or not: article in a scientific journal (and related supporting information), monograph, book or book chapter, conference proceedings and "grey literature" (informally published material not having gone through a standard publishing process, e.g., reports and highlights). A **Persistent Identifier** may be assigned to them. **Scientific Publications** typically report the **Conclusions** of a **Study** and may be supplemented by **Publication Data**. **Scientific Publications** may be attributed to some or to all the **Research Users** who are members of the **Project**.

Simulation
Computational **Data Acquisition** performed on a **Model** to manipulate its **Settings** using a specific computational and/or theoretical **Technique** in order to study, predict or optimize the behaviour and performance of existing or proposed features and properties of a physical **System** that would otherwise be too complex, too large/small, too fast/slow, too dangerous, unaccessible, or unacceptable to engage or control. Examples of **Simulations** are: multiscale simulation, finite-element simulation, molecular dynamics simulation, discrete dislocation dynamics simulation.

System
Physical or digital entity or set of entities with distinctive properties (structural, chemical, dimensional, functional or others) which is the subject of one or more actions or investigations. According to this definition, **Input**, **Precursor**, **Sample**, **Sample Component**, and **Model** are particular types of **System**.

Study
Set of all the processes and activities performed by one or more **Research Users**, who are part of the same **Project**, with the purpose of verifying, falsifying or establishing the validity of a hypothesis and supporting the answer to some scientific research question (i.e., **Conclusions**). The output of a **Study** is usually reported in one or more **Scientific Publications** and may be supplemented by **Publication Data**.

Technique
Any experimental, theoretical or computational method used during **Data Acquisition** or during any of the processes included in the **Data Analysis Lifecycle** to acquire, process or analyze **Research Data** about a **System** or a set of them with an **Instrument**.

The relations: graphical representation

The terms of the Glossary are related to each other at different levels and can be grouped focusing on specific aspects. In the following graphical representations, each pair of related terms is connected by a line; for the sake of simplicity, the arrow represents one relation (e.g., "is part of") while the inverse one (e.g., "has part") is omitted. Further details on the relevant related terms are included in the definitions, and an explicit visualization of all the properties is beyond the purpose of this document.

Being aware that the overall picture and the formal representation should be provided by an ontology, for the current visualization purposes the different conceptual blocks are reported separately, to highlight the different thematic parts with the due level of detail.

Overview

Figure 1 gives an overview of the Glossary. A Study is performed by one or more Research Users, who are part of a Project. The Study consists of one or more processes, which are: Fabrication, Preparation (Sample Preparation or Model preparation, in case of experimental or computational workflow, respectively), Data Acquisition (Measurement in case of experimental workflow, Calculation or Simulation in case of computational workflow), and the Data Analysis Lifecycle, which can include Data Processing, Data Analysis and/or Data Interpretation. Each process consists of a set of actions. The outputs of a Study are Conclusions, published on Scientific Publications, and Research Data, which can be collected in one or more Datasets and published as Publication Data supporting the Scientific Publications. Different types of Metadata describe both the Scientific Publication and the Publication Data (including Research Data and Datasets individually).

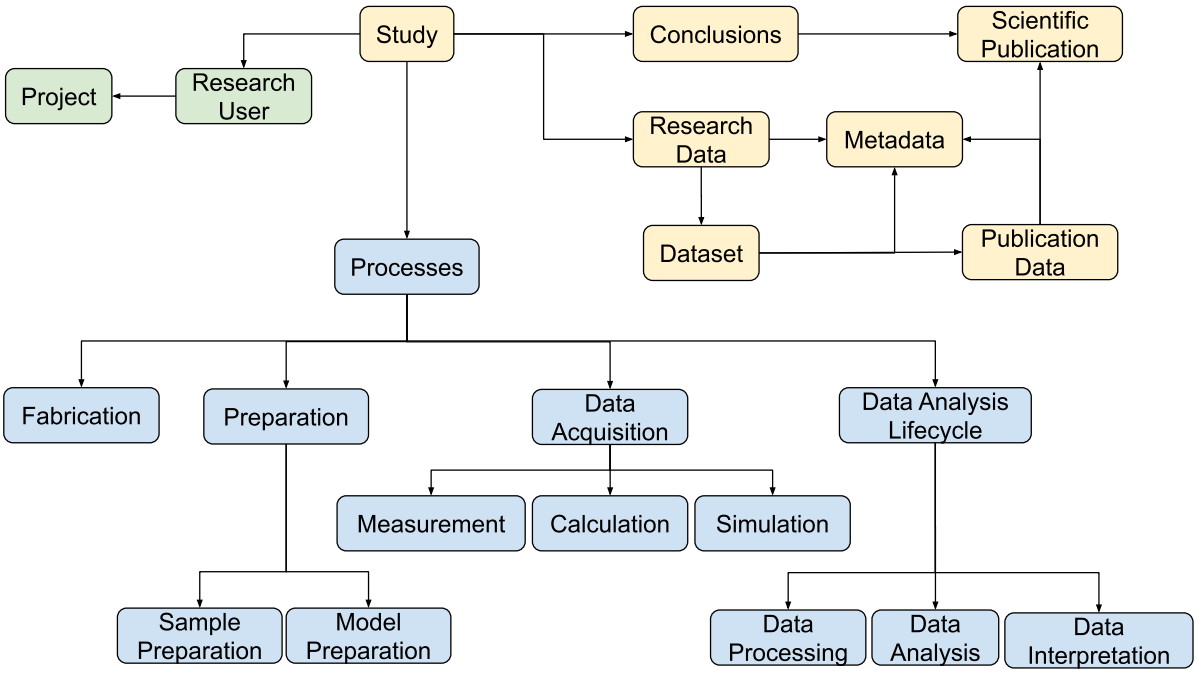

Figure 1: Overview of the Glossary. Green boxes: agents, yellow boxes: entities, blue boxes: processes.

Experimental workflow

Figure 2 shows a basic experimental workflow, in which each process is performed only once, using one Technique. Multiple loops are possible, even on more than one System. Not necessarily all the illustrated processes apply to all cases. As a generic experimental workflow, one or more Inputs undergo a Fabrication to produce one or more Precursors, which undergo a Sample Preparation to produce one or more Samples. The output Sample, consisting of one or more Sample Components, undergoes a Measurement and Raw Data is produced. In any of these processes, the use of Consumables may be needed. A Measurement can also be performed during Fabrication or Sample Preparation, using a Technique, which requires an Instrument and other Equipment, located in a Laboratory hosted by an Institution. The Settings describe any constant or varying controlled conditions of the Instrument, the Equipment, or the general environment in which the processes take place.

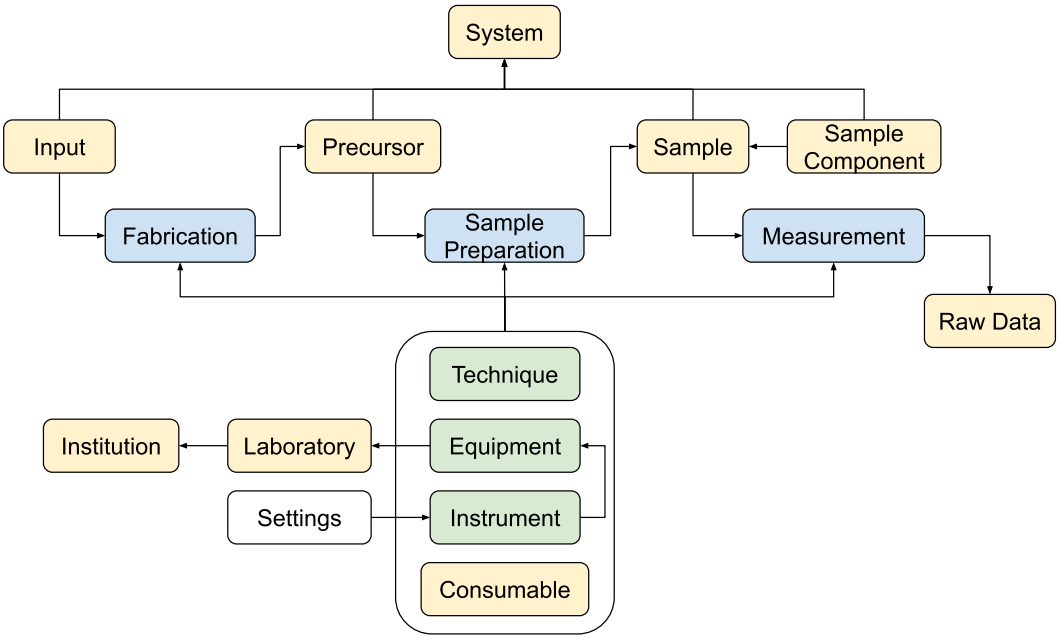

Figure 2: Experimental workflow. Green boxes: agents, yellow and white boxes: entities, blue boxes: processes. The rounded-corner box groups the terms connected to each of the processes, to simplify the graphical representation.

Computational workflow

Figure 3 shows a basic computational workflow. Multiple loops are possible, as well as different combinations of Research Data, Systems and Techniques. As above, not necessarily all the illustrated processes apply to all cases. As a generic computational workflow, Research Data undergoes Data Processing in order to obtain Processed Data, which is then used to perform the Model Preparation. The resulting Model, optionally together with Processed Data, is the input of any type of Data Acquisition, which may be a Simulation or a Calculation, and Raw Data is produced. During any of these processes, a Technique is used, which requires a (usually virtual) Instrument (e.g., a computational resource or an HPC infrastructure) and possibly other Equipment, located in a Laboratory hosted by an Institution, and/or virtually or remotely accessed, if the Laboratory is a virtual place. Research Software is additionally included in Data Processing and Model Preparation, while it is not in Simulations or Calculations: any software used during Data Acquisition is considered part of the Instrument and should be described as such. The Settings describe any constant or varying controlled conditions of the Instrument, the Equipment, and the Research Software (whenever relevant) or the general environment in which the processes take place.

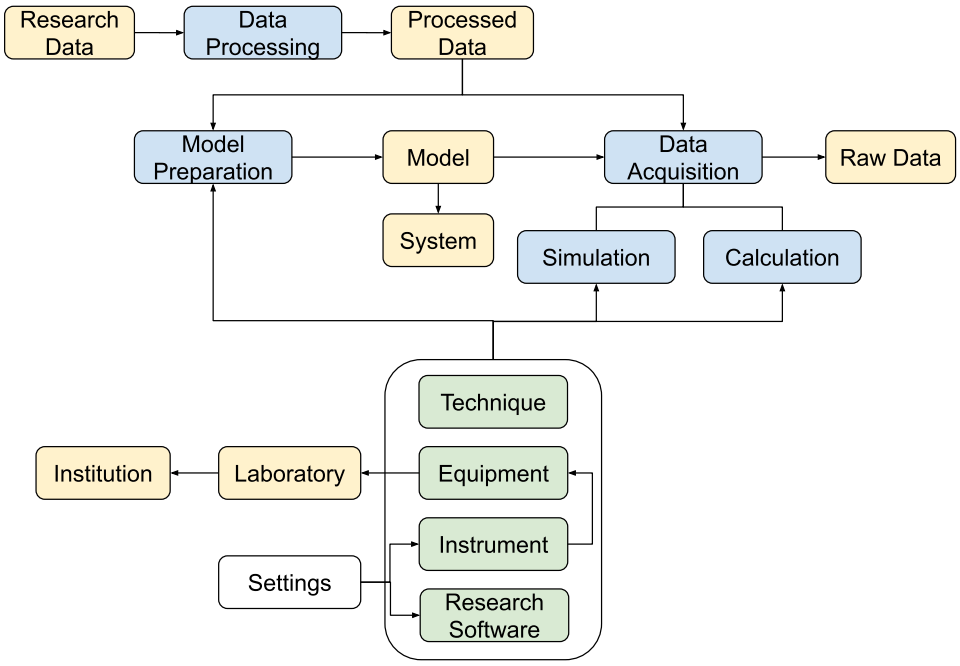

Figure 3: Computational workflow. Green boxes: agents, yellow and white boxes: entities, blue boxes: processes. The rounded-corner box groups the terms connected to each of the processes, to simplify the graphical representation.

Data Analysis Lifecycle

Figure 4 illustrates the processes which may be performed in a simple Data Analysis Lifecycle. Multiple loops are possible, involving some or all the processes, even in a different order. A typical Data Analysis Lifecycle starts with the Data Processing of Raw Data to prepare it for one or more further processes. The Processed Data is usually an intermediate result, used as input of Data Analysis to obtain Analyzed Data. Data Interpretation is then performed on Analyzed Data to draw Conclusions that support the answer to some specific research question. The processes in the Data Analysis Lifecycle may require the use of Reference Data (as the main input data or for comparison) and the employment of Research Software, whose configuration parameters are described by Settings.

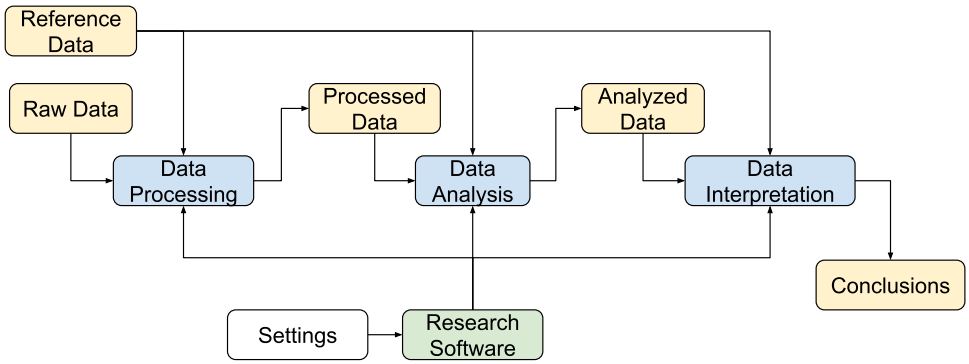

Figure 4: Data Analysis Lifecycle. Green boxes: agents, yellow and white boxes: entities, blue boxes: processes.

Data

Figure 5 offers a schematic representation of the different types of data considered in the Glossary and their related entities. Research Data is defined as any kind of data collected, created or examined to be analyzed or considered as a basis for reasoning, discussion or calculation in a research context, with the purpose of generating, verifying and validating original scientific claims that support the answer to some specific research question. According to this definition, Raw Data, Processed Data, Analyzed Data and Reference Data are particular types of Research Data. Any Research Data may be described by Metadata and, together with it, may be part of a Dataset. A Dataset may also consist of other Datasets. Publication Data consists of one or more Datasets that have undergone quality assessment, as well as the relevant scientific and administrative Metadata.

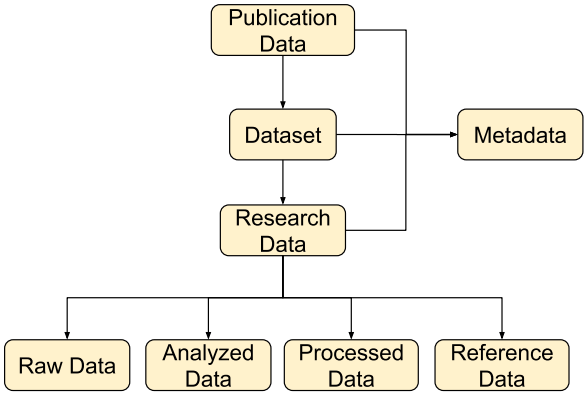

Figure 5: Schematic representation of the different types of data considered in the Glossary and their relations.

Data and metadata management

Figure 6 reports the schematic representation of the different types of data, already shown in Figure 5, including the information systems considered in the Glossary for data and metadata management. In particular, a distinction is made between a Data Collaboration Platform, intended as a sharing system for Datasets and Research Data which are still subject to active research, and a Data Repository, particularly suitable for depositing Publication Data which are not likely to be altered again and are associated with a Persistent Identifiers. A Metadata Repository is instead an information system specifically suited to store, manage and provide access to Metadata, to assign it a Persistent Identifier, and to link it to the data it describes.

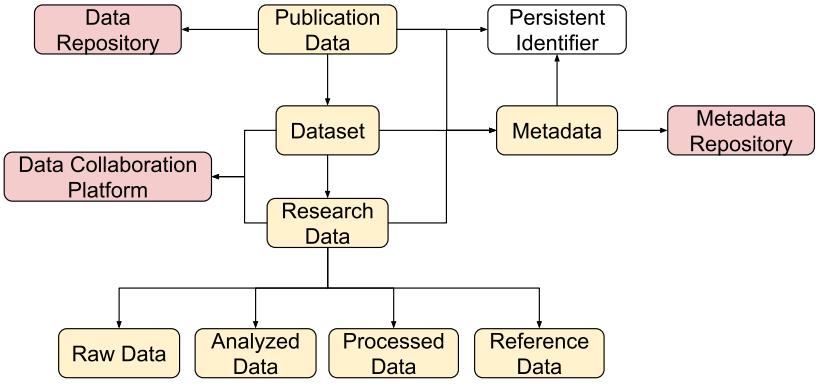

Figure 6: Schematic representation of the different types of data considered in the Glossary and their relations, including the information about the management infrastructure. Yellow and white boxes: entities, red boxes: information systems.

Acknowledgements

This work was carried out with the support of: the Joint Laboratory Model and Data-driven Materials Characterization (JL MDMC), a cross-centre platform of the Helmholtz Association; the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot; the research program 'Engineering Digital Futures' of the Helmholtz Association of German Research Centers; the Helmholtz Metadata Collaboration Platform.

References

[1] Metadata WG: https://jl-mdmc-helmholtz.de/mdmc-activities/metadata-working-group/
[2] NFFA-EUROPE Pilot: https://nffa.eu
[3] JL MDMC: https://jl-mdmc-helmholtz.de
[4] Skosmos: https://skosmos.org
[5] Crossref schema version 5.0 (2020): https://gitlab.com/crossref/schema/-/blob/5.0/schemas/common5.0.xsd
[6] Dublin Core, DCMI Metadata Terms: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[7] DataCite Metadata Schema: https://schema.datacite.org
[8] Schema.org: https://schema.org
[9] V. Bunakov et al., Metadata for Experiments in Nanoscience Foundries. In: Kalinichenko L., Kuznetsov S., Manolopoulos Y. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2016. Communications in Computer and Information Science 706, 248-262. Springer (2017). DOI: 10.1007/978-3-319-57135-5_18
[10] Matthews, B., et al., Using a Core Scientific Metadata Model in Large-Scale Facilities. International Journal of Digital Curation 5(11), 106-118 (2010). DOI: 10.2218/ijdc.v5i1.146
[11] S. Sufi, B. Matthews, A Metadata Model for the Discovery and Exploitation of Scientific Studies. In: Talia, D., Bilas, A., Dikaiakos, M.D. (eds) Knowledge and Data Management in GRIDs 135-149. Springer US (2007). DOI: 10.1007/978-0-387-37831-2
[12] CODATA CASRAI, RDM Terminology WG: https://codata.org/codata-casrai-rdm-terminology-working-group/
[13] M. Könnecke et al., The NeXus data format. Journal of Applied Crystallography 48(1), 301-305 (2015). DOI: 10.1107/S1600576714027575
[14] National Institute of Standards and Technology (NIST) Materials Resource Registry Vocabulary: https://matportal.org/ontologies/NMRRVOCAB
[15] A. Medina-Smith et al., A Controlled Vocabulary and Metadata Schema for Materials Science Data Discovery. Data Science Journal 20(1), 18. DOI: 10.5334/dsj-2021-018
[16] Elementary Multiperspective Material Ontology: https://emmo-repo.github.io
[17] S. Arndt et al., Metadata4Ing: An ontology for describing the generation of research data within a scientific activity. DOI: 10.5281/zenodo.5957103
[18] Wikipedia: https://www.wikipedia.org
[19] Materials Design Ontology: https://matportal.org/ontologies/MDO-FULL/?p=summary
[20] Physics Subject Headings: https://physh.org/browse