MDMC Glossary of terms

This glossary, compiled by the MDMC Metadata Working Group, defines and explains the high-level terms used in the context of the MDMC (meta)data management.

The definitions make use in a consistent way of glossary terms, which are written in **bold** with Capital Initial Letters.

This glossary is intended to be a living document, subject to updates if required by the community. For any inquiries, please contact Dr. Rossella Aversa ([rossella.aversa@kit.edu](mailto:rossella.aversa@kit.edu)).

## Analysed Data

**Research Data** which is the primary output of any kind of **Data Analysis** performed on **Research Data**, typically on **Processed Data**, by one or more **Research Users**, possibly using **Research Software**. **Analysed 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. **Analysed Data** may be stored in a **Data Collaboration Platform** and/or in a **Data Repository**. **Analysed Data** may be part of a **Dataset**.

## Conclusions

The primary output of **Data Interpretation** performed by one or more **Research Users** on **Research Data**, typically on **Analysed Data**, possibly using **Research Software**. **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 steps of a **Study** or about future work. **Conclusions** form an important part of a **Study** debrief and are usually reported in **Scientific Publications**.

## Correlative Characterization

The action of characterizing one or more **Samples** (depending on the particular research context) and connecting the different types of information from co-referenced (in time or space) multi-dimensional **Research Data**. This may include the output of multiple **Measurement Techniques** 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 **Sample** areas across multiple length scales or over time.

## Data Analysis

Data treatment performed by one or more **Research Users**, who collect, model and analyse **Research Data**, typically **Processed Data**, to extract insights that support the answer to some specific research question (i.e., **Conclusions**). **Data Analysis** may include: linear combination, data fitting, data modeling, pattern extraction. **Data Analysis** may be performed using **Research Software**. The output of **Data Analysis** is **Analysed Data**.

## Data Analysis Lifecycle

Set of processes carried out by one or more **Research Users**, who systematically perform actions and apply methods, e.g., statistical and/or logical techniques, on **Research Data** in order to produce synthesized knowledge to, e.g., detect pattern, determine relationships, develop explanations, test hypotheses, prove theories, suggest the **Conclusions** of the **Study**. **Data Analysis Lifecycle** includes (but is not limited to): **Data Processing**, **Data Analysis**, **Data Interpretation**. These processes may consist of different steps, may be iterative and may be combined in chains or workflows. **Data Analysis Lifecycle** may be performed using **Research Software**.

## 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

Process performed by one or more **Research Users**, who assign meaning to **Research Data**, typically **Analysed Data**, and determine the **Conclusions** of the **Study**, possibly in comparison with **Reference Data**. **Data Interpretation** supports decision-making about the next steps of the **Study** or about future work. **Data Interpretation** may be performed using **Research Software**.

## Data Processing

Data treatment performed by one or more **Research Users** on **Research Data**, typically **Raw Data**, to prepare it for another step of the **Data Analysis Lifecycle**, e.g., **Data Analysis** or **Data Interpretation**. **Data Processing** usually consists of routine actions. It may include: filtering, denoising, transformation, fusion or compression of existing **Research Data**, as well as calibration, normalization, statistical data reduction, background subtraction, correction of artifacts. **Data Processing** may be performed using **Research Software**. 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 **Measurement 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 **Raw Data** (including the output of computational **Experiments**), **Processed Data**, **Analysed Data**, or other **Datasets**. The components of a **Dataset** remain individually identifiable.

## Equipment

Any kind of item, device, machine or other tool (also virtual) used by one or more Research Users to perform one or more **Sample Component Syntheses**, **Sample Preparations** and/or **Measurements**. Usually, the **Equipment** is located in a **Laboratory** hosted by an **Institution**. According to this definition, an **Instrument** is a particular type of **Equipment**.

## Experiment

Identifiable and reproducible activity with a clear start time and end time, which may include a set of one or more **Sample Component Syntheses**, **Sample Preparations** and/or **Measurements**, performed by one or more **Research Users**. An **Experiment** may be a simulation (computational **Experiment**) or a combination of computational and physical **Measurements**.

## Instrument

Identifiable piece of **Equipment** used by one or more **Research Users** to perform a **Measurement** and to generate **Raw Data**. **Instrument** is located in a **Laboratory** hosted by an **Institution**. **Instrument** may also stand for a software, a software module and/or a particular configuration of it, used to perform a simulation run (computational **Measurement**).

## Institution

Hierarchical entity which hosts one or more **Laboratories**, including the virtual ones.

## Laboratory

Place (could also be virtual) hosted by an **Institution**, where one or more **Instruments**, as well as the **Equipment**, are located and the **Measurement** is performed. **Laboratory** may be the hardware and/or the software platform and/or the services which allow to order and manage computational **Experiments**. In this case, the software platform (virtual **Laboratory**) serves the purpose of managing software modules (virtual **Instruments**).

## Measurement

Identifiable and reproducible activity, performed by one or more **Research Users**, who generate a single self-consistent unit of **Raw Data** about a **Sample** or a set of them using an **Instrument** under constant or varying controlled conditions, depending on the particular research context. **Measurement** is specific to **Instrument**: an investigation on the same **Sample** conducted using a different **Instrument** implies a different **Measurement**. A **Measurement** may also be performed during the **Sample Component Fabrication** and/or the **Sample Preparation**, e.g., to characterize the intermediate and/or final resulting **Sample**(s) or **Sample Component**(s). A computational **Measurement** may be a part of a simulation (computational **Experiment**), e.g., a simulation run using a particular model, configuration or input(s).

## Measurement Technique

Technique or technology corresponding to the method used during a **Measurement** to collect **Raw Data** about a **Sample**, a **Sample Component** or a set of them with an **Instrument**.

## Metadata

Any descriptive data intended to help 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** describes information pertaining to the research **Projects**, including (but not limited to) **Research Users**, **Studies**, **Experiments**, **Measurements**, **Instruments**, **Samples**, and corresponding **Data Analysis Lifecycle**. **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**.

## 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**.

## Processed Data

**Research Data** which is the primary output of any kind of **Data Processing** or manipulation of **Raw Data** performed by one or more **Research Users**, possibly using **Research Software**, in order to prepare it for another step of the **Data Analysis Lifecycle**, e.g., **Data Analysis** or **Data Interpretation**. **Processed 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. **Processed Data** may be stored in a **Data Collaboration Platform** and/or in a **Data Repository**. **Processed Data** may be part of a **Dataset**.

## Project

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

## Publication Data

**Dataset**(s) generated by one or more **Research Users** 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 **Raw Data**, **Processed Data** and/or **Analysed Data**, as well as the relevant **Metadata** about the **Experiment**(s) and the **Data Analysis Lifecycle**. **Publication Data** may be attributed to some or to all the **Research Users** who are members of the **Project**.

## Raw Data

**Research Data** which is the primary output of a **Measurement**, collected by one or more **Research Users** using an **Instrument**, before any subsequent **Data Processing**. **Raw Data** may be the output of a simulation run (computational **Measurement**). **Raw 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. **Raw Data** may be stored in a **Data Collaboration Platform** and/or in a **Data Repository**. **Raw Data** may be part of a **Dataset**.

## Reference Data

**Research Data** not produced during the current **Study**, which 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 analysed 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 statistics, output of **Experiments** and/or individual **Measurements**, observations resulting from fieldwork, survey results, recordings and images. According to this definition, **Raw Data**, **Processed Data**, **Analysed 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 stored in a **Data Collaboration Platform** and/or in a **Data Repository**. **Research Data** may be part of a **Dataset**.

## Research Software

Software used to generate, process, analyse or access **Research Data** during any of the processes included in the **Data Analysis Lifecycle** (possibly including data rendering, visualization, plotting). Depending on the research context, **Research Software** can be used during **Data Processing**, **Data Analysis** or **Data Interpretation** taking as input **Raw Data**, **Processed Data** or **Analysed Data**, respectively, and giving as output **Processed Data**, **Analysed Data** or any kind of synthesized knowledge (e.g., **Conclusions**), respectively. Any acquisition software used during **Measurements** is considered part of the **Instrument** and should be described as such. Any software used to perform simulation runs (computational **Measurements**) and to generate **Raw Data**, it is considered an **Instrument** and should be described as such.

## Research User

Person, usually member of a **Project**, who conducts any part of the **Experiment**(s), or performs any of the steps of the **Data Analysis Lifecycle** during the course of one or more **Studies**, in order to collect and/or analyse **Research Data**, or is interested in reusing **Research Data** collected 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 (e.g., principal investigator, **Instrument** scientist, **Project** member).

## Sample

Identifiable entity (typically a piece of material) with distinctive properties (structural, chemical, dimensional, functional and others), composed by 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** may also stand for a model, configuration or input (or any combination of them) of a simulation run (computational **Measurement**).

## Sample Component

Identifiable entity (typically a piece of material) with distinctive properties (structural, chemical, dimensional, functional and others), which is fabricated during the **Sample Component 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 molecules.

## Sample Component Fabrication

Fabrication of a **Sample Component** in controlled conditions, performed by a commercial enterprise, one or more **Research Users** or a third party. **Sample Component Fabrication** may require the use of **Equipment** and **Instrument**(s). A **Measurement** may also be performed during the **Sample Component Fabrication**, e.g., to characterize the intermediate and/or final resulting **Sample Component**(s).

## Sample Carrier

A piece of **Equipment** used for carrying one or more **Samples** and/or one or more **Sample Holders** 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

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

## Sample Preparation

Identifiable and reproducible set of actions (physical changes or chemical reactions) typically carried out by one or more **Research Users** to produce one or more **Samples** and/or to make the **Sample**(s) fit to perform a **Measurement**. The actions may be performed on (or between) one or more **Sample Components** or **Sample**(s). **Sample Preparation** may require the use of **Equipment** and **Instrument**(s). A **Measurement** may also be performed during the **Sample Preparation**, e.g., to characterize the intermediate and/or final resulting **Sample**(s) or **Sample Component**(s).

## 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**.

## Study

Set of one or more **Experiments** and corresponding **Data Analysis Lifecycle** performed by one or more **Research Users** who are part of the same **Project**.