Workshop on Data Lifecycles in the World of Materials Modelling and Characterisation

Workshop Report

Alexandre Ouzia[1,*], Geoffrey Daniel[2], Jesper Friis[3], Emanuele Ghedini[4], Peter Imrich[5], Julian de Marchi[6], Yoav Nahshon[7], Franz Martin Rohrhofer[8], Sophie Schmid[9], and Alexandra Simperler[10]

[1] Heidelberg Materials AG, Global R&D, Germany

[2] Commissariat à l'Énergie Atomique et aux Énergies Alternatives, France

[3] SINTEF, Norway

[4] University of Bologna, Italy

[5] KAI GmbH, Austria

[6] NLR-Netherland Aerospace Centre, The Netherlands

[7] Fraunhofer IWM, Germany

[8] Know Center Research GmbH, Austria

[9] Institute for Mechanics of Materials and Structures, Technische Universität Wien, Austria

[10] EMMC ASBL - European Materials Modelling Council, Belgium

* Corresponding Author: alexandre.ouzia@heidelbergmaterials.com

Funded by the European Union

Contents

1.Introduction

The Horizon Europe projects MatCHMaker[1], AddMorePower,[2] AID4GREENEST,[3] CoBRAIN,[4] KNOWSKITE-X[5] and D-STANDART[6] came together in Vienna on the 7th of April 2025 (see Appendix A) to provide their approaches to, and issues with each of the stages of the data lifecycle.[7] In total, 27 workshop attendees (both representants of the projects and external attendees) discussed future developments for improvement and best practice for how they may plan, acquire, process, analyse, preserve, and share their data. The projects cover a range of industries, and thus, are using a variety of characterisation and materials modelling techniques which may not be reciprocated by a sister project. However, what they have in common is dealing with a plethora of data that undergo a lifecycle from being created to bearing knowledge.

Representants of the above projects provided insights in how they tackle their respective data lifecycles, including:

- Planning
  - Was there a prioritised plan for what data was needed, and how to extract knowledge from it?
  - Whether and how the projects worked with Data Management Plans?
  - Did they look for existing semantic assets to prepare for interoperability?
  - Did they purchase software and or equipment to produce data?
- Acquiring
  - How is data acquired?
  - Are meta data acquired?
  - What types of data are acquired?
  - Are the data described as they are acquired?
- Processing
  - How are data extracted?
  - Are there automatic workflows in place?
  - How is provenance ensured?
  - Are processes that cut out or join samples traced? If so, how is it done?
- Analysing
  - What tools did/will they develop to extract knowledge from their data?
- Preserving
  - How do they document their "new" data?
  - Are they using versioning?
  - In what sort of repository does the preservation happen?
- Sharing
  - Are their data shared with 3rd parties?
  - Do they publish their code/apps/workflows?

[1] https://cordis.europa.eu/project/id/101091687
[2] https://cordis.europa.eu/project/id/101091621
[3] https://cordis.europa.eu/project/id/101091912
[4] https://cordis.europa.eu/project/id/101092211
[5] https://cordis.europa.eu/project/id/101091534
[6] https://cordis.europa.eu/project/id/101091409
[7] e.g., https://nfdi-matwerk.de/solutions/via-data-life-cycle

- How and to what degree is the data made FAIR (Findable, Accessible, Interoperable, Reusable)?

This was done via a pre-workshop survey (see Appendix B) and during the workshop with six talks highlighting key points of the projects providing valuable insights on data life cycles. Thereafter, all workshop participants were joining an activity using the information given in these talks and their professional experience to jointly map out the key stages of the data life cycle in materials R&D (from data generation to long-term reuse) and identify pain points (i.e., challenges, bottlenecks, uncertainties) and suggest good practices/solutions (i.e., already existing successes or desirable potential improvements).

2. The Data Lifecycle in the Projects According to the Pre-workshop Survey

The survey reveals that although various data-lifecycle tools exist, Excel[8] remains the default choice for many participants. This choice stems as much from habit and user-friendliness as from any lack of available alternatives. Also, most respondents were academic researchers in the materials area, which often has lower data volumes compared to life sciences. Hence, options such as relational databases or electronic lab notebooks are often not considered.

The characterisation data stem from a frequent use of electron microscopy (SEM, TEM, EBSD, EDX) and mechanical testing (strength, hardness, creep), sometimes enhanced by tribological or electrochemical methods. Certain projects are using advanced in-situ techniques such as synchrotron XRD or peruse external databases (e.g., NIMS[9]). Modelling data stem from finite element and phase-field methods to AI or ML tools for image segmentation or property prediction, often linked to experimental data via in-house scripts. Yet the practical integration of modelling and characterisation remains largely manual: data is exchanged via email or spreadsheets, and multiple proprietary instrument formats require time-consuming conversions. Except for a few advanced users, there is no established, automated data flow between experiments and simulations.

Some projects have consistent, high-volume data generation and clear problem definitions, making them prime candidates for structured data-lifecycle pipelines, robust metadata standards, or even "ontology-like" approaches. Others are more exploratory or conceptual, where fundamental questions are still evolving. Attempting to impose elaborate data management in these nascent stages can disrupt creativity and motivation. Many real-world efforts lie somewhere in between, suggesting an approach where minimal standards (such as consistent naming or .csv exports) are introduced early on, with advanced systems phased in once the project scope, data volume, and collaborative needs become clear.

Despite a handful of data management plans (DMPs) or standard operating procedures (SOPs), most respondents rely on ad hoc or partial solutions for documentation, data storage, and long-term preservation. Excel, custom scripts, and local repositories are ever-present, while cloud-based options (Zenodo, HPC platforms, GitHub) or LIMS solutions see patchy adoption. Formal metadata remains rare and only a few teams systematically unify file structures. Issues like cost, IP restrictions, and unclear data ownership often increase the gaps. Some note that science has progressed for centuries

[8] https://www.microsoft.com/en-gb/microsoft-365/excel
[9] https://www.nims.go.jp/eng/

using “informal methods,” raising the question of whether advanced data tools are truly necessary—and, if so, for which use cases.

Foremost among these challenges is the lack of data interoperability and standardisation. Diverse formats and minimal metadata make it difficult to merge microscopy images, mechanical test logs, and simulation outputs into cohesive workflows. Equally pressing is the absence of time or resources dedicated to data stewardship: academic researchers, under pressure to publish, often deprioritise data management tasks. Industrial confidentiality and IP concerns add another barrier to sharing data, and skill gaps in data-lifecycle practices further complicate adoption. Scalability also emerges as a problem: some teams generate vast datasets, but have no robust strategy for storing or processing them at scale.

Proposed solutions include adopting “common sense” export formats and metadata protocols, offering more training or practical guidelines (such as short workshops and step-by-step manuals), and allocating time or budget within project plans specifically for data tasks. Many respondents stress the value of periodic cross-team “data checkpoints” to align experimentalists, model developers, and data managers on what should be collected and how. For high-throughput groups, better tooling, e.g., automated ingestion pipelines, HPC workflows, or advanced AI, may significantly reduce the burden of repetitive analyses. Above all, a more agile approach is encouraged, replacing rigid, “waterfall” data-management strategies with iterative feedback cycles that evolve alongside the science itself.

Beyond technical fixes, organisational support remains critical. Without recognition of data stewardship as a legitimate project deliverable - backed by deadlines, budgets, or management directives - researchers will continue to view data management as an unwelcome administrative task. Likewise, IP policies and proprietary data constraints require careful negotiation if truly open, cross-project sharing is to become feasible. Many advanced users want simpler instructions or hands-on guidance for setting up robust data workflows, and novices echo the same need.

Despite universal agreement on the potential benefits - improved reproducibility, better synergy between experiments and models, more efficient collaboration - there is no single “best” data-lifecycle approach. The suitability of advanced, structured solutions depends on a project’s maturity, throughput, and stability of problem definition. Teams operating at a lower scale or in the early conceptual phase may find that a phased or hybrid approach is more appropriate, starting with minimal but effective standardisation practices. Ultimately, Excel’s popularity underscores that simplicity and familiarity often outweigh complexity, especially when formal data-management tasks are seen as secondary to the core research. If the goal is to move beyond ad hoc methods, success depends on agile project management, tailored frameworks, and genuine support for data stewardship at both the cultural and organisational levels.

3. The Data Lifecycle in the Projects Presented during the Workshop

In one of their use cases, project **MatCHMaker** is working on characterisation and modelling data workflows for low-carbon cement optimisation. This involves four partners, who (i) define the design of the experiments and prepare samples, (ii) acquire SEM or TEM images, (iii) compute phase assemblages thereof, and (iv) from the latter, compute the strength of the materials. For these

partners, solving the scientific problem is the top priority, while the formulation of the problem continues to evolve. During this evolvement, the teams may require more experimental data. This can result in an iteration of acquisitions, or that the most previous versions are discarded and a new set of characterisation experiments is performed. To formulate the scientific problem clearly, a core team of experts in the different fields must work together and interact for long hours every week to learn from one another, speak the same language, and then build the data life cycles tools. The latter only makes sense once there is a clear understanding and stable vocabulary in the formulation of goals, research questions and problems. Hence, the four protagonists cocreating this workflow, advice to jointly formulate exactly, what the scientific problem is, which they try to solve. However, as with all R&D, the problem-solving process remains unpredictable and the data lifecycle naturally is all but smooth and needs to be adapted continuously. Once their workflow reaches a higher Technology Readiness Level (TRL) and has a high-volume throughput with consistent data and many users, the path to a more sophisticated data lifecycle is open. In a second talk, the MatCHMaker team introduced a CHADA[10] template for SEM and how to represent it in their knowledge base. This required them to make CHADAs machine readable and represent workflows semantically. Their tools, once developed, aim to simplify the documentation of workflows in a FAIR way.

AID4GREENEST is working to revolutionise steel manufacturing by harnessing AI and digitalisation. To this end, they are developing a Process Editor digital tool for representing processes in a FAIR manner. A particular area of interest for using this tool is the documentation of characterisation workflows. The Process Editor supports multiple schemas and standards to ensure compliance, particularly with the CHADA standard. Advanced semantic technologies are being employed for FAIR data management, which are labour-intensive and require collaboration between ontology and domain experts. Typically, ontology development approaches take about six months to become practically useful, necessitating the involvement of ontology experts who actively engage with domain specialists. To address this challenge, AID4GREENEST is employing an innovative bottom-up approach to ontology development. For this purpose, an additional digital tool called the Semantics Manager has been developed. This tool collects and organises vocabulary terms across domains, enabling experts to communicate effectively and share a common language. This approach is highly rewarding as it ensures a well-structured, modular foundation for the data lifecycle and promotes the reusability of ontologies across various domains.

D-STANDART produces and gains knowledge from composite materials characterisation and modelling data to improve the lifetime estimation of composite structures. The beneficiaries also work with ontologies and use an ontology of materials characterisation, called CHAMEO.[11] They are using CHADAs to describe their characterisation workflows and MODAs[12] to document their materials modelling workflows. To make this palatable for an industrial uptake, their data lifecycle is standardised and a consistent and viable digital thread holds all pieces together. One of their toughest challenges was to convert partner specific metadata to a consistent format which was greatly helped by developing domain-specific translation tools. D-STANDART made good use of existing developments such as CHADA and CAMEO and adapted and extended as was needed for their project. There are ongoing efforts to automatise tools related to the data lifecycle to enable more stakeholders to take part and profit from their approach. One of these tools will be a web-hosted digital thread service to capture characterisation results and make them searchable. But not only the human-in-the-loop is considered; their data are indexed and as such fit for ML purposes.

[10] https://www.cencenelec.eu/media/CEN-CENELEC/CWAs/RI/2025/cwa17815_2025.pdf
[11] https://github.com/emmo-repo/domain-characterisation-methodology
[12] https://www.cencenelec.eu/media/CEN-CENELEC/CWAs/RI/cwa17284_2018.pdf

The engineering of thermal spray coatings lays at the heart of project **CoBRAIN**, where both characterisation and materials modelling are used to produce data. The beneficiaries are looking for novel formulations to avoid critical or toxic materials. They created DMPs and reused conceptualisation templates, a semantic asset developed by OntoTrans[13] based on the Elementary Multiperspective Material Ontology (EMMO)[14] as top-level ontology. These templates enable an ontologist to extract knowledge from a domain expert in a formal and standardised way. Data came in formats that were specific to the data owners and often not ready for semantic interoperability, hence one project partner refactors them. When it came to data processing, CoBRAIN found unprecise or non-unique data identifiers as a severe bottleneck. One outcome of the project will be a knowledge base of thermal spray coatings which can be searched via SPARQL queries.

KNOWSKITE-X is developing electrode materials based on mixed oxides with perovskite structure. From the beginning of these project, the beneficiaries started with DMPs and had the interoperability of data in mind. They run an internal data management platform built on an integrated rule-oriented data system (iRODS),[15] which enables accessing, managing, and sharing data. Hence, in KNOWSKITE-X, the data lifecycle is run via iRODS, which was adapted to the project's needs. For example, there is a workflow editor that can be used to create and view CHADAs and MODAs and an ML platform, that can train databased materials models. The platform profits from a Python API and can be connected to external frameworks. Despite their mature setup, the beneficiaries still encounter challenges. One is inconsistency in workflows, as different stakeholders bring in individual approaches and data formats. They also may have different requirements as in how data should be stored. Another interesting challenge is the amount of data – such systems tempt their users to have them ingest vast amounts of data, even though all data may not be needed to be stored long-term. The project partners see a steady discussion kept alive throughout their project vital to make the most out of their data.

4. The Workshop Activity

The goal of this activity was to map out the key stages of the data lifecycle in materials R&D (from generation to long-term reuse) and identify:

- Pain points (challenges, bottlenecks, uncertainties)
- Good practices/solutions (existing successes, potential improvements)

This exercise clarified where participants see the biggest issues and the most promising strategies in moving data across the characterisation/modelling workflows.

Six flipcharts were provided and each of them represented a main part of the data lifecycle, i.e., Data Generation, Data Processing, Data Curation, Data Sharing, Data Publication, and Data Reuse.

[13] https://cordis.europa.eu/project/id/862136
[14] https://github.com/emmo-repo/EMMO
[15] https://irods.org/

Data Generation

Pain points

- Data quality might be unknown/hard to estimate
- Data provenance (i.e., was the model trained correctly, the characterisation experiment properly gauged, was the model understood as being a model, was something improved in the data generation workflow which makes a previous data set redundant, …)
- Data postprocessing (outliers, signal-to-noise, all regions explored, …)
- Data volume (ML/AI can produce data very fast and fill data storage)

Solutions

- Automation of the data generation process
- Provide protocols for the generation of both simulated and experimental data

Data Processing

Pain Points

- Inaccurate physical interpretation (i.e., inconsistent artifacts in the acquisitions, signal-to-noise misinterpreted, …)
- Unstructured data, many different formats
- Often interpretation by one or more human experts is needed
- Often no version control is logged
- Processing scripts/tools without quality control/updates

Solutions

- Provision of physics-based descriptors
- Definition of standardised data processing workflows including provenance logging
- Processing scripts/tools should be clearly referenced (FAIR software)

Data Curation

Pain points

- Incomplete data (meta data missing, historical data, low quality, undocumented format changes, …)
- No regulation when to purge data and how to replace old data with better newer data
- No well-established toolset for describing meta data for data creation (which model, methodology, characterisation machine vendor, …)

Solutions

- Use common vocabularies and ontologies
- Admit only FAIR data
- Curation needs the perspective of both data producer and data consumer
- Tools for curators

Data sharing within the project

Pain points

- Different metadata/formats/quality
- Accessibility (open vs closed data, IP, data protection regulations)
- In an R&D setting, relevant data change frequently
- Synthetic data may not always represent genuine corporate data

Solutions

- Ensure that data is FAIR
- Provide sufficient metadata and context for the data to be useful
- Agree on a platform solution and access thereof (hubs, iRODS, inter/national repositories, …)

Data Publication – sharing data externally

Pain Points

- IP
- Data provenance
- How long before data becomes "outdated"
- Link a good repository to a publication: make data citeable
- The academic publication system is not geared towards data

Solutions

- Work with embargos to honour IP
- FAIR data including full provenance for reproducibility
- Public and free of charge EU hosted repositories with DOIs

Data Reuse

Pain Points

- Data credibility and trust due to lack of metadata
- One cannot easily test for reproducibility or relevance
- Data quality is unknown
- Data format may not be useful

Solutions

- Establish data quality rating factor, e.g., R-factor for XRD data
- Documentation including rich metadata and ontologies
- Choose machine readable format
- Choose an approved repository, such as NOMAD[16] or EOSC[17] for materials science data, etc.

5. Conclusions and Outlook

A recent workshop in Vienna (April 7, 2025) brought together six Horizon Europe projects - MatCHMaker, AddMorePower, AID4GREENEST, CoBRAIN, KNOWSKITE-X, and D-STANDART - to discuss every stage of the data lifecycle in materials modelling and characterisation: planning, acquiring, processing, analysing, preserving, and sharing. A pre-workshop survey underscored the prevalence of ad hoc data tools (e.g., Excel) over more advanced frameworks. Although many participants recognized the benefits of robust data management, i.e., improved reproducibility, streamlined integration between modelling and characterisation, efficient collaboration, major obstacles included limited time, low recognition of data stewardship, and intellectual property (IP) constraints. During the workshop's presentations and interactive sessions, participants mapped common pain points (e.g., inconsistent metadata, data provenance, lack of interoperability) and

[16] https://nomad-lab.eu/nomad-lab/
[17] https://eosc.eu/eosc-about/

proposed key improvements: adopting minimal but standardised metadata practices early on, allocating budget for data-focused roles, and embedding "data checkpoints" throughout a project's lifecycle.

The project exemplars highlighted both the variety of needs and the adaptability required. For instance, MatCHMaker's iterative workflows for low-carbon cement rely on flexible approaches until the problem definition stabilises; AID4GREENEST invests significantly in ontology-based solutions for AI-driven steel manufacturing; D-STANDART uses the CHAMEO ontology and digital thread concepts for composite materials; CoBRAIN applies EMMO-based frameworks to unify thermal spray coating data; and KNOWSKITE-X leverages an iRODS platform to handle large datasets for perovskite electrode materials. Despite these diverse strategies, the workshop revealed broad agreement on the importance of combining domain expertise, common vocabularies or ontologies (e.g., CHADA, MODA, CHAMEO), and modern data stewardship roles.

Building on these insights, there is no one-size-fits-all approach. Projects differ significantly in scope, maturity, and throughput; some are highly exploratory, benefiting little from rigid data-handling rules at early stages, while others produce large, steady data streams that demand rigorous pipelines from the outset. A practical recommendation is to adapt data management tools to each project's maturity level. One could begin with a concise DMP that focuses on small-scale demonstration. Early adoption of "common sense" standardisation—such as clear file-naming conventions and minimal but consistent metadata—provides a foundation that can be expanded as the project grows. In addition, an agile approach is invaluable, because data quality and usefulness typically evolve alongside the research. Thus, processes must be revisited and updated at logical intervals. Notably, DMPs are "living" documents that can be used as checkpoints for the data steward to monitor partner engagement as a key performance indicator.

Although strong technical solutions for data management, ranging from SQL-based systems to complex semantic ontologies, are readily available, widespread adoption often remains hindered by non-technical barriers. In many R&D environments, data stewardship tasks languish at the bottom of to-do lists, overshadowed by immediate research milestones and a lack of formal recognition, which can make data management feel tedious. Historically, science has progressed without elaborate schemas, yet today's interdisciplinary, data-heavy landscape demands more scalable approaches. Consequently, the role of a dedicated data steward becomes critical: this individual not only ensures that each researcher maintains the required metadata and adheres to FAIR guidelines but also conducts periodic "checkpoints" to prevent data management from becoming a perpetual motivation killer. Their responsibilities extend well beyond oversight and training; they must continuously advocate for resources, foster communication among domain experts and IT teams, and update policies such as the DMP in real time. Structured training sessions—one at a project's outset and additional targeted workshops at intermediate stages—help maintain momentum, reinforce best practices, and keep data stewardship aligned with evolving research needs.

Researchers and industry partners in materials science may take inspiration from the Allotrope Foundation's work in chemical and life sciences, where community-driven data standards revolutionize how data are acquired, shared, and leveraged. In materials science, a similar push could begin with developing standard templates for frequently used techniques, building on efforts such as CHADA (for characterisation) and MODA (for modelling). Several EMMO-compliant domain ontologies - like CHAMEO or the nanoindentation testing or microscopy ontologies - already support standardized metadata for specialised methods. Projects such as MatCHMaker, AddMorePower, AID4GREENEST, CoBRAIN, KNOWSKITE-X, and D-STANDART are working toward robust knowledge

platforms and tools, demonstrating that a coordinated approach can make data lifecycles in materials modelling and characterisation more reliable, reproducible, and scalable.

In short, by striking a balance between flexibility and rigor and by assigning clear roles, especially for a data steward, R&D teams can transition from ad hoc practices to streamlined, sustainable data lifecycle management. The recent workshop exemplifies how collective discussions and concrete tools help tackle the persistent organizational and technical bottlenecks. This shift not only enhances reproducibility and collaboration but ultimately brings researchers one step closer to realising the full potential of their data.

6. Acknowledgement

This study has received financial support from the European Union: MatCHMaker (GA 101091687), AddMorePower (GA 101091621), AID4GREENEST (GA 101091912), CoBRAIN (GA 101092211), KNOWSKITE-X (GA 101091534) and D-STANDART (GA 101091409).

7. Disclaimer

All statements of fact, opinion, or analysis expressed in the report are those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. The information used and statements of fact made are not guarantees, warranties or representations as to their completeness or accuracy. The authors assume no liability for any short term or long terms decision made by any reader based on analysis included in this report. In many cases, the opinion expressed in the reports is our current opinion based on prevailing trends and is subject to change.

8. Acronyms, Abbreviations and Elucidations

AI – Artificial Intelligence

Beneficiary - Beneficiaries are the legal entities involved in an EU project covered by a grant agreement.

CHADA – Characterisation Data

CHAMEO - Characterisation Methodology Domain Ontology

DMP - Data Management Plan

EBSD - Electron Backscatter Diffraction

EDX - Energy-Dispersive X-ray Spectroscopy

EMMC – European Materials Modelling Council

EOSC - The European Open Science Cloud

FAIR – Findable, accessible, interoperable, and reusable

HPC – High Performance Computing

IP – Intellectual Property

iRODS - integrated rule-oriented data system

LIMS - Laboratory Information Management System

ML – Machine Learning

MODA – Modelling Data

NIMS - National Institute for Materials Science, located in Japan

R&D – Research and Development

SEM - Scanning Electron Microscopy

SOP - Standard Operating Procedure

SPARQL - a query language specifically designed for retrieving and manipulating data stored in the Resource Description Framework format.

TEM - Transmission Electron Microscopy

TRL – Technology Readiness Level

XRD - X-ray diffraction

9. Appendix A: Workshop Schedule

The meeting was facilitated by the EMMC and took place in Vienna, 7th April 2025, 1-6 pm CET.

Time/CET	Speaker	Topic
12:00 – 1:00	Registration & Networking	
1:00 – 1:15	Nadja Adamovic (TU Wien, AT) **MatCHMaker**	Welcome & Introduction
1:15 – 1:45	Alexandre Ouzia (Heidelberg Materials, DE), Geoffrey Daniel (CEA, FR) and Sophie Schmid (TU Wien, AT) **MatCHMaker**	*"Characterisation and modelling data workflows for low carbon cement optimisation."*
1:45 – 2:15	Yoav Nahshon (Fraunhofer IWM, DE) **AID4GREENEST**	*"Interoperable CHADA: A Semantic Approach for Managing and Exploiting Characterisation Data and Protocols"*
2:15 -2:45	Julian de Marchi (NLR-Netherland Aerospace Centre, NL) **D-STANDART**	*"Case study: CHADA v2 population with materials characterisation data from the D-STANDART project"*
2:45 – 3:15	Emanuele Ghedini (Unibo, IT) **CoBRAIN**	*"EMMO for Manufacturing: The CoBRAIN Knowledge Base for Thermal Spraying Process, Modelling and Characterisation"*
3:15 – 3:45	Coffee break	
3:45 – 4:15	Jesper Friis (SINTEF, NO) **MatCHMaker**	*"CHADA template for scanning electron microscopy and how to represent it in the knowledge base"*
4:15 – 4:45	Franz Martin Rohrhofer (Know Center Research GmbH, AT) **KNOWSKITE-X**	*"Managing Data for Machine Learning: The KNOWSKITE-X Approach to a Shared and Adaptive Knowledge Base"*
4:45 – 6:00	Discussion & Networking	

10. Appendix B: Pre-workshop Survey

Survey on current practices, needs, and bottlenecks regarding data lifecycle analysis

Section 1: General Information

Question 1.1. Which project are you affiliated with?

- AddMorePower
- D-STANDART
- AID4GREENEST
- Knowskite-X
- CoBrain
- MatCHMaker

Question 1.2. What is your primary role in the project?

- Data Manager
- Project Coordinator
- Model Developer
- Researcher
- Other*

Question 1.3. Which stage(s) of the data lifecycle are you primarily involved in? (Select all that apply)

- Planning
- Analysing
- Acquisition
- Preserving
- Processing
- Sharing

Question 1.4. How would you describe your familiarity with the integration of modelling and characterisation workflows?

- Novice
- Intermediate
- Advanced

Section 2: Current Practices

Question 2.1. What characterisation methods or materials modelling techniques does your project primarily use?

Question 2.2. How do you typically document and manage the workflows for integrating modelling and characterisation?
Standard operating procedures
Informally through notes/email
Custom protocols
The practice is not implemented yet

Question 2.3. What tools or platforms are currently being used to handle data within your project? (Select all that apply)

- Excel/Spreadsheets
- Data repositories (e.g., Zenodo, institutional data)
- LIMS (Laboratory Information Management System)
- Custom scripts/software
- Other*

Section 3: Challenges and Bottlenecks

Question 3.1. What are the main challenges in integrating modelling and characterisation workflows?

Question 3.2. Have you faced bottlenecks in the following areas? (Select all that apply)

- Data interoperability between modelling and characterisation tools
- Lack of standardisation in data formats
- Scalability of data storage and management
- Difficulty in visualising integrated data
- Other*

Question 3.3. What proportion of your project's effort is spent on addressing these bottlenecks?

- Less than 10%
- 10-25%
- 25-50%
- More than 50%

Section 4: Needs and Opportunities

Question 4.1. What features or tools would most improve your workflows for integrating characterisation and modelling?

Question 4.2. Would your project benefit from standardised approaches or shared tools across projects in the consortium?

- Yes
- No
- Maybe

Question 4.3. Are there specific training or resources you think are missing for your team?

Section 5: Collaboration and Feedback

Question 5.1. Have you collaborated with sister projects on shared workflows or tools?

- Yes
- No
- Not yet, but we are planning to

Question 5.2. Would you be interested in contributing to or using a shared repository of best practices for integrating data workflows?

- Yes
- No

- Maybe

Question 5.3. Any additional comments or suggestions regarding integrating modelling and characterisation or the workshop itself?

Section 6: Specific Data Lifecycle Stages

PLANNING

Question 6.1. Do you create detailed data management plans (DMPs) before starting a project?
- Yes
- No
- Partially
- The practice is not implemented yet

Question 6.2. What challenges do you face in planning data management for characterisation projects?

ACQUISITION

Question 6.3. What tools or methods do you use to collect characterisation data? Additional help available automated instruments, manual measurements, external datasets, etc.?

Question 6.4. How do you ensure data quality during acquisition?
- Calibrated equipment
- Standard operating procedures (SOPs)
- Real-time monitoring
- The practice is not implemented yet
- Other*

Question 6.5. Are there any recurring issues during data acquisition?

PROCESSING

Question 6.6. What tools or software do you use for data pre-processing? Additional help available custom scripts, third-party software, vendor-specific tools, etc.

Question 6.7. How do you handle missing or noisy data?
Manual correction
Automated tools
The practice is not implemented yet
Other*

Question 6.8. What challenges do you encounter during data processing?

ANALYSING

Question 6.9. What statistical or computational techniques do you use most frequently for analysis?

Question 6.10. Do you document your analytical workflows for reproducibility?
- Yes
- No
- Partially

- The practice is not implemented yet

Question 6.11. Are there any gaps or needs in your analysis workflows?

PRESERVING

Question 6.12. Where do you archive characterisation data after the project ends?
- Local storage
- Institutional repository
- Cloud services
- The practice is not implemented yet
- Other*

Question 6.13. Do you use metadata standards for preserved data?
- Yes
- No
- Partially
- The practice is not implemented yet

Question 6.14. What issues do you face in long-term data preservation?

SHARING

Question 6.15. How do you share data with collaborators or other stakeholders?
- e-mail
- Cloud-based platforms
- Institutional repositories
- Public databases
- The practice is not implemented yet
- Other*

Question 6.16. Do you use licenses to govern how your shared data can be reused?
- Yes
- No
- The practice is not implemented yet

Question 6.17. What barriers prevent or hinder data sharing in your context?

Section 7: Organisational Practices and Challenges

Question 7.1. What are the main factors preventing you or your organisation from fully implementing data lifecycle strategies? (Select all that apply)
- Limited time and resources
- Lack of appropriate tools or infrastructure
- Insufficient knowledge or training on data lifecycle practices
- Data management is a lower priority compared to other tasks
- Uncertainty about which tools or methods to use
- Organisational resistance to adopting new practices
- Other*

Question 7.2. Do you feel that you and your colleagues have adequate time to dedicate to data lifecycle activities?

- Yes, we have sufficient time
- Somewhat, but it could be improved
- No, we lack adequate time
- The practice is not implemented yet

Question 7.3. How would you describe your organisation's support for data lifecycle management?

- Strong support with clear policies and resources
- Moderate support but lacking in some areas
- Minimal support; it's mostly up to individual teams
- No support; data lifecycle management is not a focus
- The practice is not implemented yet

Question 7.4. What initiatives would help you and your organisation better apply data lifecycle strategies? (Select all that apply)

- Access to specialised tools and software
- Training and professional development opportunities
- Clear guidelines and best practices documentation
- Allocation of dedicated time for data management tasks
- Increased collaboration with other teams or projects
- Management recognition of the importance of data lifecycle
- Other*

Question 7.5. In your opinion, what are the key benefits of effectively implementing data lifecycle practices in your projects?

Question 7.6. Are there any specific challenges unique to your team or organisation that affect data lifecycle management?