Digital Transformation in Materials Science: A User Journey of Nanoindentation, Image Analysis and Simulations

RESEARCH PAPER

HANNA TSYBENKO
SARATH MENON
FEI CHEN
ABRIL AZOCAR GUZMAN
KATHARINA GRÜNWALD
STEFFEN BRINCKMANN
TILMANN HICKEL
TIM DAHMEN
VOLKER HOFMANN
STEFAN SANDFELD
RUTH SCHWAIGER

*Author affiliations can be found in the back matter of this article

ABSTRACT

A robust digital infrastructure, built upon overarching frameworks and software tools, is essential for the ongoing digital transformation in materials science and engineering. This user journey demonstrates the seamless integration of distinct technical solutions for data handling and analysis, enabling (a) the pursuit of a specific scientific question and (b) adherence to FAIR principles. The scientific study selected for this user journey focuses on comparing different measures of the elastic modulus of a typical engineering material. The user journey involves three research groups replicating real-world collaborative research scenarios. Specifically, it integrates existing digital solutions for experimental data management (PASTA-ELN), simulation workflow execution (pyiron), and image processing workflow execution (Chaldene). Within the auxiliary data management workflow, generated data and metadata are systematically stored in repositories, with metadata aligned to the MatWerk Ontology. Key insights from this user journey include lessons learned from scientists' perspectives and recommendations for improvement, such as machine-readable experimental protocols, standardized workflow representation, and automated metadata extraction.

CORRESPONDING AUTHOR:
Steffen Brinckmann
Forschungszentrum Jülich, Institute of Energy Materials and Devices - Structure and Function of Materials (IMD-1), 52425 Jülich, Germany
s.brinckmann@fz-juelich.de

KEYWORDS:
FAIR; workflow; materials science

TO CITE THIS ARTICLE:
Tsybenko, H., Menon, S., Chen, F., Guzman, A.A., Grünwald, K., Brinckmann, S., Hickel, T., Dahmen, T., Hofmann, V., Sandfeld, S. and Schwaiger, R. 2025 Digital Transformation in Materials Science: A User Journey of Nanoindentation, Image Analysis and Simulations. *Data Science Journal* 24: 33, pp. 1–18. DOI: https://doi.org/10.5334/dsj-2025-033

1 INTRODUCTION

Recent decades of research data management in materials science have been shaped by several key factors. First, materials science has increasingly become a data-driven discipline within the engineering sciences, driven by the rapid accumulation of heterogeneous data that often varies in format, quality, and quantity (Rodrigues *et al.*, 2021; Scheffler *et al.*, 2022). Second, the field is inherently multi- and interdisciplinary, with data generated and exchanged across experimental and computational workflows. These workflows typically involve multiple collaborating teams with diverse expertise, enabling thorough interpretation and validation of research findings. Third, widespread recognition of the reproducibility crisis in academic research (Baker, 2016) has spurred substantial community efforts to establish new standards and practices aimed at enhancing the transparency and repeatability of both data and scientific workflows. Furthermore, the potential for materials data to be reused beyond its original purpose—such as in computer-aided materials discovery—significantly enhances its value (DeCost *et al.*, 2020; Himanen *et al.*, 2019). This potential underscores the need for research data to be findable, machine-readable, and accessible following publication.

These challenges were partially addressed through the establishment of the FAIR (Findable, Accessible, Interoperable, Reusable) principles—a set of data management criteria designed for the scientific community (Go-FAIR, 2024; Wilkinson *et al.*, 2016). A core emphasis is placed on rich, standardized, and systematically documented metadata to enhance the ability to discover and exchange data, and reusability. Additionally, globally unique persistent identifiers (PIDs) and clearly defined data access protocols further support these principles. While the original guidelines primarily targeted (meta)data as research outputs, recent initiatives have expanded their scope to encompass research software (Barker *et al.*, 2022; Chue Hong *et al.*, 2022) and entire scientific workflows (Celebi *et al.*, 2020; de Visser *et al.*, 2023; Goble *et al.*, 2020; Nicolae *et al.*, 2023; Wilkinson *et al.*, 2022). These principles provide an essential theoretical framework for reproducible research practices (Deutsche Forschungsgemeinschaft, 2022), but their practical implementation depends on available resources, existing standards, and the specific requirements of research domains, funding agencies, and institutions.

In practice, FAIR data management is facilitated by various software tools and technologies across the research life cycle. Electronic laboratory notebooks (ELNs) support metadata documentation while centralizing (meta)data storage from various experimental sources. Research software for numerical modeling and data analysis often integrates FAIR data management components within comprehensive computational frameworks. These tools frequently interface with workflow management systems for formulating, scheduling, and executing computational workflows. Metadata schemas and ontologies further enrich and unify semantic dataset descriptions through standardized terms and formalized relationships. Data repositories enable data publication, preservation, discovery, and sharing by providing storage capabilities and assigning PIDs. Collectively, these tools form the backbone of digital research data infrastructures in materials science (Scheffler *et al.*, 2022).

For any specific research project, the sequential use of software tools can be considered a customizable workflow. However, such implementations present distinct challenges for researchers: software may require specific hardware setups for accessibility within laboratory environments, or it may lack interoperability because of different file formats and different metadata nomenclature. Additional adoption barriers include the need for user training, software customization, and limited usability of user interfaces (Higgins *et al.*, 2022; Kanza *et al.*, 2017). These challenges are amplified in collaborative research projects, where increased data volumes and diversified software complicate workflow management. Therefore, during software development, it is crucial to assess user interactions within realistic workflows that generate, analyze, and share data. One widely used method for identifying user needs and potential issues is user journey mapping (Stickdorn and Schneider, 2012), an approach rooted in agile software development practices.

In this article, we implement a user journey to investigate scientists' subjective experiences and perspectives when using various software tools to facilitate transparent, reproducible workflows and produce FAIR materials science data. This user journey encompasses three scientific workflows executed by collaborating groups to address an overarching research question, alongside an external data management workflow for FAIR (meta)data

storage, exchange, and publication. To achieve this, we employ a suite of software and technologies supported by the NFDI-MatWerk Consortium. These include solutions for experimental research data management (PASTA-ELN, 2024), image processing workflow execution (Chaldene (Chen *et al.*, 2022)), and simulation workflow execution (Janssen *et al.*, 2019; pyiron, 2024). The data management platform Coscine and a GitLab (2024) repository were utilized for storing and sharing workflow outputs. Furthermore, metadata from these workflows was aligned with the MatWerk Ontology (2024), converted into both machine- and human-readable formats, and ultimately integrated into the MSE Knowledge Graph (2024).

Through this user journey, we gain insights into how scientists interact with these tools and navigate the various stages of research. We also identify specific challenges encountered when managing research data in collaborative projects involving multiple integrated software solutions. By learning from these experiences, we aim to better align software design and functionality with user requirements, enhance usability, and optimize data production pipelines. Ultimately, we demonstrate a realistic problem-solving process facilitated by research data management software. We include a glossary at the end of the paper to provide definitions of key materials science terms used in the subsequent sections.

2 BUILDING BLOCKS OF THE USER JOURNEY

This work investigates the scientific application of software tools, focusing on their role in facilitating collaborative research workflows. The goals of the user journey implementation can be defined as follows:

- **Follow the scientists:** Track the collaborative efforts of researchers working towards solving a scientific challenge in the domain of materials science. In this user journey, the goal is to determine Young's modulus of Aluminum by experiments and simulations. As multiple methods and analysis tools are employed, interoperability represents one of the major challenges.
- **Promote FAIR principles:** Employ solutions that advance Findable, Accessible, Interoperable, and Reusable (FAIR) data practices and workflows.
- **Document workflows and interactions:** Record the resulting workflows and investigate the scientists' points of interaction with the software (see section 2.2) and other user journey personas (see section 2.3).
- **Analyze user experiences:** Evaluate the scientists' experiences to identify existing challenges and opportunities for optimization.

2.1 APPROACH

The general modeling approach for the user journey focuses on representing how materials scientists interact with software tools to generate FAIR data. This approach emphasizes realistic research practices and working conditions, covering the main stages of a scientific project, including the end-to-end pipeline of data handling throughout the project life cycle. This life cycle encompasses data collection, data processing, data analysis, data storage, and data sharing. By following these steps, the research project additionally follows the FAIR principles, as detailed in the Appendix. Certain steps, such as the reuse of published data, were intentionally excluded, as they would initiate a new iteration of the research cycle with a different set of agents, thereby extending the user journey beyond its original scope. A crucial aspect of recreating authentic research conditions is to frame a problem suitable for a highly collaborative environment. For instance, a comparative study requires multiple scientific methods, engagement from several research groups, and extensive data exchange. Moreover, integrating experimental and computational methods to compare and validate results aligns with the project's interdisciplinary nature. This report provides only a brief description of the individual software tools, as they are expected to evolve significantly in the future. The primary focus is on the scientists' workflow rather than the tools themselves. Additionally, these tools lack dedicated interoperability features. Therefore, this study examines the inherent interoperability of structured data and explores how the absence of purpose-built software interoperability impacts research.

The research question follows the state of the art of experimental and computational materials science. It focuses on comparing the elastic properties of an aluminum alloy (EN AW-1050A, 99.5 wt.% Al), a standard alloy grade used for sheet metal work. The elastic modulus is determined through three different methods corresponding to three distinct workflow components:

1. **Experimental workflow:** Indentation-based measurements of a metal sample and the evaluation of Young's modulus by the Oliver-Pharr method.
2. **Data analytic workflow:** Image processing of confocal images and determining the contact area from the height profiles using the Sneddeon equation.
3. **Computational workflow:** Molecular statics simulations to determine the energy of different atomistic configurations and evaluate the elastic moduli.
4. **Data management workflow:** Handling external data to demonstrate effective collaboration, data storage, and metadata harmonization.

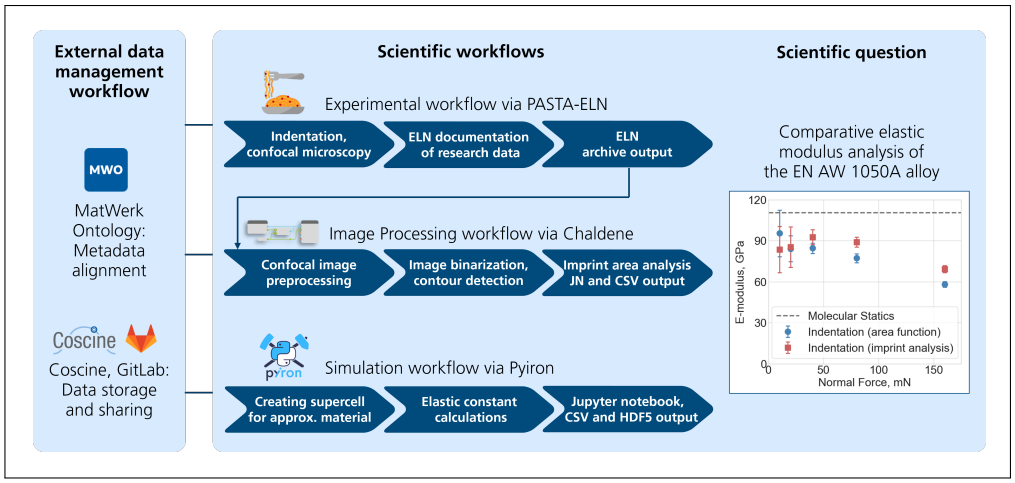

Figure 1 Overview of the workflows in the user journey and its scientific challenge. The process begins with the preparation of aluminum samples, which are then deformed using nanoindentation. The resulting nanoindentation data is used to calculate Young's modulus. To determine the contact area, the indentation imprint is measured using confocal microscopy, and the images are analyzed accordingly. Finally, molecular statics simulations are conducted to compute the energy for different configurations, allowing for the calculation of Young's modulus for the aluminum alloy.

This third workflow component underscores the importance of efficient data management among collaborators. Figure 1 provides an overview of the main user journey components.

2.2 SOLUTIONS FOR DATA MANAGEMENT

To achieve the objective of quantifying the elastic properties of an aluminum alloy, three classes of software tools are required. First, experimental data must be curated and organized; ELNs provide structured data capture and provenance. Second, quantitative image analysis performed by domain scientists requires advanced data-processing environments—Jupyter Notebooks support flexible analyses but present a usability barrier for researchers without Python experience. Third, determination of elastic constants from many energy-minimization simulations requires scalable workflow engines to orchestrate and parallelize large simulation ensembles. To promote accessibility and reuse, we use an ontology to define a shared, machine-readable vocabulary (classes, properties, and constraints) and a knowledge graph to store and link instance-level data across workflows. The ontology ensures semantic interoperability, while the knowledge graph enables integrated queries, discovery, reasoning, and provenance tracing across experimental, image-analysis, and simulation datasets—directly advancing the FAIR principles. We selected the specific software tools to reflect the expertise and development priorities of the contributing teams, allowing evaluation and improvement of real-world implementations; this emphasizes that the architecture is tool-agnostic and can be realized with different toolchains.

PASTA-ELN (2024) is an open-source ELN software designed to provide a centralized framework for research data management during experimental workflows. It focuses on organizing and analyzing raw experimental data stored on the researcher's hard drive. Two key features are the automated (meta)data capture by extractor add-ons during data file integration into ELN projects. These metadata entries can be fully annotated according to the scientific domain as well as the FAIR principles. Additionally, PASTA-ELN automatically generates

digital representations of traditional laboratory notebooks. The adaptable structure allows the software to be used in a wide range of scientific domains.

Chaldene (Chen *et al.*, 2022) is a visual programming language for data analysis and scientific image processing workflows, built upon Jupyter Notebook features for interactive development and the scikit library for image processing. In Chaldene, the code blocks are represented as visual programming cells that can be assembled into directed acyclic graphs using nodes and connectors. This approach facilitates the creation and automatic execution of computational workflows that use a PNG or JPEG file for input.

pyiron (2024) and Janssen *et al.* (2019) is an open-source, Python-based framework for computational workflows, providing functionalities for developing and executing various simulation tasks. Interactive development is facilitated through the Jupyter Notebook interface. The software includes a large number of simulation tools for materials scientists out of the box. pyiron employs "jobs", objects of an abstract class, to standardize the steps and elements of simulation protocols.

GitLab (2024) is a platform for collaborative development, code management, and data version control, enabling users to track changes to workflow outputs. However, GitLab does not guarantee their long-term data preservation (GitLab, 2024). To address this, workflow outputs were also stored in the Coscine (2024) repository developed at RWTH Aachen University. Coscine can store files of any type with a file size limit of 4TB. Each project and resource is assigned a Persistent Identifier (PID) and can be archived for 10 years, ensuring long-term data findability and accessibility in accordance with the principles of good scientific practice set by the German Research Foundation (2022).

The MatWerk Ontology (2024) provides a standardized vocabulary and structure for metadata originating from diverse sources. It enables the harmonization of resource descriptions at the mid-level by offering terms such as *projects*, *authors*, *repositories*, *software*, *instruments*, and *methods*. Additionally, it includes material-specific metadata, such as *material type*, *property*, and *condition*. These standardized descriptions can populate the Materials Science and Engineering (MSE) Knowledge Graph, which is exemplified as a demonstrator and can be queried using the conventional SPARQL format. The Knowledge Graph represents the entire collected knowledge in an interconnected manner and facilitating enhanced discovery.

2.3 PERSONAS

To explore user interactions with the software tools during workflow execution, we define characteristic portraits of users representing the target groups, that is to say, personas. Modeling personas helps us better understand users' motivations, needs, and behaviors in relation to studied products or services.

The primary target group consists of *Researcher* personas, who fulfill multiple roles:

1. Project planning and execution: *Researchers* contribute to project planning and ensure timely completion of tasks.
2. Data production and analysis: They generate, collect, and analyze (meta)data, which is crucial for deriving research findings.
3. Data management and FAIRification: *Researchers* share responsibility for managing datasets and ensuring their compliance with FAIR principles to enable future reuse.

In these tasks, *Researchers* rely heavily on software solutions and the support from Research Data Management (RDM) Agents while adhering to community standards and institutional or funding agency requirements. Their main concerns regarding software tools include seamless integration into scientific workflows, ease of use, and efficient data exchange across research groups. Collaborative projects are typically cross-group or cross-institutional, with each group comprising one or more *Researchers*. To capture the diversity in expertise and domain knowledge, we introduce three *Researcher* personas reflecting different scientific disciplines and levels of data management proficiency. This approach allows us to examine (i) software workflows (e.g., a single *Researcher* using a single tool) and (ii) interactions of workflows at intersections (e.g., collaboration among multiple *Researchers*). The interaction of the Researcher personas is evaluated based on user feedback from the three individuals.

The second key user group comprises RDM Agents, who specialize in providing technical support for software solutions and act as a bridge between FAIR technology and *Researchers*. While *Researchers* drive the three scientific workflows, RDM Agents play a pivotal role in the external data management workflow. The RDM Agent persona typically comes from software development teams or has expertise in data stewardship. Their main objectives include:

1. Understanding *Researchers'* specific requirements and workflows
2. Tailoring software solutions according to these needs
3. Handling registration services, metadata structuring, and unification

Since *Researchers* often actively contribute to software development or are among its core users, they may also take on RDM Agent roles. Thus, personas can be assigned to both roles simultaneously. However, since different RDM Agents do not interact with one another, a single RDM Agent persona is sufficient to represent this user journey.

3 WORKFLOW EXECUTION

In the following sections, we visualize the workflows and elaborate on the steps *Researcher* personas take to accomplish their tasks while interacting with each other and the software. Each workflow starts with *Researcher* personas planning and selecting the necessary methods and procedures. At this stage, they also define the expected output data and metadata formats and determine appropriate storage and publication strategies.

3.1 EXPERIMENTAL WORKFLOW

The experimental workflow is represented from the perspective of *Researcher I* (Figure 2). Instead of detailing the experimental methods, we focus on *Researcher I*'s interactions with PASTA-ELN and other *Researcher* personas.

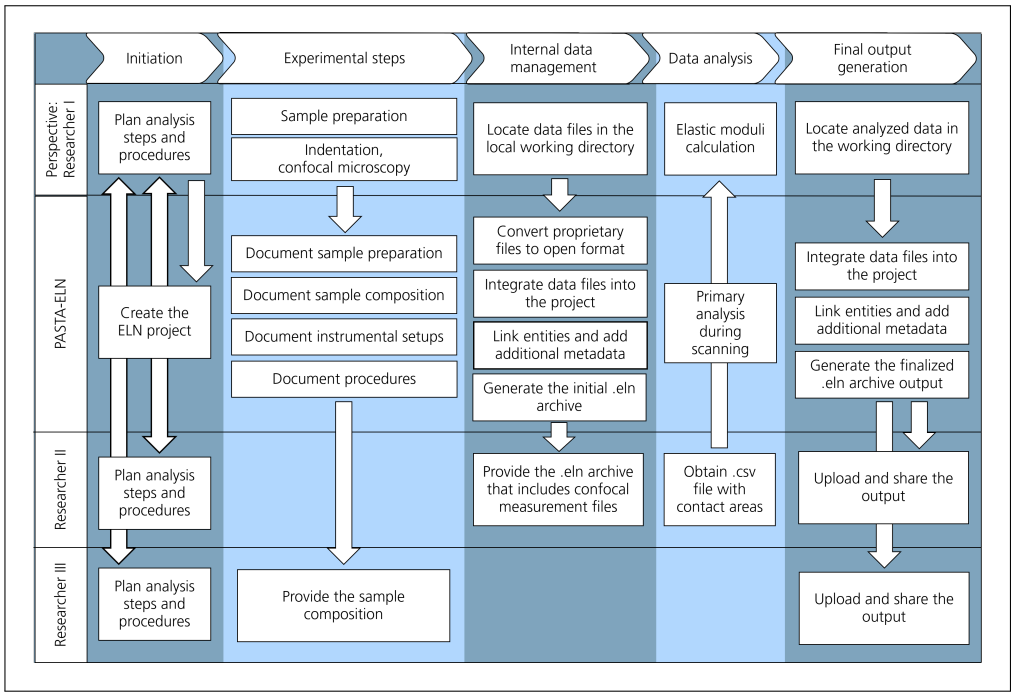

Figure 2 Experimental workflow within the user journey, illustrating the perspective of *Researcher I*, who uses PASTA-ELN for research data management. The workflow comprises five main tasks—initiation, experimental steps, data management, data analysis, and output—each with subtasks that connect to other *Researchers* and the PASTA-ELN software. Documentation and annotation subtasks ensure workflow provenance.

PASTA-ELN serves as a centralized platform for data management, documentation, preliminary analysis, and packaging of research outputs for publication as digital objects. Each manual activity performed by *Researcher I* is digitally recorded, with corresponding metadata linked to the relevant data files. At the planning stage, *Researcher I* initializes an ELN project, which automatically creates a corresponding working directory in the file system. This directory functions as a drop-box for data files.

After each step, a corresponding digital instance is manually created in the ELN project. PASTA-ELN enables users to generate instances and integrate files related to *Samples*, *Procedures*, *Instruments*, and *Measurements*. In this case, *Researcher I* creates instances for:

- Specimen body (EN AW-1050A)
- Measurement procedures (indentation, confocal microscopy)
- Instruments (Fischerscope H100 C, LEXT OLS4000)

Metadata is recorded in two formats:

- Structured format (machine-readable)
- Unstructured format (freehand comments)

Upon completion of this stage, *Researcher I* shares the specimen composition with *Researcher III*, initiating the simulation workflow.

At this stage, *Researcher I* integrates the collected data files into the ELN project by seamlessly creating digital instances of experimental objects or procedures and efficiently linking them. Since indentation and confocal microscopy data files are in proprietary formats (HAP and LEXT), they must be converted into open formats—HDF5 (Hierarchical Data Format version 5) and GWY—for further analysis. This conversion is performed using converter add-ons and Nečas and Klapetek (2012)'s work. Deciphering proprietary file formats is the biggest challenge in experimental workflows that use ELNs. To maintain data integrity, both proprietary and converted open-format files are stored within the ELN project directory. *Researcher I* then scans the working directory, which triggers the following processes:

1. Files are integrated into the ELN project as digital instances.
2. Hierarchical and descriptive metadata is saved to the Apache CouchDB (2024).
3. File extractors identify compatible formats and automatically extract data and metadata into the project.

Researcher I links the *Measurement* instances with their corresponding *Sample* and *Procedure* instances and manually adds additional metadata (Figure 3).

To initiate the image processing workflow, *Researcher I* exports the confocal microscopy images from the ELN project and shares them with *Researcher II*. The exported ELN file is a ZIP archive-like bundle containing all project metadata and data, designed for interoperability across all different ELN platforms. The underlying file structure adheres to the RO-Crate 1.1 standard (Soiland-Reyes *et al.*, 2022), which supports FAIR data publication and can be included into repositories. The export contains (i) the collected data and (ii) the metadata file (ro-crate-metadata.json), which follows the Schema.org (2024) in machine-readable JSON-LD format (Sefton *et al.*, 2023) and is linked to the files in the data collection, thus acting as a table of contents.

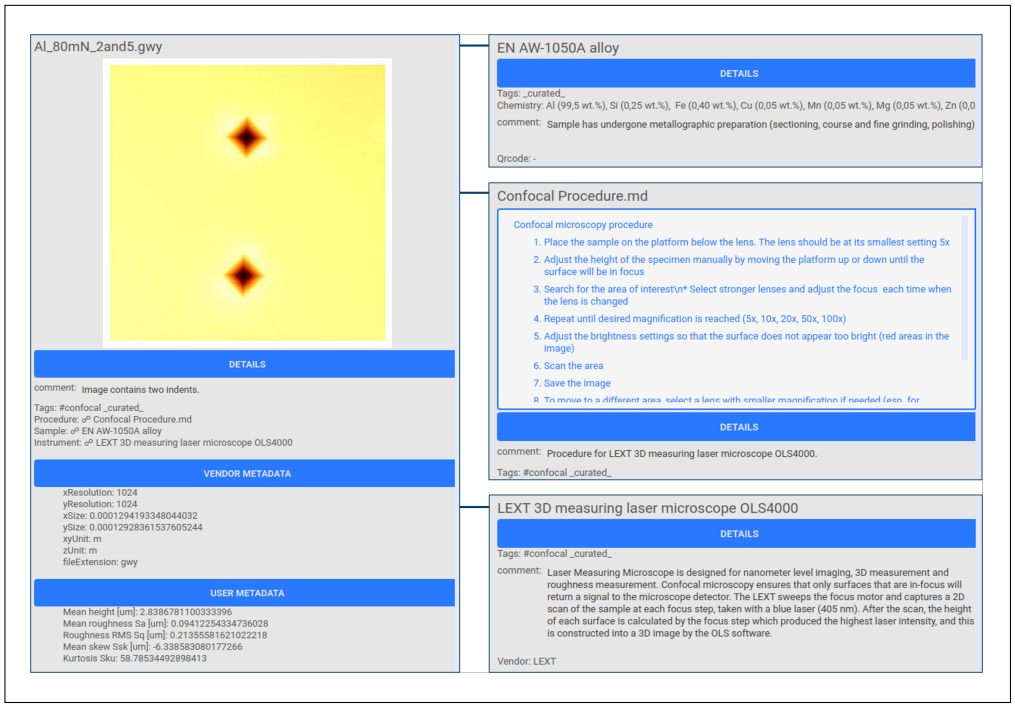

Figure 3 PASTA-ELN project screenshots showing extracted data and metadata from the confocal microscopy GWY file and the linked *Sample*, *Procedure*, and *Instrument* instances. Metadata is categorized into *Details* (general metadata), *Vendor Metadata* (extracted from the measurement file), *User Metadata* (defined by the user), and *Database Metadata* (required for ELN operation; omitted here). The measurement on the left is linked to a sample, procedure, and instrument, displayed on the right.

After receiving the estimated contact areas from the image processing workflow, *Researcher I* proceeds with elastic modulus analysis using the Oliver-Pharr Method (Oliver and Pharr, 1992, 2004). This method involves using projected contact areas to calculate the effective elastic modulus, from which the elastic modulus of the sample is determined. For this analysis step, two data sets of projected contact areas are used: one based on the indenter tip area function and the other one determined via confocal microscopy image analysis. *Researcher I* incorporates the results into the ELN project, adding annotations and linking them to relevant data files. The experimental workflow concludes with the ELN file being uploaded to repositories for data sharing (section 3.4).

3.2 DATA ANALYTICAL WORKFLOW IN CHALDENE

The second scientific workflow in this user journey involves *Researcher II*, who executes an image processing workflow (Figure 4). This workflow is developed interactively in Jupyter Notebook using the Chaldene visual programming language. Its objective is to compute the projected contact areas of indentation imprints based on the confocal microscopy data obtained from *Researcher I*.

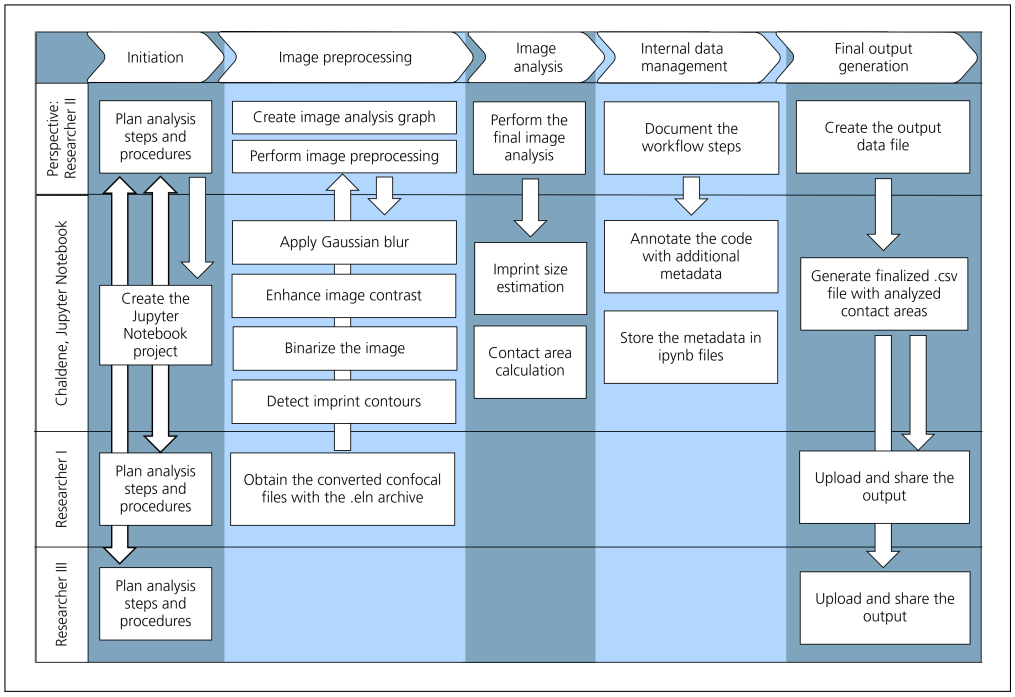

Figure 4 Image processing workflow as a part of the user journey, demonstrating the perspective of *Researcher II*, who uses Chaldene as a scientific image processing tool. The workflow consists of five tasks—initiation, image preprocessing, analysis, data management, and output—each with subtasks that connect to other *Researchers* and integrate with the Chaldene software.

In Chaldene, each workflow step is represented as a cell, equivalent to a code block in Jupyter Notebook. The cell nodes correspond to data states before and after processing and are linked through connectors, forming a graph representation of the workflow (Figure 5a). This approach ensures correct execution order and enables traceability of processing steps.

The first part of preprocessing enhances image quality to improve contour detection for projected contact area analysis. *Researcher II* performs the following steps: (i) Data acquisition and format conversion to extract confocal height data (Figure 5b) from the ELN file and convert proprietary formats into open formats, (ii) Noise reduction and contrast enhancement by applying Gaussian blur filtering for general noise removal and enhancing image contrast (Figure 5c), and (iii) Binarization for contour detection using the Isodata algorithm (van der Walt *et al.*, 2014) to automatically determine a threshold value that separates the imprint from the background pixels (Figure 5d).

The analysis stage includes two steps: (i) Contour extraction using a 'marching squares' algorithm implementation from the scikit-image library (van der Walt *et al.*, 2014) to trace imprint boundaries (Figure 5e). (ii) Area calculation by estimating the size of each contour by counting enclosed pixels and converting pixel-based areas into physical units using the pixel size obtained during file reading. *Researcher II* repeats this workflow for each confocal micrograph.

At the final stage, *Researcher II* annotates the Jupyter Notebook with metadata describing the workflow steps, saves the computed projected contact areas as comma-separated value (CSV) files for further analysis (see section 3.1), and uploads data files to GitLab and Coscine repositories for sharing with other *Researchers* (section 3.4).

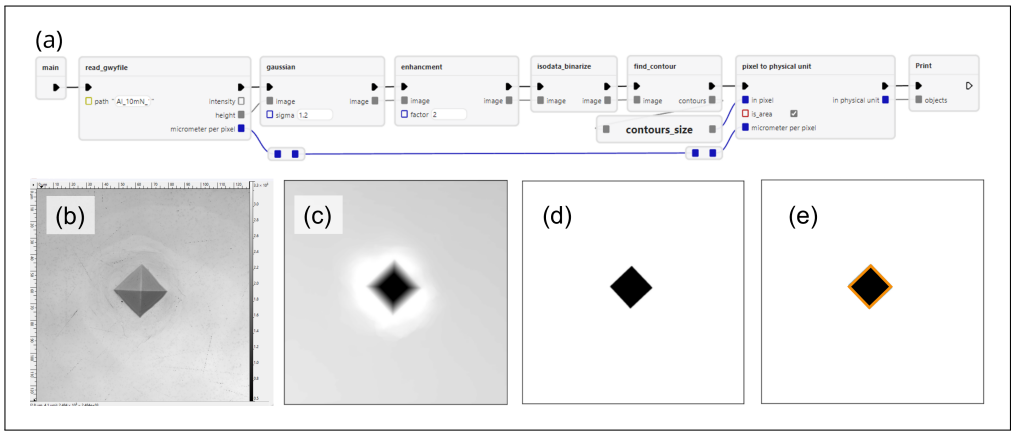

Figure 5 (a) Chaldene image processing workflow represented as a series of interconnected nodes (rectangles) linked by edges (lines). The nodes correspond to key processing stages: data ingestion, Gaussian blur, enhancement, binarization, and physical area evaluation. **(b–e)** Confocal microscopy images illustrating different stages of the workflow: (b) raw light reflection, (c) raw height profile, (d) binarized image using the Isodata algorithm, and (e) binary image contour.

3.3 COMPUTATIONAL WORKFLOW IN pyiron

The simulation workflow, carried out from the perspective of *Researcher III*, is illustrated in Figure 6. At the preprocessing stage, *Researcher III* selects parameters for rapid prototyping of the simulation workflow using Jupyter Notebooks as an interactive environment in pyiron. The composed workflow can later be scaled up for execution on multi-core processors or high-performance computing (HPC) clusters. During the next stage, *Researcher III* performs molecular statics simulations using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) to estimate the elastic constants. This stage consists of three steps: (i) Supercell creation to construct an Al face-centered cubic (FCC) supercell with 5324 atoms (Figure 7) substitutional impurities added to match the composition of EN AW-1050A (0.2 wt.% Si, 0.2 wt.% Fe, 0.05 wt.% Cu, and 0.05 wt.% Mg), (ii) Interatomic Potential Selection, here using a modified Embedded Atom Method (EAM) potential (Jelinek *et al.*, 2012), optimized for this alloy system and ensuring alignment with density functional theory (DFT) calculations for elastic constants of the pure elements and their binary combinations, and (iii) Elastic Tensor Calculation by optimizing the structure with respect to cell parameters and atomic positions, applying small deformations (maximum Lagrangian strain of 0.001) considering the structure's space group. Forces and stresses are recorded to calculate elastic constants. All the workflow steps are annotated with metadata to ensure reproducibility and stored in a Jupyter Notebook.

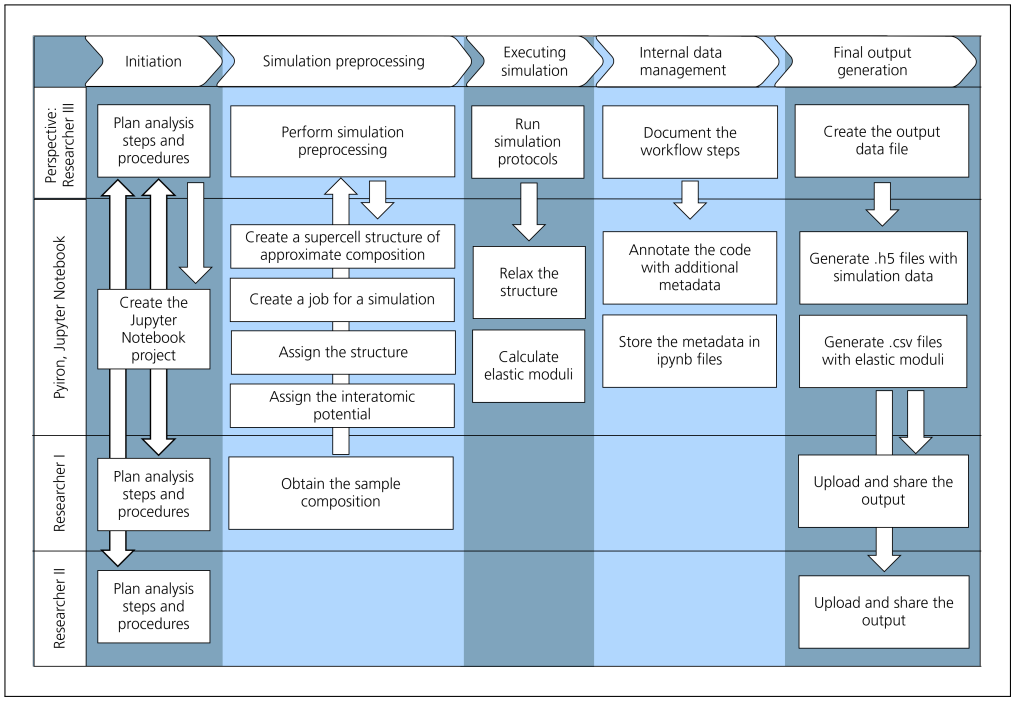

Figure 6 Simulation workflow as a part of the user journey, illustrating the perspective of *Researcher III*, who uses pyiron as the computational framework. The workflow comprises five main tasks—initiation, preprocessing, simulation execution, data management, and output—each with associated subtasks that connect to other researchers, the pyiron software, and its database. The entire workflow is implemented within a Jupyter Notebook.

At the final stage, *Researcher III* saves workflow inputs, outputs, and results in HDF5 files, extracts key results into CSV files for sharing with other *Researchers*, and uploads the workflow, software versions, and HDF5 data files to repositories (section 3.4).

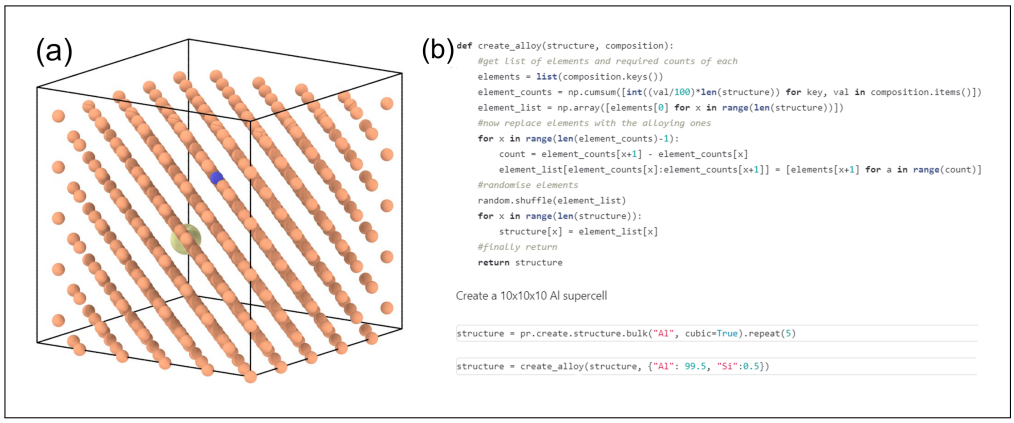

Figure 7 (a) Supercell of Al atoms used in the molecular statics simulation, with randomly placed impurity atoms. The [111] planes of the FCC structure are visible along the viewing direction. **(b)** Jupyter Notebook snippet from the simulation workflow, highlighting the creation of the supercell during preprocessing and the inclusion of impurity atoms.

3.4 METADATA AND STORAGE (EXTERNAL RDM WORKFLOW)

The external data management workflow encompasses the handling of (meta)data beyond the scientific workflows (Figure 8). It primarily focuses on the RDM Agent's role in enabling *Researcher* personas to use storage services, aligned metadata descriptions, and populate a knowledge graph.

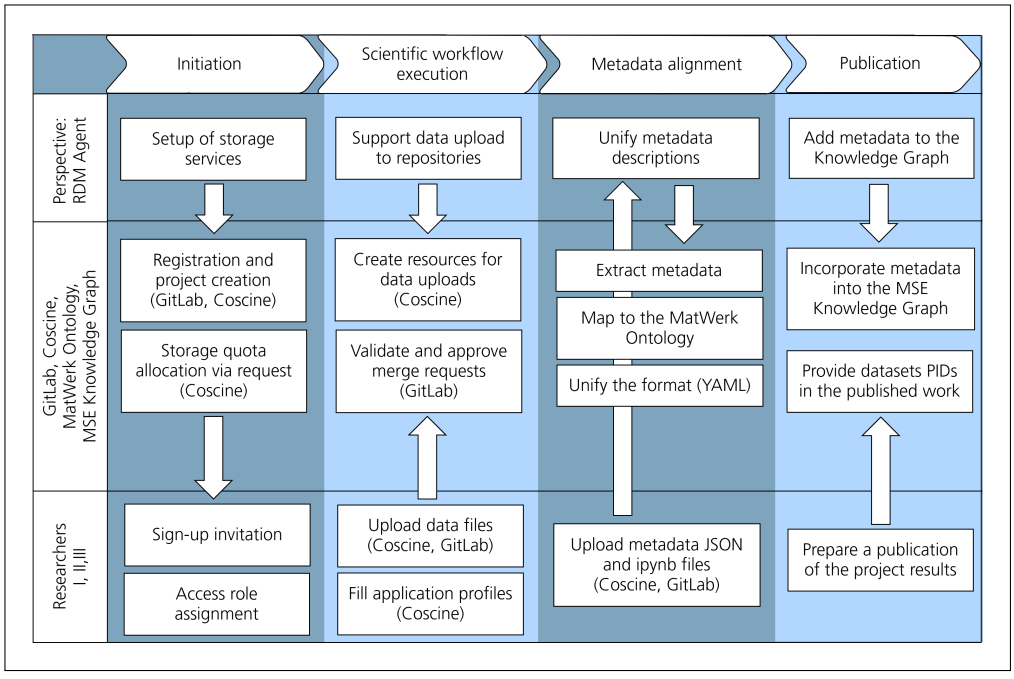

Figure 8 External data management workflow as a part of the user journey. The workflow comprises four main tasks—initiation, data upload support, metadata alignment, and publication within a knowledge graph—along with PIDs of the repository entries. Subtasks connect *Researchers*, repositories, and the knowledge graph.

In the project planning phase, the RDM Agent sets up project instances in Coscine and GitLab repositories. Once the repositories are established, the RDM Agent sends sign-up invitations to *Researchers* and assigns roles based on their respective tasks. For Coscine projects, the RDM Agent also submits an application to request a specific storage quota.

During scientific workflow execution, *Researchers* upload data and metadata files either via a GitLab merge request, approved by the RDM Agent, or directly into pre-created Coscine resources. Each project is assigned a PID, and a corresponding resource is created in Coscine. Additionally, every Coscine resource requires an appropriate application profile, serving a metadata schema in a tabular form, which *Researchers* must manually complete upon data upload. By default, the base application profile—derived from the DataCite metadata schema Kernel 4.0 (DataCite Metadata Working Group, 2016)—is used, containing fundamental metadata fields: Title, Creator, Creation Date, Subject Area, and Type.

The outputs from all three workflows are uploaded to RDS-S3 object-based resources (Tsybenko *et al.*, 2023a), while simulation workflow data is imported directly from Tsybenko *et al.* (2023b). This ensures that all data receives a PID and is annotated with additional metadata.

To ensure semantic consistency across the three scientific workflows and support the FAIR principles, the RDM Agent unifies metadata annotations and formats (Appendix). Experimental workflow metadata is automatically exported as a structured JSON-LD and bundled with data files in a single RO-Crate archive. Computational workflow metadata is combined with analysis code in two Jupyter Notebooks, requiring manual extraction into a structured format. Metadata from domain-specific file formats is then converted into the standard machine- and human-readable YAML Ain't Markup Language (YAML) format using a template (MatWerk Ontology templates, 2024) provided by the RDM Agent (Figure 9a). At the same time, extracted descriptions from the YAML format are aligned with the MatWerk 2.0 Ontology, ensuring semantic harmonization and hierarchical organization of metadata. This alignment enhances interoperability across various materials science domains.

In the final stage, the RDM Agent facilitates the publication of (meta)data, ensuring findability, accessibility, and reusability. Each published work includes the PID (Tsybenko *et al.*, 2023a), improving traceability. Metadata is integrated into the MSE Knowledge Graph (Figure 9b), allowing instances, agents, and activities to be interconnected and discoverable via SPARQL queries. Published metadata includes licensing information, defining usage permissions.

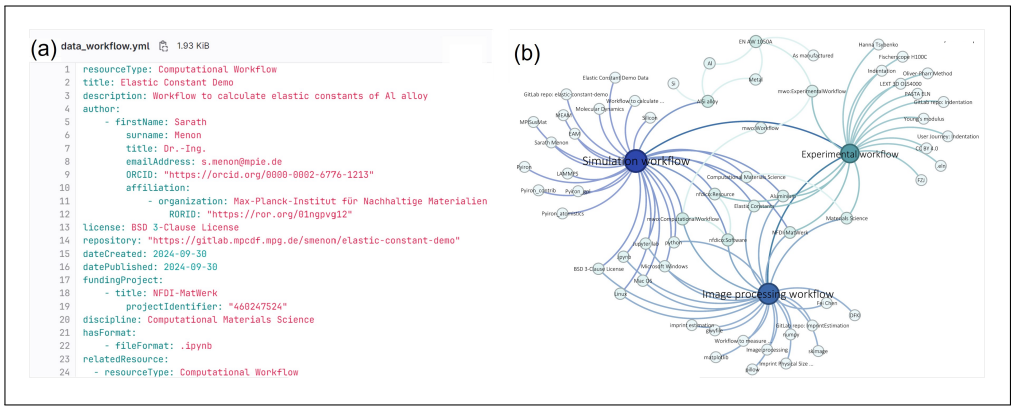

Figure 9 (a) A snippet of the YAML file with metadata from the computational workflow aligned with the MatWerk 2.0 Ontology to highlight that the YAML format is beneficial for manual data entry and automatic processing. **(b)** The visualization of the triples that populated the MSE Knowledge Graph, which highlights the strongly interconnected metadata and the hierarchy of the node-structured data.

4 LESSONS LEARNED

The purpose of this section is to assess the *Researchers*' experiences with FAIR software solutions and RDM Agents during scientific workflow execution. Based on identified limitations, we propose potential improvements for workflow optimization and software improvements. The scientific results of secondary interest are not discussed.

Our evaluation concentrates on a single, well-characterized use case; accordingly, we do not claim exhaustive validation across the full range of materials-science environments. This focused choice enabled an in-depth exploration and critical analysis of a representative scenario within the constraints of the present study. We expect the approach to generalize for two principal reasons: (1) the methodology is modular and readily extensible by substituting specific tools or accommodating alternative workflows, and (2) the core tasks—experimental execution, data analysis, and simulation—are largely domain-agnostic and thus transferable across contexts. Nonetheless, systematic cross-domain validation, encompassing diverse subdomains, experimental configurations, and dataset scales, remains essential. In future work, we plan to benchmark the pipeline on at least two additional use cases and to report the corresponding lessons learned.

4.1 GENERAL OUTCOMES

Implementing the user journey provided valuable insights into the dynamics of collaborative research project and the impact of RDM software. This approach has proven highly effective for analyzing *Researchers*' interactions with software tools, RDM Agents, and one another. From a long-term perspective, the published user journey serves as a blueprint for streamlined

FAIR data generation as it outlines the detailed steps necessary to integrate FAIR principles into research workflows. Additionally, the modular and accessible nature of the implemented workflows facilitates component reuse within the research community.

This study highlights the essential role of NFDI MatWerk Consortium tools (PASTA-ELN, Chaldene, pyiron, Coscine, MatWerk Ontology, and the MSE Knowledge Graph) in solving scientific problems and generating FAIR data (Appendix A). Each tool adheres to multiple FAIR principles, promoting holistic and transparent research data management. Consequently, these tools contribute significantly to enhancing research practices, improving data quality and reusability, and fostering collaboration in materials science.

Our analysis shows that *Researchers* rely heavily on RDM Agents, particularly for repository setup and metadata harmonization. In cases where no RDM Agent is available, *Researchers* often assume dual roles, leading to increased workloads and higher technical demands. Ideally, dedicated data stewards should fill these gaps. However, when unavailable, training programs on FAIR data concepts, available tools, and best practices could mitigate knowledge barriers. Overall, our findings indicate a strong commitment among *Researchers* to FAIR principles, ensuring the publication of FAIR data.

The user journey examined advanced collaborative interactions between *Researchers* and various software solutions. The workflows reflected state-of-the-art scientific approaches, simulating a realistic research project. However, more complex workflows—involving iterative experiments or computation, modified approaches, or additional data requirements—could offer deeper insights into software scalability and adaptability. These scenarios should be explored in future implementations.

After executing the workflow, we interviewed the scientists about the advantages and limitations of the software tools, as well as their interactions within the workflow. We found that no essential software tools were missing for this specific scientific workflow. However, we identified shortcomings in interoperability and feature availability. The limitations identified can lead to enhancements in future versions of the software tools.

We identified limitations in interoperability and feature availability across the evaluated tools. In particular, the tools lack a common input/output file format that would enable non-programming users to export, read, and analyze data. For example, producing the final materials-science overview of elastic modulus versus normal force currently requires programming expertise because data from three distinct sources must be consolidated into a single figure. Future development should therefore enable the complete workflow to be executed without writing code. Achieving this objective will require adoption of a common data-exchange format and, potentially, a standardized workflow-execution framework. Semantic web formats such as RDF offer rich machine-readable semantics that facilitate content discovery and interpretation, but they are not well-suited to packaging very large volumes of instrument or computational data. Conversely, FAIR digital-object containers such as RO-Crate and the ELN file format support rich, machine-actionable semantic metadata, can accommodate multigigabyte datasets, and can explicitly encode inter-resource relationships. However, practical challenges remain—most notably cross-dataset linking and consistent semantic interpretation—which necessitate community conventions and interoperable tooling.

4.2 EXPERIMENTAL WORKFLOW OUTCOMES

Modeling a realistic experimental workflow naturally highlighted several recurring challenges in materials science. A key issue is that experimentalists are often constrained by proprietary instrument software, which can become obsolete, lack technical documentation, and hinder data provenance tracking. *Researcher I* encountered this limitation during the experimental indentation setup. To address it, *Researcher I* exported only raw data from the instrument and meticulously documented its origin within the PASTA-ELN project.

Additionally, the collected data was stored in proprietary formats, preventing access by multiple software vendors and restricting parsing as raw text files. Consequently, experimental data analysis and metadata extraction depended on vendor-specific software. This issue was resolved by using the PASTA-ELN converter add-on, which converted the proprietary files into open formats, ensuring accessibility and interoperability. A potential improvement would be to include more metadata on file encoding formats during export.

Furthermore, automated metadata extraction upon file integration into the ELN project saved *Researcher I* considerable time compared to manual metadata entry. PASTA-ELN functionalities improved data structuring, retrieval, and sharing, allowing *Researchers* to focus on the scientific aspects. However, further workflow optimization could be achieved by introducing user-defined templates for experimental protocols. These templates should support efficient metadata input and automatically convert metadata into machine-readable formats.

Proprietary nanoindentation and confocal microscopy instruments commonly produce data in vendor-specific file formats (e.g., HAP, LEXT), which introduces two principal problems for downstream analysis and long-term preservation. First, exporting to open formats via vendor-supplied software can result in loss of metadata and a reduction in numeric precision (for example, some exports are limited to three significant digits), compromising data fidelity. Second, opaque proprietary files are frequently unreadable by standard analysis tools, impeding reuse and archival access. While LEXT files (a TIFF variant) can be interpreted by third-party readers such as Gwyddion, no public readers were available for HAP files; consequently, we developed a conversion utility to extract both primary data and associated metadata from HAP containers. Deciphering such closed formats, therefore, represents a substantial bottleneck in experimental workflows. To preserve provenance and support reproducibility, we retain both the original proprietary files and the converted open-format derivatives, and we validate conversions by comparing key numerical values and metadata fields between source and output. The conversion tool will be made available on request. Finally, we confirm that the datasets described contain no personal data; had personal data been present, appropriate anonymization and controlled-access procedures would have been implemented. These measures reduce risks to data integrity, privacy, and intellectual property while acknowledging the ongoing challenges posed by opaque proprietary formats.

4.3 COMPUTATIONAL WORKFLOWS

While molecular statics calculations are common in computational materials science, they remain challenging due to the need to select structures, apply strains, compute forces and stresses, and postprocess results. *Researcher III* faced these challenges when using different software tools. To simplify this process, pyiron provides built-in routines, allowing users to focus only on defining structure and strain ranges, significantly improving reproducibility. Additionally, directory structures are automatically created and managed, while metadata is stored alongside the workflow, enhancing reusability and transparency.

A major challenge in computational research is the lack of workflow standardization. While both pyiron and Chaldene use Jupyter Notebooks and Python, they are not inherently interoperable, making it difficult for *Researchers* to reuse code. Existing standards such as Common Workflow Language (CWL) and Nextflow provide some workflow standardization, but they do so at the expense of Python's object-oriented flexibility and rapid prototyping capabilities.

Currently, pyiron allows exporting Python code as a pyiron workflow, but this requires extra effort from *Researchers*. A more efficient solution would be a generic Python-based workflow description that is semantically interoperable across pyiron, Chaldene, and potentially other platforms such as PASTA-ELN.

Another challenge is that workflow input/output keywords are often domain-specific, making them difficult to interpret outside specialized fields. A potential solution is to incorporate terminology services or ontologies that define key terms, improving workflow clarity and FAIR compliance. Additionally, metadata records should include details on authors, institutions, and software dependencies. While the manual metadata extraction and alignment with the MatWerk Ontology partially address this issue, further improvements are needed (see section 4.4).

4.4 METADATA AND ONTOLOGY

Human- and machine-readable metadata are essential for enhancing the findability, interoperability, and reusability of research data. However, data complexity and heterogeneity pose challenges for accurate metadata recording. In this study, ELN exported files from experiments and Jupyter Notebooks from computational studies contained research data, workflows, analysis results, and documentation. While manual metadata entry using a

YAML template was feasible, it was also time-consuming and impractical for large datasets. Automated metadata extraction tools should be developed and integrated into user-facing software to address this issue.

The MSE Knowledge Graph serves as a centralized resource for querying domain knowledge, improving resource visibility and accessibility. However, populating the knowledge graph currently requires manual input. Implementing automated metadata harvesting from repositories and storage services would significantly enhance efficiency.

While the MatWerk Ontology helps harmonize metadata and improve findability and interoperability, it lacks domain-specific terminology, which is essential for precise concept definitions. A viable solution is to combine high-level ontology alignment with detailed, domain-specific metadata, ensuring that workflows are understandable beyond specific research domains. Recent efforts to develop materials science semantic artifacts could further support ontology integration (MatPortal: the ontology repository for materials science, 2024).

5 CONCLUSIONS

This study implemented a user journey simulating a realistic collaborative research project, integrating three scientific workflows and an external data management workflow. The results demonstrate how NFDI-MatWerk Consortium tools—PASTA-ELN, Chaldene, pyiron, Coscine, the MatWerk Ontology, and the MSE Knowledge Graph—enable transparent, reproducible research data management. Additionally, the findings emphasize the critical role of RDM Agents in providing technical support and ensuring adherence to FAIR principles.

While this work highlights the benefits of integrating RDM software, challenges remain, particularly in handling proprietary data formats, ensuring computational workflow interoperability, and streamlining metadata entry. Future improvements should prioritize standardizing workflows across platforms and integrating domain-specific terminology services. These advancements will enhance the FAIRness, transparency, and efficiency of scientific workflows and large-scale projects beyond the specific workflow examined in this study.

APPENDIX

Table A1 Description of the FAIR data principles implementation in the finalized datasets published in GitLab and Coscine repositories.

	FAIR DATA PRINCIPLES	DESCRIPTION OF IMPLEMENTATION
Findable	F1. (Meta)data are assigned a globally unique and persistent identifier	All the Coscine resources with uploaded (meta)data files are automatically assigned with an identifier that is globally unique and persistent. The registry service responsible for assigning and resolving the PIDs of digital objects is the Handle.Net Registry (HNR).
	F2. Data are described with rich metadata	The metadata files are bundled with the data files and contain the resources' descriptions (discipline, funding project, title, publication date, etc.). The files are in human- and machine-readable YAML format, which improves the findability of the resources.
	F3. Metadata clearly and explicitly include the identifier of the data they describe	The metadata files include the PID URL to the Coscine resources they are referring to as well as the URL to the location of the resource within the GitLab repository.
	F4. (Meta)data are registered or indexed in a searchable resource	The visibility of the Coscine project and the resources is set to 'public', therefore, the files are listed in a Coscine-wide search for the appropriate (meta)data. The project can also be searched for in the GitLab repository. In addition, the (meta)data can be discovered by SPARQL-querying of the MSE Knowledge Graph.
Accessible	A1. (Meta)data are retrievable by their identifier using a standardized communications protocol	Users can use the contact form (available via PID URL) to obtain permission from the project owner and access the (meta)data of the Coscine resources. The (meta)data are also publicly available on GitLab. In both cases, the (meta)data are retrievable through HTTP(S).
	A1.1 The protocol is open, free, and universally implementable	The http(s) is free and open and can be implemented globally to retrieve the (meta)data.
	A1.2 The protocol allows for an authentication and authorization procedure, where necessary	Coscine Users can authenticate to search for the project-specific (meta)data. Authentication and authorization are not required to access the GitLab project and the related (meta)data.
	A2. Metadata are accessible, even when the data are no longer available	The assigned PIDs of the Coscine resources allow the metadata to be long-term accessed.

(Contd.)

	FAIR DATA PRINCIPLES	DESCRIPTION OF IMPLEMENTATION
Interoperable	I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	The metadata stored in the YAML files is aligned at the top level to the MatWerk 2.0 Ontology, which uses a formal knowledge representation language OWL. The MatWerk Ontology specification is published online and is accessible to the community.
	I2. (Meta)data use vocabularies that follow FAIR principles	The MatWerk Ontology aims to follow FAIR principles. For instance, its classes and properties have globally unique and persistent identifiers (IRIs) that can be resolved using a standardized communication protocol (HTTP(S)), and it uses a formal, accessible, shared, and broadly applicable language for knowledge representation OWL.
	I3. (Meta)data include qualified references to other (meta)data	Metadata files include meaningful links to other related entities, such as applied ontology for metadata descriptions, affiliated organizations and authors, as well as the applied software package.
Reusable	R1. (Meta)data are richly described with a plurality of accurate and relevant attributes	The metadata clearly describes the content of the data. It includes license information under which data can be reused and information about the data creation context.
	R1.1. (Meta)data are released with a clear and accessible data usage license	The (meta)data includes usage rights information (BSD 3-Clause License for the source code and Creative Commons Attribution 4.0 International License for other (meta)data).
	R1.2. (Meta)data are associated with detailed provenance	The provenance metadata in YAML files includes the authorship information, date of creation, employed scientific methods, applied software, and instruments. In addition, the ELN file includes the descriptions of instruments and procedures in the experimental workflow, whereas the Jupyter Notebooks contain simulation workflow descriptions.
	R1.3. (Meta)data meet domain-relevant community standards	The dataset contains files in well-established formats common in materials science projects, for example, MD, CSV, HDF5, GWY. The datasets are organized as data and metadata bundles. The metadata files use the templates for standard descriptions and are all aligned to the MatWerk Ontology, which represents the research activities in Materials Science.

GLOSSARY

General terms

- Alloy: A metal composed of two or more chemical elements that form a single or multiphase system with distinct properties.
- Confocal microscopy: An optical imaging technique that acquires high-resolution, three-dimensional surface topography.
- Crystal plane [111]: A crystallographic plane identified by Miller indices (1,1,1); in many face-centered cubic metals this plane is associated with prominent slip.
- Elasticity: Reversible deformation in which the material returns to its original shape upon removal of the applied load.
- Face-centered cubic (FCC): A metallic crystal structure in which atoms occupy the corners and face centers of the cubic unit cell.
- Impurity atoms/substitutional impurities: Chemical elements present at low concentrations that occupy lattice sites of the host metal.
- Nanoindentation: An experimental technique for measuring mechanical properties by driving a rigid probe (often a diamond tip) into a specimen.
- Plasticity: Permanent, non-recoverable deformation that remains after removal of the applied load, arising from mechanisms such as dislocation motion.
- Tip area function: A mathematical description relating the contact depth of the probe or tip to the projected contact area.
- Weight percent (wt.%) composition (e.g., 99.5 wt.% Aluminum): Mass fraction expressed as a percentage (e.g., 99.5 wt.% Al indicates 99.5 mass percent aluminum in the alloy).

Simulation terms

- Density functional theory (DFT): A quantum mechanical method for computing ground-state electronic structure and derived properties.
- Embedded Atom Method (EAM) potential/interatomic potential: A class of semi-empirical many-body potentials used to model metallic bonding.
- Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS): An open-source, parallel molecular dynamics simulation package for atomistic modeling.

- Molecular statics/molecular statics simulations: An energy-minimization simulation in which atomic positions are relaxed to local minima.
- Supercell/supercell creation: Construction of a periodically repeated simulation cell containing multiple unit cells (and any defects or solute atoms).

Mechanics of materials

- Contact area: The projected area of physical contact between the indenter and the specimen surface during indentation.
- Elastic constants/elastic modulus/Elastic tensor: Quantities that relate stress to the elastic strain in the material.
- Hardness: A measure of resistance to plastic deformation, defined as the maximum applied load divided by the projected contact area.
- Oliver–Pharr method: A widely used analysis procedure for instrumented indentation data that extracts hardness and the reduced elastic modulus.
- Strain: A measure of relative deformation of the material defined as the change in length (or displacement) normalized by the original dimension.
- Stresses: Internal forces distributed over a certain area within a material arising from externally applied loads or constraints.

File formats

- CSV: Comma-separated values, a simple, open, text-based tabular data exchange format.
- GWY: File format used by Gwyddion and some instruments for surface topography data; an open format.
- HAP: Proprietary data format of nanoindentation equipment by the company Fischer.
- HDF5: Hierarchical Data Format version 5, a portable, self-describing binary format.
- MD: Generic designation for molecular dynamics input or data files.

ACKNOWLEDGEMENTS

The authors thank Velislava Yonkova for conducting indentation experiments and confocal microscopy measurements.

FUNDING INFORMATION

This work is part of the consortium NFDI-MatWerk, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure—NFDI 38/1—project number 460247524.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Hanna Tsybenko orcid.org/0000-0001-7691-2856
Forschungszentrum Jülich, Institute of Energy Materials and Devices - Structure and Function of Materials (IMD-1), 52425 Jülich, Germany

Sarath Menon orcid.org/0000-0002-6776-1213
Max Planck Institute for Sustainable Materials (MPI-SusMat), Computational Materials Design, Germany

Fei Chen orcid.org/0000-0001-7890-0330
German Research Center for Artificial Intelligence (DFKI), Germany

Abril Azocar Guzman orcid.org/0000-0001-7564-7990
Forschungszentrum Jülich, Institute for Advanced Simulation - Materials Data Science and Informatics (IAS-9), Germany

Katharina Grünwald orcid.org/0000-0001-7550-552X
RWTH Aachen University, IT Center, Germany

Steffen Brinckmann orcid.org/0000-0003-0930-082X
Forschungszentrum Jülich, Institute of Energy Materials and Devices - Structure and Function of Materials (IMD-1), 52425 Jülich, Germany

Tilmann Hickel orcid.org/0000-0003-0698-4891
Max Planck Institute for Sustainable Materials (MPI-SusMat), Computational Materials Design, Germany; Materials Informatics, Bundesanstalt für Materialforschung und -prüfung, Germany

Tim Dahmen orcid.org/0000-0003-4060-7192
German Research Center for Artificial Intelligence (DFKI), Germany; Fakultät Elektronik und Informatik, Hochschule Aalen, Germany

Volker Hofmann orcid.org/0000-0002-5149-603X
Forschungszentrum Jülich, Institute for Advanced Simulation - Materials Data Science and Informatics (IAS-9), Germany

Stefan Sandfeld orcid.org/0000-0001-9560-4728
Forschungszentrum Jülich, Institute for Advanced Simulation - Materials Data Science and Informatics (IAS-9), Germany

Ruth Schwaiger orcid.org/0000-0001-8940-2361
Forschungszentrum Jülich, Institute of Energy Materials and Devices - Structure and Function of Materials (IMD-1), 52425 Jülich, Germany

REFERENCES

Apache CouchDB (2024) *Apache CouchDB*. Available at: https://couchdb.apache.org/ (Accessed: 29 May 2024).

Baker, M. (2016) '1,500 scientists lift the lid on reproducibility', *Nature*, 533, pp. 452–454. Available at: https://doi.org/10.1038/533452a

Barker, M., Chue Hong, N., Katz, D., Lamprecht, A.L., Martinez-Ortiz, C., Psomopoulos, F., Harrow, J. *et al.* (2022) 'Introducing the FAIR Principles for research software', *Scientific Data*, 9, p. 622. Available at: https://doi.org/10.1038/s41597-022-01710-x

Celebi, R., Moreira, J., Hassan, A., Ayyar, S., Ridder, L., Kuhn, T. and **Dumontier, M.** (2020) 'Towards FAIR protocols and workflows: the OpenPREDICT use case', *PeerJ Computer Science*, 6, p. e281. Available at: https://doi.org/10.7717/peerj-cs.281

Chen, F., Slusallek, P., Müller, M. and **Dahmen, T.** (2022) 'Chaldene: Towards Visual Programming Image Processing in Jupyter Notebooks', *2022 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, pp. 1–3. Available at: https://doi.org/10.1109/VL/HCC53370.2022.9832910

Chue Hong, N., Katz, D., Barker, M., Lamprecht, A.L., Martinez, C., Psomopoulos, F., Harrow, J. *et al.* (2022) 'FAIR Principles for Research Software (FAIR4RS Principles)'.

Coscine (2024) *Coscine*. Available at: https://about.coscine.de/en/ (Accessed: 29 May 2024).

DataCite Metadata Working Group (2016) *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data v4.0*. Tech. rep., DataCite e.V.

de Visser, C., Johansson, L., Kulkarni, P., Mei, H., Neerincx, P., van der Velde, J., Horvatovich, P. *et al.* (2023) 'Ten quick tips for building FAIR workflows', *PLoS Computational Biology*, 19, p. e1011369. Available at: https://doi.org/10.1371/journal.pcbi.1011369

DeCost, B., Hattrick-Simpers, J., Trautt, Z., Kusne, A., Campo, E. and **Green, M.** (2020) 'Scientific AI in Materials Science: a Path to a Sustainable and Scalable Paradigm', *Machine Learning: Science and Technology*, 1. Available at: https://doi.org/10.1088/2632-2153/ab9a20

Deutsche Forschungsgemeinschaft (2022) Guidelines for Safeguarding Good Research Practice. Code of Conduct. Deutsche Forschungsgemeinschaft

German Research Foundation (2022) *Leitlinien zur Sicherung guter wissenschaftlicher Praxis*. Available at: https://doi.org/10.5281/zenodo.3923601

GitLab (2024) *GitLab*. Available at: https://about.gitlab.com/ (Accessed: 29 May 2024).

Go-FAIR (2024) *Go-FAIR Initiative*. Available at: https://www.go-fair.org/ (Accessed: 29 May 2024).

Goble, C., Cohen-Boulakia, S., Soiland-Reyes, S., Garijo, D., Gil, Y., Crusoe, M., Peters, K. *et al.* (2020) 'FAIR Computational Workflows', *Data Intelligence*, 2, pp. 108–121. Available at: https://doi.org/10.1162/dint_a_00033

Higgins, S., Nogiwa-Valdez, A. and **Stevens, M.** (2022) 'Considerations for implementing electronic laboratory notebooks in an academic research environment', *Nature Protocols*, 17, pp. 179–189. Available at: https://doi.org/10.1038/s41596-021-00645-8

Himanen, L., Geurts, A., Foster, A. and **Rinke, P.** (2019) 'Data-driven materials science: status, challenges and perspectives', *Advanced Science*, 6, p. 42. Available at: https://doi.org/10.1002/advs.201900808

Janssen, J., Surendralal, S., Lysogorskiy, Y., Todorova, M., Hickel, T., Drautz, R. and **Neugebauer, J.** (2019) 'pyiron: An integrated development environment for computational materials science', *Computational Materials Science*, 163, pp. 24–36. Available at: https://doi.org/10.1016/j.commatsci.2018.07.043

Jelinek, B., Groh, S., Horstemeyer, M., Houze, J., Kim, S., Wagner, G., Moitra, A. *et al.* (2012) 'Modified embedded atom method potential for Al, Si, Mg, Cu, and Fe alloys', *Physical Review B*, 85. Available at: https://doi.org/10.1103/PhysRevB.85.245102

Kanza, S., Willoughby, C., Gibbins, N., Whitby, R., Frey, J., Erjavec, J., Zupančič, K. *et al.* (2017) 'Electronic lab notebooks: can they replace paper?', *Journal of Cheminformatics*, 9, p. 31. Available at: https://doi.org/10.1186/s13321-017-0221-3

MatPortal: the ontology repository for materials science (2024) *MatPortal: the ontology repository for materials science*. Available at: https://matportal.org/ (Accessed: 03 October 2024).

MatWerk Ontology templates (2024) *MatWerk Ontology templates*. Available at: https://git.rwth-aachen.de/nfdi-matwerk/ta-oms/mste (Accessed: 03 October 2024).

MSE Knowledge Graph (2024) *MSE Knowledge Graph*. Available at: https://demo.fiz-karlsruhe.de/matwerk/ (Accessed: 29 May 2024).

Nečas, D. and **Klapetek, P.** (2012) 'Gwyddion: an open-source software for SPM data analysis', *Open Physics*, 10, pp. 181–188. Available at: https://doi.org/10.2478/s11534-011-0096-2

Nicolae, B., Islam, T., Ross, R., van Dam, H., Assogba, K., Shpilker, P., Titov, M. *et al.* (2023) 'Building the I (Interoperability) of FAIR for Performance Reproducibility of Large-Scale Composable Workflows in RECUP', *2023 IEEE 19th International Conference on e-Science (e-Science)*. IEEE, pp. 1–7. Available at: https://doi.org/10.1109/e-Science58273.2023.10254808

Oliver, W. and **Pharr, G.** (1992) 'An improved technique for determining hardness and elastic modulus using load and displacement sensing indentation experiments', *Journal of Materials Research*, 7, pp. 1564–1583. Available at: https://doi.org/10.1557/JMR.1992.1564

Oliver, W. and **Pharr, G.** (2004) 'Measurement of hardness and elastic modulus by instrumented indentation: Advances in understanding and refinements to methodology', *Journal of Materials Research*, 19, pp. 3–20. Available at: https://doi.org/10.1557/jmr.2004.19.1.3

PASTA-ELN (2024) *PASTA-ELN*. Available at: https://github.com/PASTA-ELN/pasta-eln (Accessed: 29 May 2024).

pyiron (2024) *pyiron*. Available at: https://github.com/pyiron (Accessed: 29 May 2024).

Rodrigues, J., Florea, L., de Oliveira, M., Diamond, D. and **Oliveira, O.** (2021) 'Big data and machine learning for materials science', *Discover Materials*, 1, p. 12. Available at: https://doi.org/10.1007/s43939-021-00012-0

Scheffler, M., Aeschlimann, M., Albrecht, M., Bereau, T., Bungartz, H.J., Felser, C., Greiner, M. *et al.* (2022) 'FAIR data enabling new horizons for materials research', *Nature*, 604, pp. 635–642. Available at: https://doi.org/10.1038/s41586-022-04501-x

Schema.org (2024) *Schema.org*. Available at: https://schema.org/ (Accessed: 03 October 2024).

Sefton, P., Carragáin, E.Ó., Soiland-Reyes, S., Corcho, O., Garijo, D., Palma, R., Coppens, F. *et al.* (2023) *RO-Crate Metadata Specification 1.1.3*, Zenodo.

Soiland-Reyes, S., Sefton, P., Crosas, M., Castro, L., Coppens, F., Fernández, J., Garijo, D. *et al.* (2022) 'Packaging research artefacts with RO-Crate', *DS*, 5, pp. 97–138. Available at: https://doi.org/10.3233/DS-210053

Stickdorn, M. and **Schneider, J.** (2012) *This is Service Design Thinking: Basics, Tools, Cases*. Wiley.

The MatWerk Ontology (MWO) (2024) *The MatWerk Ontology (MWO)*. Available at: http://purls.helmholtz-metadaten.de/mwo (Accessed: 29 May 2024).

Tsybenko, H., Menon, S., Chen, F., Guzman, A., Grünwald, K., Brinckmann, S., Hickel, T. *et al.* (2023a) *Project Data: Elastic properties of EN AW-1050A alloy: a scientific user journey built upon NFDI-MatWerk infrastructure solutions*. Available at: http://hdl.handle.net/21.11102/4781e163-229d-4a2a-91c2-75e107c21730 (Accessed: 29 May 2024).

Tsybenko, H., Menon, S., Chen, F., Guzman, A., Grünwald, K., Brinckmann, S., Hickel, T. *et al.* (2023b) *Project Data: Elastic properties of EN AW-1050A alloy: a scientific user journey built upon NFDI-MatWerk infrastructure solutions*. Available at: https://git.rwth-aachen.de/nfdi-matwerk/ta-wsd1/user-journey-indentation (Accessed: 17 October 2024).

van der Walt, S., Schönberger, J., Nunez-Iglesias, J., Boulogne, F., Warner, J., Yager, N., Gouillart, E. *et al.* (2014) 'scikit-image: image processing in Python', *PeerJ*, 2, p. e453. Available at: https://doi.org/10.7717/peerj.453

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N. *et al.* (2016) 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, 3, p. 160018. Available at: https://doi.org/10.1038/sdata.2016.18

Wilkinson, S.R., Eisenhauer, G., Kapadia, A., Knight, K., Logan, J., Widener, P. and **Wolf, M.** (2022) 'F*** workflows: when parts of FAIR are missing', in *2022 IEEE 18th International Conference on e-Science (e-Science)*, Salt Lake City, UT, USA, pp. 507–512. Available at: https://doi.org/10.1109/eScience55777.2022.00090

TO CITE THIS ARTICLE:
Tsybenko, H., Menon, S., Chen, F., Guzman, A.A., Grünwald, K., Brinckmann, S., Hickel, T., Dahmen, T., Hofmann, V., Sandfeld, S. and Schwaiger, R. 2025 Digital Transformation in Materials Science: A User Journey of Nanoindentation, Image Analysis and Simulations. *Data Science Journal* 24: 33, pp. 1–18. DOI: https://doi.org/10.5334/dsj-2025-033

Submitted: 05 May 2025
Accepted: 27 October 2025
Published: 17 November 2025

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.