CLUSTER FORSCHUNGSDATEN

SHARED RDM

FAIR&AI

AI and ML potential in RDM: Opportunities and action areas

[Authors: Alexander Bardel, Claire Jean-Quartier, Thomas Seyffertitz, Ilire Hasani-Mavriqi]

1. Abstract

The Research Data Lifecycle spans from research planning to data collection, processing, publication, archiving, and reuse. Each phase presents specific challenges that can be addressed through Artificial Intelligence (AI) and Machine Learning (ML). Based on a systematic mapping of stakeholder involvement, we identify and prioritize AI/ML applications with the highest potential impact on efficiency, quality, and FAIR compliance. We identify strategic action areas where AI/ML could support future shared RDM services and where no mature solutions currently exist. The increasing availability of generative AI and machine learning (AI/ML) methods raises new opportunities and challenges for research data management (RDM) [i]. The Shared RDM Services and Infrastructure project was launched to establish a framework that provides selected tools and infrastructures as shared services for Austrian universities and research institutions [ii]. Building on the project's original focus on interoperability, FAIR principles [iii], and the bundling of expertise across institutions, a dedicated working group was established in 2025 to explore potential intersections between AI technologies and the research data cycle. In this paper, we present a structured mapping of AI/ML opportunities along the stages of the research data lifecycle, complemented by cross-cutting issues, such as metadata interoperability, legal compliance, and quality assurance. For each stage, we identify typical activities, relevant stakeholders, gaps in current practices, potential AI/ML applications, and their expected impact. We highlight "action areas," such as the absence of mature tools for cross-domain metadata annotation and AI-supported license advising, where new shared RDM services could be developed. AI-based models and tools could serve as an incentive for existing research in the field of AI/ML, while at the same time paving the way for new, innovative approaches to be investigated in a specific sub-field of computational science or AI research. The analysis provides an avenue for both, use cases that lead directly to quick wins and long-term strategies for integrating AI into national and European research data infrastructures. We argue that AI can act as an enabling layer for FAIR and Open Science, if service design, governance, and responsible use remain in focus.

Keywords: Research Data Management (RDM), Shared Services, Generative AI, Machine learning, FAIR Principles, Metadata Interoperability, Open Science

2. Introduction

Research data management (RDM) is becoming increasingly important for researchers and requires a variety of supporting tools that can be used throughout the lifecycle of research data [iv]. In recent years, technical and organizational advancements have been made that are now being applied in various fields of research (e.g. research data repositories, analysis platforms, electronic laboratory notebooks, and data stewardship programs). The aim of the project Shared RDM [ii] is to create a framework to offer selected tools and infrastructures in the field of RDM as shared services for selected Austrian universities and research institutions. This bundling of different expertise creates a sensible use of resources and promotes interoperability, standardization and the connection to international initiatives. The project is carried out in the spirit of the EOSC [v] initiative and thus contributes to an even more reliable and easier re-use of research output. It creates a landscape of national RDM infrastructures and services that can be used as a use case/success story at international level and increases Austria's visibility.

The technology leap brought about by the increasing availability of large language models, generative AI in general, machine learning (AI/ML) methods, and the ongoing development of AI-based tools is now beginning to reshape research data management [i]. As a project, we have responded to this development by establishing a dedicated working group aiming to identify intersections between AI technologies and the research data cycle, and to provide guidance for their strategic integration into RDM services thereby focusing on the infrastructures and services developed within the project. We present a structured mapping of AI/ML opportunities across the stages of the research data lifecycle, complemented by transversal themes such as metadata interoperability, legal compliance, and quality assurance. For each stage, typical activities, stakeholders, gaps, potential applications, and impact levels are assessed. Based on this analysis, we propose a set of action areas.

In section 3 we provide the methodological framework for the following analysis. The results will be presented in section 4, and discussed in a broader context in section 5. In the final section 6 conclusions for further work on these issues will be presented.

3. Methodology

The identified AI/ML applications in RDM and consequential action areas were derived through a combination of literature review, analysis of existing RDM services and tools within Austrian universities as well as gap analysis comparing current capabilities with FAIR requirements, community and expert input gathered at a symposium conducted in the Shared RDM project.

Literature review

Literature was screened and collected via Web of Science database (WOS), using search terms in Title: ("artificial intelligence" and "research data") or ("machine learning" and "research data") or ("AI" and "research data") or ("ML" and "research data") or ("artificial intelligence" and "RDM") or ("machine learning" and "RDM") or ("AI" and "RDM ") or ("ML" and "RDM ") or ("artificial intelligence" and "data management") or ("machine learning" and "data management ") or ("AI" and "data management") or ("ML" and "data management ") which resulted in 145 hits on 2026-04-14. Among those references there were 10 review articles. Only 69 references were Open Access articles. These were further refined to “Article”, “Review article”, resulting in 50 documents. The reference list is available as supplementary file at https://doi.org/10.3217/4qm20-mze24.

Based on those documents, literature search was expanded manually using Google Scholar and OpenAlex to expand on and develop the key aspects along the research data lifecycle as indicated in Figure 1. The framework consists of six Research Data Lifecycle phases [vi]: Research Planning, Data Collection, Data Processing & Analysis, Data Sharing & Publishing, Data Archiving, and Data Reuse.

Mapping and gap analysis

All stages of the research data cycle were initially mapped with exemplified AI/ML applications from the base literature and extended with examples from the experience of authors integrating transversal themes such as metadata interoperability, legal compliance, and quality assurance. For each stage, typical activities, stakeholders, gaps, potential applications, and impact levels are complemented by all authors during repeated discussions rounds. Stakeholders have been grouped into researchers, research support, IT infrastructure, publishers, and externals, such as collaboration partners, third party consultants, external funders and suppliers. Stakeholder categories are outlined in the overview

summarizing AI/ML applications along the research data lifecycle, and are exemplified in more detail in the respective subsection of results.

Exemplary AI/ML applications are evaluated by the authors based on three key characteristics: (1) stakeholders, (2) impact (efficiency, quality, FAIR alignment), and (3) related risks. As a final step, authors completed with a gap analysis through comparison with existing tools and use cases at their associated as well as collaborating institutions and delineated action areas derived from internal discussion rounds. The research data lifecycle model was used to structure the results. integrating respective activities, processes and technical requirements at each stage. For each gap identified in the different phases of the research data lifecycle, a description can be provided using three elements: (1) description of the situation, (2) gap description, (3) related risk and related potential.

Community and expert input

In addition to the literature review, this work was informed by discussions during the FAIR & AI Symposium organized in 2025 [vii]. The symposium brought together researchers, data stewards, infrastructure providers, legal experts, and policy stakeholders to discuss opportunities and challenges at the intersection of FAIR data and AI. The event highlighted recurring themes that informed the analysis presented in this paper, including metadata quality and interoperability, trustworthy and explainable AI, legal and ethical compliance, data quality assessment, AI governance, reproducibility, and the role of research support services. While the identified action areas are not a direct outcome of the symposium, the discussions provided valuable practical perspectives that complemented the literature review and expert assessments conducted within the Shared RDM project.

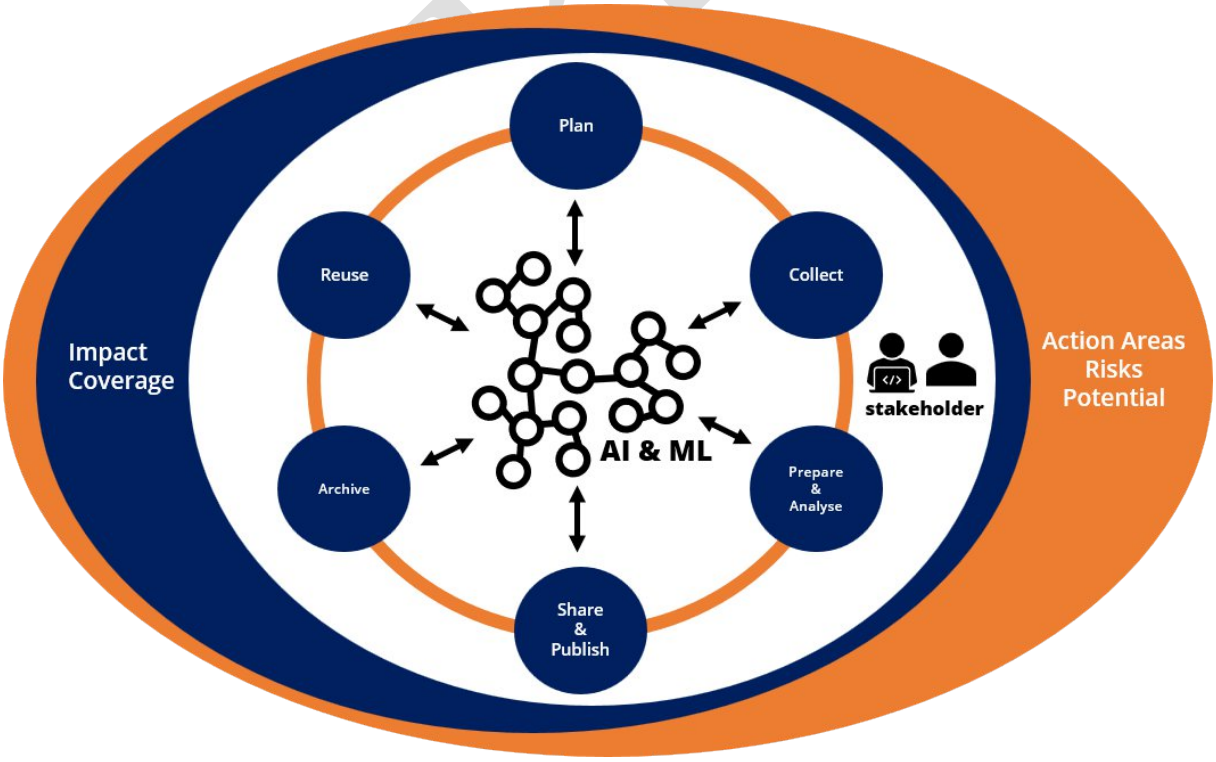

Figure 1: Graphical abstract: Mapping of AI/ML applications along the research data cycle included stakeholders, gaps, risks and potentials.

4. Analysis of AI/ML Potentials in the Research Data Lifecycle

In this section examples are listed along the research data lifecycle for the steps of research planning, data collection, data preparation and analysis, data sharing and publishing, data archiving, and data reuse.

Most exemplified AI/ML approaches for or by RDM tasks have been suggested for the phases of data collection and data preparation, such as metadata and data mining and cleaning [viii, ix, x, xi, xii, xiii], data selection and management [xiv, xv, xvi, xvii], followed by examples for supported data cleaning and labeling as well as feature and model selection, deployment and debugging [xviii]. Other examples of automated risk assessment and report generation [xix] could be attributed f.i. to research data planning.

The following subsections indicate an overview on AI/ML applications along the research data lifecycle, summarized in Table 1.

	Challenges	Stakeholders	AI/ML application	Impact	Risks	Example Reference
Planning	DMP, tools & methods, ethics & legal compliance, resources	researchers, research support, IT infrastructure, externals	automated DMP generation, risk and resource forecasts	accelerates planning, high level integration into funding portals	false interpolation or mapping	xxi, xxii
Collection	experimental setup, quality control, documentation	researchers, research support, IT infrastructure	Real-time anomaly and outlier detection, smart sampling & calibration	reduces expensive repetition of experiments, increases data integrity	false deletion/interpolation or mapping	xxvi, xxvii, xxviii
Preparation /Analysis	data cleaning, metadata enrichment, versioning & workflow management, evaluation	researchers, research support, externals	automated data cleaning & anomaly detection, NLP-based annotation, AutoML pipelines	time-saving, increased data quality and traceability	false deletion/interpolation or mapping	xxxii, xxxiii, xxxiv, xxxv
Sharing & Publication	Long-term citeability, access & usage conditions, quality assurance & peer review	researchers, IT infrastructure, publishers	automated license suggestions, semantic search & repository recommendation	increased visibility, reuse & citation, promotes open science	possibly missed community relevant infrastructure	xxxvii
Archiving	long-term formats, backup & integrity, security, sustainability planning	IT infrastructure, researchers, publishers, externals	predictive format migration, automated integrity monitoring	long-term security	real-time checks still required	xli
Reuse	Findability, Interoperability, Data citation & tracking, community building	researchers, externals, research support	recommender systems for data sets usage, automated bibliometric analyses	network effect	return on investment under debate	xliii, xliv

Table 1: Examples of AI/ML applications along the research data cycle

	Exemplary roles
Researchers	Research project management and acquisition, experimentation and analysis, students, research assistants
Research Support	RDM consultants, ethics and legal department, technology transfer

IT Infrastructure	Infrastructure s.a. server and cloud operators, administration of software and devices
Publisher	Publishing group, journals, editors, libraries, archive manager, research societies, scientific associations
Externals	Collaboration partners, third party consultants, external funders and suppliers

Table 2: Stakeholder categories and exemplary roles as indicated in the beneath AI/ML applications

4.1. Research planning

Key activities and challenges include the creation of a data management plan (DMP) with tasks as defining data formats, metadata standards, responsibilities, and retention periods. The decision on suitable collection and analysis tools takes interoperability and open source principles into account. The Ethical and legal review is based on data protection (e.g., GDPR), consent forms for human studies, and directions of the ethics committee. Finally, budgeting and securing resources adds tasks as calculating costs for storage space, infrastructure, and personnel.[xx,xxi] Relevant stakeholders, as introduced in Table 2, comprise researchers (management, concept development), research support / RDM consultant (DMP expertise, advice on standards), ethics committee / data protection officer (legal approval), and infrastructure operators (e.g., data center, cloud provider for capacity planning). AI/ML applications for supporting the above listed tasks include automated DMP generation by e.g. a historical template analysis, and risk and resource forecasts on e.g. budget, time, and infrastructure requirements.[xxii,xxiii] These applications facilitate the preparation of grant applications and DMPs. Forecasts reduce budget deviations and infrastructure bottlenecks. In summary, planning is accelerated, but this approach requires a high level of integration into funding portals.

4.2. Collecting and gathering data

Key activities and challenges comprise the experimental setup or study design towards ensuring validity and reproducibility, data collection including sampling, sensor technology, surveys and interviews. Additionally, real-time quality control by validation and plausibility checks during the collection are of importance, as well as the documentation of all steps covering complete logs, and versioning. [xxiv,xxv,xxvi] Relevant stakeholders involve researchers conducting experiments and field work, technical staff and laboratory technicians setting up and maintaining equipment, students or research assistants collecting data and keeping records, as well as infrastructure operators for network and storage for raw data. AI/ML applications supporting the collection process comprise real-time quality control using anomaly and outlier detection, and smart sampling & adaptive instrument calibration.[xxvii, xxviii,xxix] Such applications prevent incorrect measurements and data loss at an early stage, and optimize resource utilization in field studies and sensor networks. In sum, expensive re-measurements are avoided or reduced, and data integrity is increased.

4.3. Data preparation and analysis

Key activities and challenges comprise data cleaning such as handling missing values, outliers, and consistency checks; metadata enrichment by assignment of standardized terms (e.g., via JSON-LD,

Dublin Core); versioning and workflow management, e.g. use of Git, FAIR workflows (Snakemake, Nextflow); and statistical evaluation & visualization: Reproducible scripts (R, Python), automated reports.[xxx,xxxi,xxxii] Relevant stakeholders involve researchers during analysis and interpretation, data stewards and/or the RDM team for quality assurance and metadata standard suggestions, IT support and system administration for setting up analysis and computing environments, and statistics and methodology consultants in case of advanced analyses. AI/ML application examples are given by automated data cleaning and anomaly detection, intelligent metadata generation (NLP-based annotation), workflow optimization and adaptive workflow orchestration (AutoML pipelines).[xxxiii,xxxiv,xxxv,xxxvi] Manual cleaning and standardization are time-consuming and error-prone. Missing or inconsistent metadata slow down FAIR principles. Reproducibility increases through standardized, versioned pipelines. In summary, such applications save time for researchers, and increase data quality and traceability.

4.4. Data sharing and publishing

Key activities and challenges include the selection of suitable repositories (disciplinary vs. institutional archives), assigning persistent identifiers (e.g., DOI) and ensuring long-term citeability, defining access and usage conditions by evaluating open access requirements vs. embargo options, accompanied by a suitable license selection. Additionally, peer review and quality assurance are implicated for data papers, data journals, and review processes.[xxxvii] Relevant stakeholders are researchers as data publishers and reviewers, library and publication services as repository operators and for license advice, publishers and data journals for the publication process, and infrastructure operators for server operation and interfaces. AI/ML application examples are given by automatic license and access optimization through e.g. recommendations for licenses, and semantic search and recommendation engines for repositories.[xxxviii] Such applications increase the visibility of data sets and researchers may find potential digital objects for reuse more quickly. Additionally, optimized license recommendations could prevent legal risks. In sum, open science is promoted increasing reuse and citations.

4.5. Data archiving

Key activities and challenges comprise long-term format migration by avoiding format obsolescence (e.g., TIFF instead of proprietary formats), backup and redundancy through off-site storage and regular integrity checks (checksums), security and access concepts, such as encryption, role and authorization systems, and sustainability planning for funding of ongoing maintenance.[xxxix,xl,xli] Relevant stakeholders are represented by researchers in case of transfer of metadata and infrastructure prioritization, archivists and associated research support for policy development and format migration, infrastructure operators for storage and backup systems , as well as funding bodies regarding the budget for long-term costs. AI/ML application examples include predictive format migration such as the detection of obsolescence risks, and automated integrity monitoring by e.g. anomaly detection in checksum logs.[xlii] Such applications can prevent data loss due to obsolete formats or silent corruption and save time and effort of manual checks in large archives. Altogether, this approach is important for long-term security, but less necessary than real-time checks.

4.6. Data reuse

Key activities and challenges include tasks for findability & discoverability, such as search engine optimization of metadata, and registries (e.g. re3data); interoperability, such as usage and

development of standardized APIs, community services (e.g. Open Geospatial Consortium), and linked data; data citation & tracking, such as the collection of citation metrics and impact analysis; and for training & community building through workshops, and showcasing case studies on successful use.[xliii] Relevant stakeholders include (external) researchers & students, industry & partner organizations in regard to application development, research support and training coordinators for workshops and documentation, as well as funding agencies and politics concerning a proof of social impact. AI/ML applications comprise recommender systems in regard to the question who uses which data sets and clustering of use cases, and automated citation and impact tracking for bibliometric analyses.[xliv,xlv] These examples support (external) users in finding suitable data sets, and enable the measurement and visualization of data impact for e.g. reporting to funding agencies. In summary, benefits are based on the network effect (increased value with increasing participation), still considering an initially lower return of investment.

4.7. Selected Use Cases

The following cross-sectional use cases for RDM applications, introduced in Figure 2, highlight exemplary AI/ML approaches spanning over several steps along the research data lifecycle.

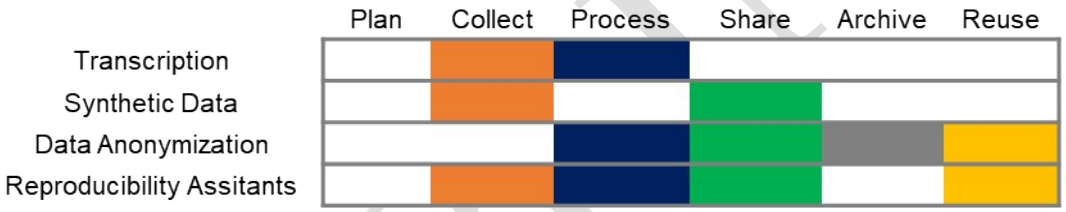

Figure 2: Selected use cases of AI/ML applications in (research) data management along the research data lifecycle

Automatic speech recognition and transcripts for preparing and analysing data

To-Text transcription is a central step in *data preparation* in – but not exclusively – qualitative social research [xlvi]. Two example tools already used in practice are the open-source software noScribe [xlvii] and aTrain [xlviii]. Both tools rely on machine learning models such as pyannote [xlix] or whisper [l] for automatic speech recognition (ASR). The impact of AI-supported transcription technology is manyfold. First, the transcription process is faster, even though it still requires the Human-in-the-loop for quality control. Second, the tools offer an export format allowing for integration in traditional transcription software like MAXQDA and ATLAS.ti [li] to allow researchers to continue directly working in those tools after transcription. Third, the software including its AI-component can be deployed locally on the desktop, hence enabling a GDPR-compliant workflow and addresses privacy concerns. Finally, for researchers with less (financial) resources, the tools mentioned above provide an alternative for automated creation of transcripts in comparison to commercial transcription services. Moreover, both example tools are open source and can be easily set up by the research support staff or by researchers itself.

Use of synthetic data for collecting and gathering data

Although synthetic data (SD) is not a new concept, with steady growth of data-driven research and the emergence of new AI/ML-methods, SD has been gaining increasing attention. SD can be defined as data that has been generated using a purpose-built mathematical model or algorithm, with the aim of

solving one or more data science tasks. [lii] Synthetic datasets mimic real datasets by preserving their statistical properties and the relationships between variables. [liii] Generating or using synthetic data can be achieved in various ways, for example, deploying deep learning architectures such as Generative Adversarial Networks, statistical simulation and others. [liv],[lv] The following examples serve to illustrate areas of applications of synthetic (research) data:

- The recent increase in *data protection regulations* has fuelled the use of synthetic data to mitigate disclosure risk. As incentive for ambitious data collection projects some journal data policies allow authors to meet the replication and data disclosure requirements by providing a synthetic data set developed from the estimated models. [lvi]
- *Licensed (research) data*: a work around in cases where data used are restricted by the data provider, either by contractual restrictions (licenses) or IP-rights restrictions, GDPR compliance or confidentiality concerns [lvii]. Synthetic data is better for privacy preservation than 'anonymised' datasets as removing identifiers is not always enough to safeguard confidentiality [lviii]. For instance, Gonzales et al. identified several use cases for synthetic data in health care research. [lix]
- In the context of *ML* the following key areas are increasingly of particular interest: (i) private data release, (ii) data de-biasing and fairness, and (iii) data augmentation for robustness. [li]

Synthea's [lx] focus is on healthcare simulation with clinical patient records. In compliance with the U.S. department of Health and Human Services it can be deployed either on-premise or as a cloud service. *Synthetic Data Vault* [lxi] allows data generation for tabular data and is designed to work on-premise, with standard CPUs.

Data Anonymization and De-identifying Personally Identifiable Information for data processing, sharing and publishing, archiving and reuse

With the emergence of generative AI, the advancement of machine learning techniques, as well as the increasing use of AI tools based on these new technologies in research and science, issues related to the management of sensitive data and compliance with legal regulations (e.g., GDPR, AI Act) have come even more into focus. Hence, anonymization or de-identification of sensitive data or so-called personal identifying information (e.g. in clinical research, as well as personal data in fields like psychology, educational research, empirical social research etc.), has become an important part of the research data cycle. [lxi],[lxii] This might affect not only the stage of data sharing and publishing, but also the archiving process and data reuse at a later stage. At the same time, the further development of AI/ML-based anonymization tools is also advancing. It is important to note that automated anonymization tools do not automatically guarantee regulatory compliant anonymity. Most tools identify and transform sensitive information, but human review remains still essential in many contexts. The recent emergence of locally deployable LLMs has significantly expanded the capabilities of anonymization systems [lxiii]. These technologies can recognize contextual references to individuals and organizations that traditional rule-based approaches may overlook. An example from the medical research field is the LLM-Anonymizer [lxiv],[lxv] that has been developed to anonymize medical documents with local, privacy preserving large language models. An extension to the LLM-Anonymizer is the locally deployable workflow termed LLMAIx [lxvi] where the anonymization tool serves as one building block in there. *Textwash* [lxvii] is also an open-source anonymization tool, specifically for qualitative and social science research. Commercial versions of the tool have been developed recently. *Microsoft Presidio* [lxviii] is a general-purpose, cross-disciplinary open-source framework for detecting, redacting,

masking, and anonymizing sensitive data across text, images, and structured data [lxix]. *SEAL* [lxx] and *RUPTA* [lxxi] while not yet production ready systems, are emerging research frameworks and as such may shape the next generation of privacy-preserving AI technologies.

Reproducibility assistants for collecting, analysing, sharing data, and reusing data

Reproducibility of scientific publications is essential for the scientific community to grow and make use as well as reuse published scientific output. The process of reproducing scholarly output plays a crucial role for researchers and likewise peer reviewers [lxxii]. Both, researchers in general, as well as developers specifically, can be supported by AI assistants to increase reproducibility of computational scientific analyses [lxxiii,lxxiv]. At the same time, the transparent use of AI assisted data analysis and code generation can be automatically documented and reported using distinct AI agents such as Git-Bob workflows/pipelines [lxxv]. On the other hand, reviewers also benefit from AI-supported reconstructions of implementations as well as assisted peer review [lxxvi]. Reproduction of scientific works is often laborious and first automation efforts have been described recently [lxxvii]. Thereby, LLM-based automatic extraction of scholarly articles' hypotheses, experiments and interpretations allow to capture representations and replication tests with the current limitations of multimodal data, visual depictions, and discrepancies between human and computational agents. For example, Bibal et al. [lxxviii] developed an AI reproducibility assistant termed *OpenPub* [lxxix], allowing for workflow reconstruction from thirty hours to one by automatically rebuilding missing steps according to their feasibility study [lxxx]. This efficiency gain illustrates the value of AI-generated proactive guidance tailored to both authors and readers. This among other examples marks a shift in kind, not simply degree: "*AI is becoming an integral layer of scientific infrastructure, a tool extending cognition in the same way the telescope extended vision or the computer extended calculation*" [lxxxi].

5. Discussion and identified action areas

The above exemplified application scenarios of AI/ML approaches are supported by and actively support the sequential steps in managing research-related digital objects. A more comprehensive view on the topic suggests various gaps identified in the given list of examples framed within the context of the research data management lifecycle. A description can be provided by using four elements: description of the situation, detailed gap description, related risk, and related potential. A summary of action areas for exemplary AI/ML applications along the research data cycle is summarized in Figure 3 and detailed in the respective subsections beneath.

In this section, potential fields for further research and development are explicated. Development and application of such AI-based models and tools would thereby involve not only research support and other stakeholders using AI/ML techniques to RDM tasks but also specific sub-fields of computational science or AI research designing and developing underlying models and tools. This could serve as a stimulus for existing research in the field of AI/ML and simultaneously pave the way for new, innovative approaches that can be applied to RDM related tasks and hence improve the research process.

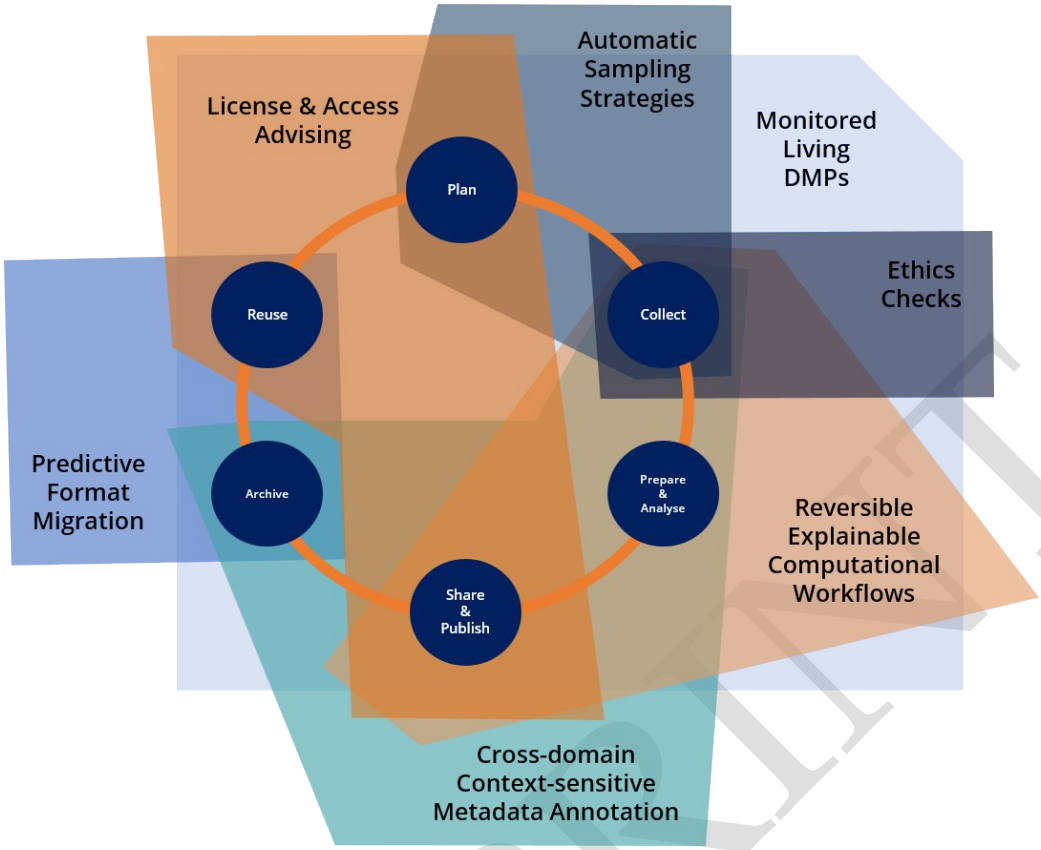

Figure 3: Mapping of action areas for potential AI applications to research data management phases; specified application examples span over multiple steps along the research data cycle.

Cross-domain, context-sensitive metadata annotation

Many subject areas use their own metadata schemas (e.g., biodiversity vs. geosciences vs. medicine). Unfortunately, cross-domain discoverability is poorly developed, which is particularly problematic given that many current research questions can only be addressed through an interdisciplinary approach. Free-text terms, local technical jargon, and non-standardized units hinder searching, network formation, and linking to knowledge graphs. Rich context is also contained in code, notebooks, instrument protocols, and file headers, but this is rarely evaluated for metadata, or manual curation is slow and inconsistent. Many repositories limit themselves to a minimum of metadata fields (simplified Dublin Core/DataCite fields) and offer only limited subject-specific information (units, instruments, methods, variables). Since there are no universally applicable checklists for researchers, different schemas are used, and repositories rarely offer context-specific assistance.

An AI/ML-powered, context-aware metadata annotator capable of deriving the subject area and academic context from files, code, and documentation—and selecting the appropriate schema or profile—could already lead to a massive improvement in the discoverability of research data. In the process, entities, variables, units, locations, and methods would be extracted and normalized to link them to relevant ontologies, thereby generating validated, machine-processable metadata that can be ingested into both generic and discipline-specific repositories. In doing so, domains and the schemas used should be automatically recognized, and compliance with FAIR sharing guidelines or those of publishers/funders should be verified at the same time. To achieve interoperability, variable names must be harmonized according to community standards and assigned to the appropriate measurement types and value ranges. Context-relevant information stored in files, headers, and

notebooks must be used to derive variables, units, instruments, and processes and to establish connections between them.
Inadequate descriptions of datasets result in poor discoverability, leaving them invisible to cross-domain search queries. Semantic discrepancies hinder integration, and missing information on units, methods, and context lead to incorrect reuse and failed reproduction. Data sets that are linguistically limited to a very narrow geographic area severely restrict their global discoverability and reusability.
Comprehensive, standardized, and multilingual metadata improves discoverability in repositories and knowledge graphs. Ontology-linked variables, units, and methods enable cross-domain queries and data integration. Furthermore, machine-processable context (methods, instruments, sampling, PIDs) allows reusers to assess usability and reproduce analyses.

Automatic license and access advising

To be on the safe side, researchers often resort to “All rights reserved” or choose a Creative Commons license without considering the consent requirements, third-party contributions, funder guidelines, or institutional policies—and without understanding how these factors interact. In addition, different types of data (e.g., software, raw data, and documentation) are grouped under a single license, which is incompatible in most cases. Compounding the issue is the fact that terms of use for data sources (APIs, publishers), embedded third-party content, and export control restrictions are difficult to identify and make the reuse of found data impossible or discourage its use.
There is a lack of an AI/ML-powered rights and access advisor that is integrated into the acquisition process—specifically, DMP and publication workflows—and that recommends appropriate licenses and access controls along with explanations, generates machine-readable metadata on rights, and coordinates controlled access as needed. It would be particularly helpful to analyse files and metadata to identify existing licenses and detect cases where no license is present, as well as outdated or custom terms.
Failing to develop a licensing strategy entails various risks that are not limited to legal issues. For instance, it stands to reason that the incorrect use or selection of licenses can easily lead to copyright violations, breaches of terms of use, non-compliance with GDPR, or export control issues. Excessive restrictions on open data, such as choosing a CC-NC/ND license, unnecessarily limit reuse, while insufficient protection when releasing sensitive data can result in financial consequences and reputational damage. The clear origin of data is also crucial; if there is an unclear link between consent/IRB approval and access conditions, this can quickly become a problem, and it becomes nearly impossible to justify the situation to auditors or funders.
Researchers and curators can both benefit from a suitable ML/AI solution, as clear, appropriate licenses provide legal clarity for reuse and automatically align with the recommendations and guidelines of funding agencies and institutions. At the same time, this leads to increased efficiency, as manual reviews are reduced to a minimum and researchers are presented with the correct licenses from the very beginning.

Monitored living DMPs with policy compliance checks

DMPs are still often viewed by researchers as static documents that are created once as part of a grant application, are rarely updated, and have little relevance to their day-to-day work. In fact, the requirements of funding agencies, journals, institutions, and laws sometimes conflict with one another or change without notice, which places an excessive burden on researchers. Although DMPs address these issues, the assignment of PIDs, metadata, deposit in repositories, and long-term archiving

planning only take place during the hectic phase before publication and could easily be integrated into ongoing operations with early planning. Since DMPs are mostly used statically rather than dynamically, important information about the project timeline is missing—information that could serve as evidence of who made which decisions, when, and why.

Dynamic, "living" DMP services that already incorporate an AI assistant could not only be machine-readable but also continuously monitor progress. In doing so, they should take into account both the project scope and institutional guidelines and funding agency requirements, and generate recommendations for action as well as a traceable chain of origin. To do this, these assistants must have access to a curated, versioned library containing the requirements of funding agencies, journals, institutions, and legal regulations, while also identifying potential conflicts or contradictions. The integration of structured fields for data types, sensitivity, repositories, licenses, retention, long-term archiving, and costs would be the absolute minimum and must be continuously updated. The integration of other tools (e.g., GitHub or Nextcloud) that serve as triggers for new events (new dataset, new code release, consent update) must be automatically detected by the AI and prompt a response from the user. This should be followed by an automatic FAIR assessment with suggestions for improvement and early-stage licensing advice for data, code, and metadata. The situation-dependent involvement of human curators or the legal department—in cases where the AI assistant is uncertain—must be a boundary condition of the assistant based on justifiable criteria.

The lack of ongoing assessment of project progress increases the risk of non-compliance with requirements and the associated potential for delayed publication. This results in necessary but costly last-minute corrections. Consequently, there is a risk of violations of consent regulations, invalid licenses, and uncontrolled handling of sensitive data. Finally, missing PIDs, insufficient metadata, and the absence of a retention plan also lead to significantly reduced reusability and contradict the due diligence required by funding agencies and publishers.

The timely deposit of project results in approved repositories with machine-readable permissions and access controls reduces bottlenecks in publication and meets funding agency requirements. Standardized licenses, PIDs, and domain metadata build trust and promote reuse. Increased efficiency through automated checks and recommendations reduces the workload for curators and PIs and simplifies the verification of provenance and compliance status.

Predictive format migration for long-term archiving

There is a shared consensus among all trustworthy repository operators that research data should be preserved at the bit level. In order to be able to use the data in the future, it will be necessary to migrate it to other formats. Currently, there is no forward-looking assessment of when certain datasets will actually need to be migrated, nor of the uniform goals and quality standards by which validation should be conducted. In practice, format changes are only addressed when users report unreadable files, which leads to rushed projects and inconsistent results

A solution would be an AI/ML-powered predictive migration planner that forecasts the obsolescence of formats and proactively plans migration with transparent justification. In doing so, it incorporates both internal policies and external sources into its analysis. These may include the popularity of software and packages, the prevalence and age of repository formats, user usage and access patterns, as well as mentions in citations/registers.

Neglecting migration strategies for research data will inevitably impact ongoing efforts to reuse research data. Although bit integrity remains guaranteed, it will not be possible to extract results or reproduce the research. This, in turn, leads to unnecessary additional costs, as experiments or simulations must be repeated to obtain important data, or regularly results in a "crisis migration"—

which is technically still possible but places an additional, unreasonable burden on staff, infrastructure, and the budget. A lack of planning leads to inconsistencies in migrations and can also be a source of security risks, as compromised formats continue to be used.
AI/ML solutions that actively address these challenges would provide long-term support to both researchers and infrastructure operators. Researchers would be more strongly encouraged to produce results in a curated and reusable manner, as it would be guaranteed that these results will remain usable over a long period of time and that their research can be validated, reproduced, and utilized for future analyses. Additionally, decision-making processes can be better documented, which increases overall audit readiness. Repository operators can incorporate format migrations into their multi-year roadmaps and plan for the corresponding computing and storage requirements. In this context, an institutional or cross-institutional transformation strategy across repositories can lead to standardized practices and reduce duplication of effort.

Explainable and reproducible computational workflows

Simulation and data analysis are fundamental components of modern research, yet there is still room for improvement when it comes to documenting and clearly visualizing dependencies. For example, it is not uncommon for inputs, parameters used, initial values, and software versions to go unrecorded, and for generated code and data to lack persistent identifiers. Runtime environments vary and contain hidden dependencies, such as in notebooks/scripts with unspecified package versions and undocumented requirements for the operating system, GPU, and libraries. Some projects use Conda, others Docker, and still others nothing at all, and requirements on HPC systems are not taken into account. If containers are used, they are often not examined, leading to security and compliance gaps, as licenses for dependencies are unknown, images are unsigned, or, in the worst case, sensitive data ends up in the code. Automation (CI/CD) for rebuilding/testing environments is currently limited and does not include machine-readable metadata for auditors and reusers.
Analysing code, notebooks, and metadata to automatically generate reproducible environments, suggest appropriate containerization strategies for each platform, create workflow definitions, and capture end-to-end provenance with minimal curation effort would be a range of tasks that could be handled by a specialized AI/ML solution. For example, the solution could analyse repositories to identify languages and dependencies in Python, R, Julia, Java, Node, MATLAB, etc., and generate predefined environments (requirements.txt). It could also suggest recommendations for base images and create and run pipelines using sample data.
If software, code, or workflows are not reused because results cannot be reproduced due to a lack of documentation, this leads to a loss of trust, time, and resources. Unknown platform dependencies and vulnerabilities cause pipelines to break on HPC/cloud systems and require custom configurations. This situation is further complicated by staff turnover, resulting in the loss of institutional knowledge. Potential security and compliance issues arising from undisclosed licenses hinder reuse and lead to reduced citations.
Fully specified, portable environments and workflows that can be re-run on laptops, HPC systems, and in the cloud—with clear traceability and PIDs—greatly increase the visibility and dissemination of one's own research, which has a direct positive impact on the reputation of researchers and institutions. Reviewers at academic journals and peers can validate results relatively quickly and cost-effectively and base future projects on existing work(forks) rather than relying on new developments. Last but not least, cross-institutional interoperability is promoted as standard workflow languages become widely adopted, and images that are automatically checked for vulnerabilities and signed enable more secure operation of the infrastructure.

Automated ethics and bias checks during data collection

Especially research in the humanities and social sciences relies heavily on the collection and analysis of personal data, which is subject to special data protection and can have serious consequences for all parties involved in the event of a data breach. Currently, in most cases, approvals for research questions are obtained on a one-time basis from the relevant committees, and the scope of consent is not machine-readable. There is also no ongoing review of compliance, especially as instruments evolve or new areas are covered. Considerations regarding biases and ethical aspects are not recorded as structured metadata (data sheets/data cards), which hinders reuse and peer review. This also means that hidden sampling and measurement biases leading to under-/over-representation of groups, as well as suggestive or ambiguous question wording, remain undetected until the analysis stage. Data protection and the scope of consent are generally taken very seriously, nevertheless, overstepping the scope of consent (or local laws) as well as restrictions (e.g., data sovereignty of indigenous peoples, export controls) can occur unintentionally and are not encoded in the metadata.
An AI/ML-powered solution that is directly integrated into survey platforms, data loggers, and annotation tools—and that performs real-time checks, recommends corrective actions, and records machine-readable metadata and provenance data regarding ethics and bias—could offer substantial improvements. The priority must be placed on safeguards for informed consent and data protection, and it is essential to promptly verify whether the incoming sample is consistent with the target parameters (from the DMP or external statistics). Any resulting warnings regarding gaps in representativeness, recommendations for adaptive recruitment or oversampling, or changes to the survey instrument must be proactively communicated to the researchers. Machine-processable metadata on ethics and bias must be automatically generated and stored in a structured format. When handling personal data, human-in-the-loop review must be conducted with decision logs and justifications, and the AI/ML solution must be clearly labelled as a “support tool” that must under no circumstances act entirely on its own.
The lack of documentation regarding ethical considerations, biases, and similar issues prevents reviewers and reusers from identifying systematic underrepresentation and the use of biased instruments, which inevitably leads to inaccurate or unreliable results for certain subgroups. This results in poor reusability, as reproducibility is not guaranteed and costly data collection must be repeated. Additional violations of consent regulations, insufficient anonymization, and breaches of sovereign rights and export control regulations lead to legal consequences and a loss of trust.
The use of AI/ML-generated datasets—which contain machine-processable consent boundaries, information on biases, and provenance data—enables reusers to assess suitability and reproduce results responsibly. Proactive quality and fairness analysis (real-time guidance) towards trustworthiness reduces bias at the source, improves validity and generalizability, and protects participants at the same time.

In summary, this analysis on exemplary application scenarios serves as a starting point for future developments. Key insights from the community and expert input during the FAIR and AI symposisum influenced the analysis and are aligned with the above indicated action areas: (1) There is a disparity between FAIR metadata and data quality. (2) Sensitive data requires integrated governance and infrastructure. (3) Data stewards are becoming critical intermediaries. (4) AI governance and literacy are emerging RDM competencies. (5) Reproducibility and provenance remain central despite AI advances.

It should be emphasized that the Shared RDM Services and Infrastructure project concludes/concluded in July 2026, which limits the possibility of quickly transforming the identified AI/ML opportunities into

concrete applications within the current project's timeframe. This plan is supported by the fact that almost all the project participants have expressed a strong wish to continue the initiative beyond the formal end of the project, transforming it into a sustainable community effort. In this sense, the outcomes documented in this paper are not only tied to the project but also serve as a long-term orientation framework for Austria's RDM landscape and its connection to European and international initiatives. While from an institutional perspective the project operates on a national level, the present conceptualization and gap analysis may fit for other countries as well, because the cycle-stages are being regarded from a research perspective without geographical boundaries.
Besides, the relevant stakeholders and their roles need to be addressed as they - like for example publishing business, infrastructure providers or funding agencies - also are affected by AI. Moreover, not only options, but also risks and potential barriers to future AI/ML support in RDM must be included in developing a strategic roadmap.

6. Conclusion and Outlook

AI/ML approaches are currently often described to increase efficiency in data-driven science which can support scientific discovery and decision making. In regard to RDM several methods can be applied particularly to support data collection, preparation and analysis, such as AI/ML-based transcripts, but also other applications along the research data cycle, such as rendering sensitive data publications possible through AI/ML-assisted anonymization and synthetic data generation techniques. Notwithstanding, efficiency could come at the expense of quality. Therefore, we highlight some examples for AI/ML approaches that potentially increase the quality in data-driven research, such as increased reproducibility of computational workflows, and automated version and provenance documentation for transparency. Some future developments are outlined for example facilitating the work with personal data through compliance checkers, or fostering interdisciplinary reuse through cross-domain, context-sensitive metadata annotation. A further step could involve building on agent-based AI workflows throughout the data lifecycle, which could facilitate data-driven research processes, provided that, inter alia, the (already well-established) individual components are orchestrated properly. Although this topic goes beyond the scope of this article, it is worth looking in this direction for future research, as there are a few promising solutions tailored to specific disciplines, some of which are still at an early stage of development.

The outlined exemplary AI/ML applications and use cases as well as the deducted action areas for potential applications highlight AI as an enabling layer for FAIR and Open Science, while stressing that sustainable service design, governance, and responsible use are essential for its success. Conversely, FAIR data sharing is pivotal to developing and implementing AI methods, and both viewpoints will be necessary to maximize the yield by reuse of digital objects in research.

7. Acknowledgements

We thank all participants and speakers of the FAIR & AI symposium [vii] for valuable discussion rounds on the interplay between AI/ML and FAIR. We are also grateful for all input from topical discussions in the AI-RDM working group on AI and RDM from the project Shared RDM Services & Infrastructure [lxxxii] and its related funding from the Federal Ministry Women, Science and Research Republic of Austria as part of the (Digital) Research Infrastructure Call.

8. Literature

[i] Azeroual, Otmane, and Joachim Schöpfel. "New Developments in Research Data Management-The Potential of AI." Encyclopedia of Libraries, Librarianship, and Information Science (2025): 206-211. https://doi.org/10.1016/B978-0-323-95689-5.00253-4

[ii] https://forschungsdaten.at/en/sharedrdm/ cited in FAIR Data Austria – Paving the Way for Enhanced Research Data Management and Collaboration. Alexander Bardel, Ilire Hasani-Mavriqi. 49-64Bd. 18 Nr. Sonderheft Forschung (2023): Digitalisierung in der Forschung – Projekte österreichischer Hochschulen 2020–2024

[iii] Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018. (2016). https://doi.org/10.1038/sdata.2016.18

[iv] Reichmann, Stefan, et al. "Between administration and research: Understanding data management practices in an institutional context." Journal of the Association for Information Science and Technology 72.11 (2021): 1415-1431. https://doi.org/10.1002/asi.24492

[v] Amanda Calatrava, Hernán Asorey, Jan Astalos, Alberto Azevedo, Francesco Benincasa, Ignacio Blanquer, Martin Bobak, Francisco Brasileiro, Laia Codó, Laura del Cano, Borja Esteban, Meritxell Ferret, Josef Handl, Tobias Kerzenmacher, Valentin Kozlov, Aleš Křenek, Ricardo Martins, Manuel Pavesio, Antonio Juan Rubio-Montero, Juan Sánchez-Ferrero. A survey of the European Open Science Cloud services for expanding the capacity and capabilities of multidisciplinary scientific applications. Computer Science Review, Volume 49, (2023), 100571, https://doi.org/10.1016/j.cosrev.2023.100571

[vi] And Cox, Andrew Martin, and Winnie Wan Ting Tam. "A critical analysis of lifecycle models of the research process and research data management." Aslib Journal of Information Management 70.2 (2018): 142-157. https://doi.org/10.1108/AJIM-11-2017-0251

[vii] Söser, B. (2025, Dezember 2). FAIR & AI Symposium @ TU Graz. Graz University of Technology. https://doi.org/10.3217/mp2jw-6af34

[viii] Lu L, Zhong Y, Luo S, et al. Dilemmas and prospects of artificial intelligence technology in the data management of medical informatization in China: A new perspective on SPRAY-type AI applications. Health Informatics Journal. 2024;30(2). https://doi.org/10.1177/14604582241262961

[ix] Elghaish, F., Chauhan, J. K., Matarneh, S., Rahimian, F. P., & Hosseini, M. R. (2022). Artificial intelligencebased voice assistant for BIM data management. Automation in Construction, 140, Article 104320. https://doi.org/10.1016/j.autcon.2022.104320

[x] Hachimi CE, Belaqziz S, Khabba S, Sebbar B, Dhiba D, Chehbouni A. Smart Weather Data Management Based on Artificial Intelligence and Big Data Analytics for Precision Agriculture. Agriculture. 2023; 13(1):95. https://doi.org/10.3390/agriculture13010095

[xi] Cedric Renggli, Frances Ann Hubis, Bojan Karlaš, Kevin Schawinski, Wentao Wu, and Ce Zhang. 2019. Ease.ml/ci and Ease.ml/meter in action: towards data management for statistical generalization. Proc. VLDB Endow. 12, 12 (August 2019), 1962–1965. https://doi.org/10.14778/3352063.3352110

[xii] Campos MLM, Silva E, Cerceau R, Cruz SMS, Silva FAB, Gouveia FC, Jardim R, Kotowski N, Lopes GR and Dávila AMR (2021) Towards Machine-Readable (Meta) Data and the FAIR Value for Artificial Intelligence Exploration of COVID-19 and Cancer Research Data. Front. Big Data 4:656553. https://10.3389/fdata.2021.656553

[xiii] Lu L, Zhong Y, Luo S, et al. Dilemmas and prospects of artificial intelligence technology in the data management of medical informatization in China: A new perspective on SPRAY-type AI applications. Health Informatics Journal. 2024;30(2). https://10.1177/14604582241262961

[xiv] Serey, J., Quezada, L., Alfaro, M., Fuertes, G., Vargas, M., Ternero, R., Sabattin, J., Duran, C., & Gutierrez, S. (2021). Artificial Intelligence Methodologies for Data Management. Symmetry, 13(11), 2040. https://doi.org/10.3390/sym13112040

[xv] Jankovic SD, Curovic DM. Strategic Integration of Artificial Intelligence for Sustainable Businesses: Implications for Data Management and Human User Engagement in the Digital Era. Sustainability. 2023; 15(21):15208. https://doi.org/10.3390/su152115208

[xvi] Balayn, A., Lofi, C. & Houben, GJ. Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. The VLDB Journal 30, 739–768 (2021). https://doi.org/10.1007/s00778-021-00671-8

[xvii] Urban R, Haluzová S, Strunga M, Surovková J, Lifková M, Tomášik J, Thurzo A. AI-Assisted CBCT Data Management in Modern Dental Practice: Benefits, Limitations and Innovations. Electronics. 2023; 12(7):1710. https://doi.org/10.3390/electronics12071710

[xviii] Chengliang Chai, Jiayi Wang, Yuyu Luo, Zeping Niu, and Guoliang Li. 2023. Data Management for Machine Learning: A Survey. IEEE Trans. on Knowl. and Data Eng. 35, 5 (May 2023), 4646–4667. https://doi.org/10.1109/TKDE.2022.3148237

[xix] Vadim Strogonov, Jaroslav Pollert; Artificial intelligence-enhanced web application approach to data management in the WIDER UPTAKE project. Journal of Hydroinformatics 1 April 2025; 27 (4): 686–699. doi: https://doi.org/10.2166/hydro.2025.248

[xx] Johnson, A., Lindquist, T., Murray, M., Ranganath, A., Freeborn, L., Knuth, S., Schnell, B., Klopsch, S. O., Sabeti, V., Wittenberg, J., Lindholm, D., Regan, K., Elsborg, D., & Viggio, A. (2025). University of Colorado Boulder - Machine Actionable Plans (MAP) Pilot Project Report. University of Colorado Boulder. https://doi.org/10.25810/TKNV-JT07

[xxi] Michener, W. K. (2015). Ten simple rules for creating a good data management plan. PLoS computational biology, 11(10), e1004525. https://doi.org/10.1371/journal.pcbi.1004525

[xxii] Miksa, T., Oblasser, S., and Rauber, A.. 2021. Automating Research Data Management Using Machine-Actionable Data Management Plans. ACM Trans. Manage. Inf. Syst. 13, 2, Article 18 (June 2022), 22 pages. https://doi.org/10.1145/3490396

[xxiii] Dong, X., Qiu, W. A method for managing scientific research project resource conflicts and predicting risks using BP neural networks. Sci Rep 14, 9238 (2024). https://doi.org/10.1038/s41598-024-59911-w

[xxiv] Raju, C. M., Elpa, D. P., & Urban, P. L. (2024). Automation and Computerization of (Bio)sensing Systems. ACS sensors, 9(3), 1033-1048. https://doi.org/10.1021/acssensors.3c01887

[xxv] Jain, N. (2021). Survey Versus Interviews: Comparing Data Collection Tools for Exploratory Research. The Qualitative Report, 26(2), 541-554. DOI: https://doi.org/10.46743/2160-3715/2021.4492

[xxvi] Bharech, S., Yang, Y., Selzer, M. et al. ML-extendable framework for multiphysics-multiscale simulation workflow and data management using Kadi4Mat. Sci Data 12, 962 (2025). https://doi.org/10.1038/s41597-025-05027-3

[xxvii] Strogonov, V., & Pollert, J. (2025). Artificial intelligence-enhanced web application approach to data management in the WIDER UPTAKE project. Journal of Hydroinformatics, 27(4), 686-699. https://doir.org/10.2166/hydro.2025.248

[xxviii] Musik S, Sasin-Kurowska J, Panczyk M. Bridging the Past and Future of Clinical Data Management: The Transformative Impact of Artificial Intelligence. Open Access Journal of Clinical Trials. 2025;17:15-33. https://doi.org/10.2147/OAJCT.S509921

[xxix] Wen, N., Zhou, Y., Wang, Y., Zheng, Y., Fan, Y., Liu, Y., Wang, Y., & Li, M. (2025). Dynamic Sensor-Based Data Management Optimization Strategy of Edge Artificial Intelligence Model for Intelligent Transportation System. Sensors, 25(7), 2089. https://doi.org/10.3390/s25072089

[xxx] Tamm, H.C., Nikiforova, A. (2026). From Data Quality for AI to AI for Data Quality: A Systematic Review of Tools for AI-Augmented Data Quality Management in Data Warehouses. In: Perspectives in Business Informatics Research. BIR 2025. Lecture Notes in Business Information Processing, vol 562. Springer, Cham.

https://doi.org/10.1007/978-3-032-04375-7_3

[xxxi] Goble, C., Cohen-Boulakia, S., Soiland-Reyes, S., Garijo, D., Gil, Y., Crusoe, M.R., Peters, K., Schober, D.; FAIR Computational Workflows. Data Intelligence 2020; 2 (1-2): 108–121. https://doi.org/10.1162/dint_a_00033

[xxxii] Jean-Quartier C, Stryeck S, Thien A, Vrella, B., Kleinschuster, J., Spreitzer, E., Wali, M., Mueller, H., Holzinger, A., Jeanquartier, F. Unlocking biomedical data sharing: A structured approach with digital twins and artificial intelligence (AI) for open health sciences. DIGITAL HEALTH. 2024;10. https://doi.org/10.1177/20552076241271769

[xxxiii] Hachimi, C. E., Belaqziz, S., Khabba, S., Sebbar, B., Dhiba, D., & Chehbouni, A. (2023). Smart Weather Data Management Based on Artificial Intelligence and Big Data Analytics for Precision Agriculture. Agriculture, 13(1), 95. https://doi.org/10.3390/agriculture13010095

[xxxiv] Celik, B., Sandt, R., dos Santos, L. C. P., & Spatschek, R. (2022). Prediction of Battery Cycle Life Using Early-Cycle Data, Machine Learning and Data Management. Batteries, 8(12), 266. https://doi.org/10.3390/batteries8120266

[xxxv] Huang R and Tao S (2025) A human-centered automated machine learning agent with large language models for multimodal data management and analysis. Front. Artif. Intell. 8:1680845. doi: 10.3389/frai.2025.1680845

[xxxvi] Bibal, A., Minton, S. N., Khider, D., & Gil, Y. (2025). AI Copilots for Reproducibility in Science: A Case Study. AAAI Workshop on Reproducible Artificial Intelligence (RAI2026), arXiv preprint arXiv:2506.20130. https://doi.org/10.48550/arXiv.2506.20130

[xxxvii] Figueiredo AS (2017) Data Sharing: Convert Challenges into Opportunities. Front. Public Health 5:327. https://doi.org/10.3389/fpubh.2017.00327

[xxxviii] Xu, W., Wu, X., He, R., & Zhou, M. (2023, May). Licenserec: Knowledge based open source license recommendation for oss projects. In 2023 IEEE/ACM 45th International Conference on Software Engineering: Companion Proceedings (pp. 180-183). IEEE. https://doi.org/10.1109/ICSE-Companion58688.2023.00050

[xxxix] Obande, B. O., Emmanuel, H., & Tsenongu, C. T. (2024). Strategies and Technologies Employed by Libraries and Archives to ensure Longevity and Accessibility to Digital Content. Direct Research Journal of Engineering and Information Technology, 12(2), 65-70. https://journals.directresearchpublisher.org/index.php/drjeit/article/view/323

[xl] Kirupa Shankar, K.M., Santhi, V. Integrating machine learning and encryption for effective data management in blood bank supply chains. J Cloud Comp 14, 56 (2025). https://doi.org/10.1186/s13677-025-00779-0

[xli] Olawale, O. P. and Ebadinezhad, S., Cybersecurity Anomaly Detection: AI and Ethereum Blockchain for a Secure and Tamperproof IoHT Data Management, in IEEE Access, 12, (pp. 131605-131620), 2024, https://doi.org/10.1109/ACCESS.2024.3460428.

[xlii] Joshi, A., Mahapatra, R.P., and Devarajan, G.G., An Improved Mechanism to Maintain Data Integrity and Anomaly Detection in Cloud Storage, Journal of Machine and Computing, 6(1), pp. 296-311, 2026. 10.53759/7669/jmc202606022

[xliii] Reyes-Lillo, D., Rovira, C., & Morales-Vargas, A. (2025). Factors for enhancing visibility in digital repositories: Metadata quality, interoperability standards, persistent identifiers, and SEO-GEO optimization. In J. Guallar, M. Vállez, & A. Ventura-Cisquella (Coords). Digital communication. Trends and good practices (pp. 119-133). Ediciones Profesionales de la Información. https://doi.org/10.3145/cuvicom.09.eng

[xliv] Yazdi, M.A., Politze, M., Heinrichs, B. (2023). Research Data Reusability with Content-Based Recommender System. In: Conte, D., Fred, A., Gusikhin, O., Sansone, C. (eds) Deep Learning Theory and Applications. DeLTA 2023. Communications in Computer and Information Science, vol 1875. Springer, Cham. https://doi.org/10.1007/978-3-031-39059-3_10

[xlv] V. Pereira, M. P. Basilio, and C. H. T. Santos, “PyBibX – a Python library for bibliometric and scientometric analysis powered with artificial intelligence tools,” Data Technologies and Applications, vol. 59, no. 2, pp.302–337, 2025. https://doi.org/10.1108/DTA-08-2023-0461

[xlvi] Dröge, K. (2025): Datenaufbereitung durch Transkription - In: Gras, Juliana [Hrsg.]; Schieferdecker, Ralf [Hrsg.]: Einführung in Qualitative Sozialforschung. Grundlagen für Studierende pädagogischer Studiengänge. Bad Heilbrunn: Verlag Julius Klinkhardt 2025, S. 224-234 - DOI: https://doi.org/10.35468/6188-15

[xlvii] Dröge, K. (2025). noScribe. AI-powered Audio Transcription (Version 0.7) [Computer software]. https://github.com/kaixxx/noScribe

[xlviii] Haberl, A., Fleiß, J., Kowald, D., & Thalmann, S. (2024). Take the aTrain. Introducing an interface for the Accessible Transcription of Interviews. Journal of Behavioral and Experimental Finance, 41, 100891. https://doi.org/10.1016/j.jbef.2024.100891

[xlix] H. Bredin et al., "Pyannote.Audio: Neural Building Blocks for Speaker Diarization," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7124-7128, https://doi.org/10.1109/ICASSP40776.2020.9052974

[l] A. Radford, J. Wook Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," In Proceedings of the 40th International Conference on Machine Learning (ICML'23), Vol. 202. JMLR.org, Article 1182, 28492-28518. https://dl.acm.org/doi/10.5555/3618408.3619590

[li] Hart, T., & Achterman, P. (2017). Qualitative analysis software (ATLAS. ti/Ethnograph/MAXQDA/NVivo). The international encyclopedia of communication research methods, 1. https://doi.org/10.1002/9781118901731.iecrm0194

[lii] Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, N., Weller, A. (2022): Synthetic Data--what, why and how? arXiv preprint arXiv:2205.03257. https://doi.org/10.48550/arXiv.2205.03257

[liii] Quintana, D. S. (2020): A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. eLife, 9, e53275. https://doi.org/10.7554/eLife.53275

[liv] Goyal, M., & Mahmoud, Q. H. (2024). A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. Electronics, 13(17), 3509. https://doi.org/10.3390/electronics13173509

[lv] N. Patki, R. Wedge and K. Veeramachaneni, "The Synthetic Data Vault," 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, 2016, pp. 399-410, https://doi.org/liv10.1109/DSAA.2016.49

[lvi] Desai P.S. (2013): Marketing Science Replication and Disclosure Policy. Marketing Science 32(1), pp. 1–3. http://dx.doi.org/10.1287/mksc.1120.0761

[lvii] p.35 in Standards for Reporting on Empirical Social Science Research in AERA Publications: American Educational Research Association: American Educational Research Association. (2006). Educational Researcher, 35(6), 33-40.

[lviii] Avraam, D., Jones, E., & Burton, P. (2022). A deterministic approach for protecting privacy in sensitive personal data. BMC medical informatics and decision making, 22(1), 24. https://doi.org/10.1186/s12911-022-01754-4

[lix] Gonzales, A., Guruswamy, G., & Smith, S. R. (2023). Synthetic data in health care: A narrative review. PLOS digital health, 2(1), e0000082. https://doi.org/10.1371/journal.pdig.0000082

[lx] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, Scott McLachlan, Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record, Journal of the American Medical Informatics Association, Volume 25, Issue 3, March 2018, Pages 230–238, https://doi.org/10.1093/jamia/ocx079

[lxi] Neamatullah I, Douglass M, Lehman LH, Reisner A, Villarroel M, Long WJ, Szolovits P, Moody GB, Mark RG, Clifford GD. Automated De-Identification of Free-Text Medical Records. BMC Medical Informatics and Decision Making, 2008, 8:32. https://doi.org/10.1186/1472-6947-8-32

[lxii] Trienes J., Trieschnigg D., Seifert C., Hiemstra D. (2020). Comparing Rule-based, Feature-based and Deep

Neural Methods for De-identification of Dutch Medical Records. In: Proceedings of the 1st ACM WSDM Health Search and Data Mining Workshop (HSDM), 2020. https://doi.org/10.48550/arXiv.2001.05714

[lxiii] Manzanares-Salor B., & Sánchez, D. (2026). A comparative analysis, enhancement and evaluation of text anonymization with pre-trained Large Language Models. Expert Systems with Applications, 297, 129474. https://doi.org/10.1016/j.eswa.2025.129474

[lxiv] Wiest, I. C., Leßmann, M.-E., Wolf, F., Ferber, D., Treeck, M. V., Zhu, J., Ebert, M. P., Westphalen, C. B., Wermke, M., Kather, J. N. (2025). Deidentifying Medical Documents with Local, Privacy-Preserving Large Language Models: The LLM-Anonymizer. NEJM AI, 2(4), AIdbp2400537. https://doi.org/10.1056/AIdbp2400537

[lxv] Wiest, I. C., Wolf, F., Leßmann, M.-E., van Treeck, M., Ferber, D., Zhu, J., Boehme, H., Bressem, K. K., Ulrich, H., Ebert, M. P., Kather, J. N. (2025). A software pipeline for medical information extraction with large language models, open source and suitable for oncology. npj Precision Oncology, 9(1), 313. https://doi.org/10.1038/s41698-025-01103-4

[lxvi] Wiest, I.C., Wolf, F., Leßmann, ME. et al. A software pipeline for medical information extraction with large language models, open source and suitable for oncology. npj Precis. Onc. 9, 313 (2025). https://doi.org/10.1038/s41698-025-01103-4

[lxvii] Kleinberg B., Davies T., Mozes M. (2022): Textwash -- automated open-source text anonymization. https://doi.org/10.48550/arXiv.2208.13081

[lxviii] Atri, S. (2025). Enterprise-Scale PII De-Identification with Microsoft Presidio Anonymizer: Architecture, Use Cases, and Best Practices. International Journal of AI, BigData, Computational and Management Studies, 6(4), 176-180. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V6I4P120

[lxix] Asimopoulos D., Siniosoglou I., Argyriou V., Karamitsou T., Eleftherios Fountoukidis E., Goudos S.K., Moscholios I.D., Psannis K.E., Sarigiannidis P. (2024). Benchmarking Advanced Text Anonymisation Methods: A Comparative Study on Novel and Traditional Approaches. https://arxiv.org/abs/2404.14465v1

[lxx] Kim, K., Jeon, H., Shin, J. Self-Refining Language Model Anonymizers via Adversarial Distillation. 2025. 39th Conference on Neural Information Processing Systems (NeurIPS 2025). https://doi.org/10.48550/arXiv.2506.01420

[lxxi] Yang, T., Zhu, X., Gurevych, I. Robust Utility-Preserving Text Anonymization Based on Large Language Models. (2024), arXiv preprint arXiv:2407.11770, https://doi.org/10.48550/arXiv.2407.11770

[lxxii] Nüst, D., & Eglen, S. J. (2021). CODECHECK: an Open Science initiative for the independent execution of computations underlying research articles during peer review to improve reproducibility. F1000Research, 10, 253. https://doi.org/10.12688/f1000research.51738.2

[lxxiii] Shah, S. M. H., Hopfgartner, F., & Bleier, A. (2026). Automating computational reproducibility in social science: Comparing prompt-based and agent-based approaches. arXiv preprint arXiv:2602.08561. https://doi.org/10.48550/arXiv.2602.08561

[lxxiv] Bahaidarah, L., Hung, E., Oliveira, A. F. D. M., Penumaka, J., Rosario, L., & Trisovic, A. (2022). Toward reusable science with readable code and reproducibility. In 2022 IEEE 18th International Conference on e-Science (e-Science) (pp. 437-439). IEEE. https://doi.org/10.1109/eScience55777.2022.00079

[lxxv] Haase, R. Towards transparency and knowledge exchange in AI-assisted data analysis code generation. Nature Computational Science 5, 271–272 (2025). https://doi.org/10.1038/s43588-025-00781-1

[lxxvi] Riehl, K., Marin, A. L., Zacharof, N., Wu, F., Langer, P., Jakob, R., Kouvelas, A., Fontaras, G., Makridis, M. A. (2026). ARA: Agentic Reproducibility Assessment For Scalable Support Of Scientific Peer-Review. arXiv:2605.02651. https://doi.org/10.48550/arXiv.2605.02651

[lxxvii] Snelleman, T., Lawrence, P. L., Hoos, H. H., & Gundersen, O. E. (2025). Automated Reproducibility Has a Problem Statement Problem. arXiv preprint arXiv:2601.04226. https://doi.org/10.48550/arXiv.2601.04226

[lxxviii] Bibal, A., Minton, S., Khider, D., & Gil, Y. (2025). AI Copilots for Reproducibility in Science: A Case Study. AAAI Workshop on Reproducible Artificial Intelligence (RAI2026), Singapore.

https://doi.org/10.48550/arXiv.2506.20130

[lxxix] openpub. (2025, September 22). reproducibility_copilot. Bitbucket. Retrieved June 9, 2026, from https://bitbucket.org/inferlink/reproducibility_copilot

[lxxx] Gundersen, O. E.; Cappelen, O.; Mølnå, M.; and Nilsen, N. G. 2025. The unreasonable effectiveness of open science in AI: A replication study. Proceedings of the AAAI Conference on Artificial Intelligence, 39(25): 26211–26219. https://doi.org/10.1609/aaai.v39i25.34818

[lxxxi] Vardeman II, C., & Brower, D. (2025). Tech Notes: AI and the Data Lifecycle in NSF Major Facilities: From Experimental Tool to Embedded Intelligence. Virtual Workshop – AI Meets CI: Intelligent Infrastructure for Major & Midscale Facilities. January 12-14, 2026. https://doi.org/10.5281/zenodo.17873016

[lxxxii] Söser, B.; Bardel, A., Hasani-Mavriqi, I., Guseva, M., Stork, C., Köbler, J., Sanchez-Solis, B., Hikl, A., Jambura-Türtscher, J., Soran, N. (2026) Shared RDM – Framework Conditions for Shared RDM Services in Austria. https://doi.org/10.3217/5x82z-1e385