2026 International Digital Curation Conference

Accepted Papers

'&%$£ In = &%$£ Out: How Controlled Vocabularies and Metadata Standards Are Fundamental for Developing Open Research Indicators

In 2024 the UK Reproducibility Network (UKRN) initiated a set of pilots involving institutional members and solution providers to establish good practice in institutional monitoring of Open Research through the creation of robust indicators. The Open Research Indicators Pilot was sector led, with institutions and solution providers working together to develop, test, and evaluate prototype machine learning solutions with valid, reliable, and ethical indicators for measuring Open Research. The University of Bristol was the lead for the 'Openness of Data' pilot and assessed providers' data to ascertain the usefulness of machine learning for this purpose.

A Citation Analysis of Government of Canada Open Data in Academic Literature: Leveraging AI for Open Data Archive Impact Assessment

This presentation introduces the first comprehensive analysis of how Government of Canada open data is cited in academic literature, addressing a critical challenge digital curators face: how to demonstrate the impact and value of open data collections.

Using a fine-tuned BERT language model trained on over 3,000 manually verified citation examples, this study overcame the problem of inconsistent data citation standards. This study leveraged AI to identify 3,953 citing articles with 91% accuracy, significantly outperforming traditional keyword-matching methods at 73% accuracy.

The study reveals key usage patterns across disciplines, identifying environmental science, agriculture, and immigration studies as primary users of Canadian government data. The study's findings provide digital curators with evidence-based insights for strategic collection development and resource allocation decisions, while the open-source methodology offers the community immediately deployable tools for impact assessment.

In an era of budget cuts where archives must continually justify their value, this study demonstrates how AI can enhance traditional bibliometric approaches to provide more comprehensive and accurate measures of collection impact, directly addressing contemporary challenges in digital curation."

Acting in the Best Interest of the Other: An Ethics of Care in Digital Curation

This study explores how digital repositories approach qualitative research data curation through an ethics of care lens, particularly when handling data containing identifiable participants. Through 44 semi-structured interviews with educational researchers and teacher-educators who produce and reuse video records of practice (VROP), the research examines perceptions of care in repository practices and the relationships between repositories and their designated communities.

Our findings indicate that (1) data producers and reusers in education view repositories as sites of care, (2) they view data curation as a form of care, and (3) they expect repositories to act in the best interest of the participants represented in research data, thereby enacting an ethics of care.

Interviewees emphasized that repositories must extend beyond technical compliance to embrace ethical commitments that preserve participant dignity throughout the data lifecycle. They sought repositories whose values aligned with their own ethics of care, particularly regarding protection of vulnerable populations. The study identifies care as both a relational process that develops over time and a framework that should inform repository policies from data selection through access decisions.

These findings extend current understanding of designated communities beyond consumers of data to include groups whose ethical frameworks should inform repository practices, with implications for qualitative data repositories containing data with identifiable participants.

Auditing the Human BioMolecular Atlas Program (HuBMAP) Human Reference Atlas (HRA): An Evaluation of Core Digital Objects

Data auditing has become increasingly critical for large-scale biomedical repositories as they serve diverse research communities while maintaining scientific rigor and compliance with established standards. The Human BioMolecular Atlas Program (HuBMAP) aims to map the human body at single-cell resolution through curated spatial and molecular data. The Human Reference Atlas (HRA), a central output of HuBMAP, includes datasets such as Anatomical Structures, Cell Types and Biomarkers (ASCT+B) tables, 2D Functional Tissue Unit (FTU) illustrations, 3D reference organ models, and Organ Mapping Antibody Panels (OMAPs).

This study reports the first comprehensive, third-party audit of the HRA, conducted from March to July 2024 to assess data quality, internal consistency, and adherence to Standard Operating Procedures (SOPs). The audit methodology combined systematic evaluation of metadata completeness and correctness with visual inspection protocols designed to assess user experience and functional utility across different digital object types. Using a combination of visual inspections, file metadata analyses, and spreadsheet comparisons across 34 ASCT+B tables, 22 2D FTU illustrations, 70 3D reference models, and 21 OMAP datasets, the audit demonstrated overwhelmingly positive results, with compliance rates of 94-100% across most evaluation criteria. Findings indicate that HuBMAP maintains robust curation standards, with structural issues present in fewer than 10% of ASCT+B tables. This audit provides a replicable model for future quality assurance activities in large-scale biomedical data infrastructures and highlights the importance of continuous audit processes for ensuring data integrity, transparency, and usability in contemporary digital curation contexts.

Curation and Reproducibility in an Artificial Intelligence World: Challenges and Solutions for Scientific Research

Much has been written about artificial intelligence, with astonishingly rapid progress in computer sciences. In the social sciences, concerns have been raised that artificial intelligence may impact the actual production of scientific output. Most of the discussion has been about the writing of texts, and estimates suggests that the number of articles created and possibly submitted with the help of AI systems is non-trivial. Less interest has been devoted to the use of AI as part of the legitimate scientific production process. Yet use of AI methods in legitimate scientific work is also increasing. With the earlier “replication crisis” still in mind, the question for curators is whether and how to curate AI-supported research tools, input data, and outputs. This article will approach the topic from the perspective of a “data editor”, responsible for verifying reproducibility and supporting curation of research compendia for a prominent learned society in economics.

Dealing with Unprecedented Scale and Complexity: Lessons from Archiving HS2 Digital Archaeological Data

Construction of High Speed 2, the UK's largest linear infrastructure project, brought the need to undertake the most extensive archaeological programme the country had ever seen. Huge challenges to how its archaeological outcome would be recorded and preserved derived from the monumental geographic and temporal scale of the project. As specific government-mandated requirements were imposed on the overall scheme, this created a complex regulatory environment for the large number of parties involved. Discrepancies between companies with distinct methodologies and individual reporting standards posed a threat for the consistency of the records and therefore their preservation, access and reuse.

The Archaeology Data Service (ADS) was tasked with setting standards for data deposition, digital preservation, and access to all archaeological data created by HS2. With over 31 terabytes of data in an array of formats, the project shone a light on the limitations of existing frameworks when managing large-scale, heterogeneous datasets while presenting a significant opportunity for innovation in archiving practice, infrastructure, and rationale. The scale and complexity of HS2 introduced both technical and epistemological risks to how we provide long-term digital preservation to the data entrusted in our care while ensuring it remains findable, accessible, interoperable, and re-usable.

This paper analyses and reflects upon how the ADS has been transformed by the demands of HS2, not only in its technical capacity but in its understanding of the infrastructural, organisational, and ethical dimensions of large-scale digital curation. Challenges offered a proving ground in which future approaches to archaeological data management and archiving could be tested. This led to new tools, adaptable procedures, better workflows, and more nuanced perspectives on the value of curated data. Our capacity to ensure data integrity and accessibility in perpetuity has expanded, demonstrating the project’s long-term infrastructural benefit to the sector on a wider level."

Digitally Preserving e-Theses: Challenges and Opportunities

The digital preservation of e-Theses presents several challenges. This is partly because the status of doctoral theses within the scholarly communication ecosystem is unclear. The shift from physical submission of theses to digital submission is also relatively new, spurred on by the Covid-19 pandemic. This paper is focused on policies, practices and workflows as they relate to e-Thesis submission and digital preservation. The research underpinning the paper is based on interviews, surveys and focus groups with doctoral colleges, institutional repositories and doctoral researchers within UK universities.

Introducing FJORD: Orchestrating Research Data Management through Enhanced FAIRness

Research data management needs better, integrated approaches. Data Management Plans (DMPs) are key, but fragmented systems limit their FAIR principles potential. This paper introduces FJORD, a

framework evolving DMPs to integrate and orchestrate research ecosystems at an institutional level. FJORD enhances DMPs into supplementary tools, leveraging a project proposal platform to manage, monitor, and track intellectual assets via knowledge graphs, aiding researchers across university systems. The framework exports machine-actionable DMPs (maDMPs) compatible with RDA and existing DMP tools. By providing unified metadata management and system interoperability, FJORD improves FAIRness tracking and assessment, streamlines administration, and accelerates scientific discovery, while maintaining data sovereignty. This paper details FJORD's tenets, architecture, and impact on institutional data strategies and governance.

Leveraging LLM for Semantic Search and Curation in a National Research Data Catalog
We present a suite of operational services (TRL 7-9) that leverage Artificial Intelligence to augment, not replace, human expertise. We have developed a prototype national catalog for French research data that integrates hybrid search capabilities with a suite of AI-driven tools for metadata enhancement and quality assessment. The catalog combines traditional faceted search with a multilingual semantic search engine, using bi-encoder models for efficient retrieval and cross-encoders for precise reranking. To tackle metadata inconsistency, we utilize right-sized, open-source LLMs like Mistral Small to align entities to controlled vocabularies (e.g., ROR) and generate standardized classifications (e.g. scientific disciplines).

This approach minimizes computational costs and environmental impact while ensuring transparency by always distinguishing between original and AI-generated metadata. Acknowledging metadata can be of low quality, we have also built a novel curation analysis tool using a few-shot LLM to assess the semantic substance of descriptions. Our roadmap focuses on evolving these tools into a proactive "FAIR by Design" ecosystem.

Moving Data Citation Forward: The Multilayered Approach of a Social Sciences Data Repository
Data citation is a crucial, but often neglected, component of data curation and data management. This contribution shows how data citation can be moved forward by different stakeholders. The stakeholders included here are data repositories, researchers, and academic journals. For all of them proper data citation practices offer benefits. However, often, data are not cited correctly, implying that the producer of the original data does not get appropriate credit for the work underlying the data production process. Improper data citation also implies that repositories cannot track their impact.

In this contribution, we analyse several materials, including survey data and journal policies, and elaborate on the steps that are necessary to improve data citation practices, and the respective roles of repositories, researchers, and journals. We derive lessons learned and recommendations for best practice in data curation work, regarding which we want to engage with the IDCC community. In line with the theme of the IDCC 2026 and with current needs and discussions in data citation, our contribution will also reflect on how AI can help repositories, researchers, and journals in improving data citation practices. The focus of this contribution is on social science data in Switzerland; yet the lessons learned are also relevant for other disciplines and national contexts.

On Curating HTR Training Datasets for Romanian Language with use of Transcribathon Tool

This paper presents a workflow for HTR dataset generation in Romanian using Transcribathon's Correct HTR feature. Leveraging citizen-science transcriptions aligned with Transkribus outputs, our case study on Jurnalul lui Dumitru Nistor reduced CER from 15.26% to 0.13%. The approach enables efficient dataset curation and supports scalable model development in low-resource languages.

One Platform, Many Pathways: Reducing Administrative Burden in RDM Through Service Integration

The administrative load on researchers is continuously growing. They are confronted with an increased demand to comply with administrative bureaucracy such as filling in Data management Plans (DMPs), Ethics forms, and much more. Most of these activities involve researchers filling in multiple forms, and sometimes these forms also require filling in the same information repeatedly. In the Netherlands, many universities are facing similar issues where the Research Data Management (RDM) landscape is getting diverse as more and more data management solutions are becoming available. Researchers are also confronted with several rules' regulations and a wider formalization of RDM practices. This comes with new requirements around data management, including the requirement by funders to write a Data Management Plan. This paper presents the research data management platform developed at Vrije Universiteit Amsterdam (VU), designed to reduce the administrative burden for researchers through automated task flows. This provides guidance based on the research lifecycle and connects with support and topic-related support staff.

With the Research Data Management Administration Platform (RDMA), VU Amsterdam created a solution that integrates curation, tools, and services within the research project process. In this paper, we describe the co-creation design process, the challenges involved in scaling across faculties, and the implementation of the platform. We reflect on the importance of tools like the RDMA platform and its impact on the reduction of administrative bureaucracy, whilst offering a model and support for institutions seeking to embed such systems.

Optimizing Historical Data Governance through GEO-Coding: Framework, AI Applications, and Case Study in Bangladesh

This technical paper explores the development and strategic role of Geographic Entity Object (GEO) codes in Bangladesh, focusing on their integration into census and survey operations. GEO-Codes are hierarchical geographic identifiers that represent administrative units from Division down to Village and Enumeration Areas (EAs), enabling accurate data collection, integration, service delivery, and smart governance. Each and every dataset can be tracked using this scientific GEO-code.

The Bangladesh Bureau of Statistics (BBS) initiated the GEO-Code system in 1978 to prepare for the 1981 Population Census. Collaborating with the Land Record and Survey Department, BBS digitized and systematized the Mouza lists and maps into a scientific coding system. Over the years, this system has evolved into a sophisticated, multi-tiered structure used across all national censuses and socio-economic surveys.

Initially, in 1961, Bangladesh had 4 Divisions and 19 Districts. By 2022, this expanded to 8 Divisions and 64 Districts. Population growth also surged, from 50.84 million in 1961 to 169.83 million in 2022. The number of villages increased from about 68,038 in 1974 to 90,049 in 2022. GEO-Codes have been uniquely assigned to each administrative unit: Division, District, Upazila, Union/Ward, Mouza, and Village.

Historically, geo-coding systems trace back to British colonial cadastral surveys and pre-1971 East Pakistan's Thana-level census framework. Post-independence, Bangladesh transitioned to a structured multi-level coding system that has now been digitized. The first fully digital implementation occurred in the 2022 Population and Housing Census, covering over 58,846 Mouzas and 600 Upazilas.

By reviewing both historical practices and contemporary innovations, the paper emphasizes that a robust GEO-Coding system is a cornerstone for inclusive development, precise data governance and national planning. With further coordination, Bangladesh can use GEO-Codes for urban planning, AI-driven policymaking, real-time census & survey analytics and maintain robust control over all the government agencies where it is needed most."

Preserving Under Pressure: The 2016/17 Data Rescue Movement and the Limits of Emergency Curation

This paper offers a retrospective analysis of the 2016/17 Data Rescue movement, a grassroots initiative that mobilized librarians, technologists, and activists to preserve at-risk federal environmental data in response to the anticipated threats posed by the Trump administration. Drawing on 16 qualitative interviews conducted in early 2025, the study examines how participants now reflect on their motivations, methods, and the movement's legacy. It explores the ethical and affective dimensions of emergency curation, the tensions between institutional and community-driven preservation, and the shifting trust in public data infrastructures. Participants expressed a strong sense of civic duty and emotional urgency, but also critical distance from the movement's limitations, particularly its overreliance on downloading as a preservation strategy. The findings underscore that trust in infrastructure is relational and partial, shaped by both political context and social practice. Ultimately, the paper argues that digital preservation in politically volatile times must be grounded in care, accountability, and long-term infrastructural thinking, rather than reactive interventions alone.

Risk and Expertise: How Professional Roles Shape Views of Repository Certification Requirements

Trustworthy Digital Repository (TDR) certification processes are mechanisms through which digital repositories signal quality and commitment to best practices to external stakeholders. In a scholarly landscape with constrained funding and the threat of austerity measures, certification serves as a risk mitigation strategy that repositories use to articulate their value. This study examines repository staff attitudes about CoreTrustSeal (CTS) certification requirements to understand how professional expertise shapes perspectives on what makes repositories trustworthy for long-term digital preservation.

A survey was administered to all CTS certified repositories in fall 2020, with 88 responses from repository staff (53.98% response rate). Respondents ranked the three sections of the CTS Requirements by importance and answered follow-up questions about specific requirements.

Professional roles were categorized as administration, digital preservation, IT, and other, and respondents indicated whether they had experience as CTS reviewers.

Findings demonstrate that respondents consistently ranked Organizational Infrastructure as the most important section of the CTS requirements, followed by Digital Object Management, and then Technology. Responses demonstrated a relationship between respondent expertise and attitudes about the CTS requirements. Additionally, those with experience as reviewers had more consistent views than those without review experience, indicating that exposure to multiple repository contexts through the review process also influences attitudes about certification.. These results suggest that expertise does indeed play a role in attitudes about CTS certification requirements."

Scaling Data Sharing Expertise with AI: a Case Study from DataSeer and Taylor & Francis

This paper outlines the collaborative development of an AI data curation tool to support data sharing in the journal publishing workflow. As scrutiny of research increases, concerns have grown regarding reproducibility, as well as fraud and bad actors in the research lifecycle. Transparent, reproducible, and well-curated data is foundational to restoring confidence. In this paper we describe the current data sharing policy landscape at academic publishers and outline key challenges which might limit further data policy implementation and enforcement on journals. We provide insights into a new approach to data sharing compliance checks, the DataSeer SnapShot tool, and how this tool was developed with the collaboration of the Open Science-, Implementation-, and Editorial Operations teams at academic publisher Taylor & Francis. The potential future iterations of the tool and the implications of its wider implementation are also discussed.

Shades of Grey: Designing Privacy Workflows To Identify, Address And Avoid Sensitive Data Leakage Via Informal Data Transfers

In academic research, data sharing, particularly secondary data reuse, relies heavily on informal networking. 'Grey' transfers of data motivated by research purposes are common. In this paper, working through use cases presented by professionals in digital health, research governance and sensitive data management and publication, we explore the compliance challenges of informal data sharing, its detection, policy challenges such as penalties and associated risks such as accidental data breach and scientific impact. We highlight challenges of maintaining researcher awareness of best practice, given the fast-moving UK legal and regulatory landscape and the need to maintain compliance with standards required by key research partners in the EU. We then explore how good data privacy practices, privacy impact assessments, principles of privacy by design and existing frameworks might be used to support the process of engineering systems that provide the needed flexibility to researchers while minimising the risks.

Standardization Vs. Preservation? Supporting Interoperability by Enhancing Thematic Metadata at Social Science Archives

The presentation ""Standardization Vs. Preservation: Supporting Interoperability by Enhancing Thematic Metadata at Social Science Archives"" addresses the challenges of data standardization and interoperability in social science research. Emphasizing the importance of effective metadata practices, the project ONTOLISST aims to study thematic ontologies with the purpose of improving data discoverability and sharing among diverse research infrastructures.

The project, funded by the European Commission's Horizon Europe program, investigates the varied approaches to thematic metadata creation across research repositories containing social science survey data. By analyzing metadata structures and curation practices, the research seeks to identify and explain commonalities and discrepancies in metadata schemes that hinder interoperability. The study highlights the need for rich metadata documentation while navigating the complexities arising from competing standards and the diversity of data describing practices.

Drawing on data documentation received from the repositories and extensive interviews with data management experts, the project presents two kinds of outcomes: research studies and technical innovation. The results of analysis feed into the development of a semi-automated thematic metadata-generating scheme based on a simplified thesaurus (LiSST). This tool aims to facilitate the integration and accessibility of social science data, fostering connectivity across disciplines and languages. Thus the anticipated outcome is a harmonized metadata structure that upholds the rich, nuanced meanings of original research while promoting discoverability and reuse.

By focusing on the balance between standardization and preservation, ONTOLISST affirms that thoughtful approaches to thematic metadata can yield practical solutions to interoperability challenges, ultimately enhancing the usability and visibility of social science datasets in the global research landscape."

Taming the AI Curator: A Content Focused Data Description Diagnostic and Assistive Writing Tool

We designed an AI based tool to diagnose and help users write clear, accurate, and complete data descriptions. The tool's components include best practices data description guidelines, data descriptions reviewed by experts as few-shot prompts, and chain of thought reasoning to explain the diagnostic outputs. We engineered our prompts and Large Language Model choice so that a score of 8 reflects an acceptable data description. Users can double check the evaluations and the assisted descriptions to minimize scores inconsistent with expert reviewers and hallucinated outputs. The application is crafted to match the standards of our field and to be used with guided intention.

The Challenges of Implementing and Operationalising the CARE Principles

Since 2016, the implementation of the FAIR Principles encouraged re-thinking how data are managed, particularly regarding Indigenous communities which, due to the processes of colonisation, had very little impact on their knowledge and data. Hence, in 2018 the CARE Principles were drafted with the aim to tackle past injustices and support Indigenous communities in governing their data. However, the CARE implementation has not been straightforward as researchers, information experts and Indigenous Peoples faced many challenges starting from the lack of IT infrastructure, skills, funding and metadata. Hence, successful implementation of the CARE Principles requires significant financial and human resources.

It has been six years since the CARE principles were published. The aim of this paper is to analyse the obstacles of implementing them, whilst arguing that the process is one of the biggest contemporary challenges in data curation and the best solutions will eventually be found. The second part of this

paper will address these challenges in the UK by analysing data collected in autumn of 2025 from the DataCite UK consortium members and the webinar guest speakers.

The Conception and Development of Data Steward Training Programs in Hungary: The Roles of Collaboration, Necessity, Flexibility and Community-building

The official and accredited training of data stewards is a relatively new phenomenon worldwide. With the ever-growing presence of the Open Science and Open Data movements within the confines of academia, the need for professionals skilled in state-of-the-art data management and curation is axiomatic.

In Hungary two initiatives started unfolding in almost parallel fashion. The very first graduates of the first higher education level data steward training program at ELTE University received their degree in the autumn of 2024 (after three semesters of attendance). Similarly, the first ’graduates’ (more of an informal, not an accredited program) of the HUN-REN data steward training program received their certificates in January 2025. The HUN-REN data stewards started working in their newly created and acquired positions on the 1st of October.

The evolution of the HUN-REN Data Steward Network reflects a commitment to improving and institutionalizing research data management in Hungary. Each milestone represents a step forward in creating a robust and unified approach to data stewardship, ensuring that research data is managed efficiently and effectively across the network.

As the network continues to grow and develop, it serves as a model for other institutions and countries looking to enhance their own data management practices. The collaborative efforts and progressive initiatives undertaken by HUN-REN demonstrate the importance of data stewardship in the modern research landscape.

Forming the Hungarian national data steward community is based now on the informal collaboration between professionals and beginners from important institutions, marking a step towards the data steward professionalization in Hungary in which both ELTE’s DS training program and HUN-REN informal Data Steward Training program, as well as the HUN-REN Data Steward Network aim to play an important and effective role.

Towards a Shared Understanding of what is Necessary for Long-Term Archiving: EOSC EDEN Core Preservation Processes

The EU Horizon Europe project EOSC EDEN was launched in 2025 to support trustworthy digital preservation in research data infrastructures. This paper introduces a first output of EOSC EDEN, the Core Preservation Processes (CPPs), a structured yet flexible set of process descriptions that articulates the essential steps involved in digital preservation in a hands-on manner. Developed by practitioners and designed as practical guidance, the 30 CPPs describe essential, system-agnostic actions that should be ensured by trusted digital archives. By offering a shared terminology and process model, the CPPs help bridge gaps between research data management and digital preservation, providing the communities a powerful tool to support training, policy, and system

development. They shall address the challenge of making digital preservation knowledge more accessible, actionable and interoperable within the EOSC ecosystem and beyond.

Towards best practices in Research Data Management: Guiding institutional development in Irish Research Performing Organizations trough collaborative self-evaluation

This paper reports on a series of collaborative self-evaluations on support services for Research Data Management (RDM) at Irish Research Performing Organisations (RPOs). The evaluations are carried out as part of the iFrame project, which is tasked with drafting a national RDM framework for Ireland.

The aim of the evaluations is to:

1. Assess institutional maturity of the services provided and enable the identification of areas for improvement via tailored reports for individual institutions.
2. Gather inputs for a landscape report on the state of RDM service provision in Irish RPOs.
3. Identify best practices in institutions that can be integrated into the upcoming national RDM framework.

The paper commences with a contextual discussion of the Irish higher education environment and the research policy environment that advocates the development of RDM practices. This is followed by a brief review of the literature that examines RDM support and practice in RPOs, utilising maturity models. The main section of the paper outlines the research rationale, design, and methodology employed in the iFrame project, which is built around the principles of openness, collaboration, and process orientation. The paper concludes with an overview of results and anticipated impact, and provides an outlook on the next steps of the project and the development of RDM at Irish RPOs.

Towards Sustainable Curation: Evaluation of Cost and Accuracy of AI Tools in Scaling Annotation Tasks in Curation of Biomedical Literature

Here we compare the performance and cost of four language models (GPT 4, Llama 3, Gemma 2 and Mixtral 8x7b) in the lightweight task of population group curation. Our findings provide insight into potential sustainable curation practices in the presence of limited resources.