Building Responsible and Sustainable Open Data Literacy Skills for Early Career Researchers: A Decade of the SoRDS Programme

ESSAY

SHAILY GANDHI
STEVE DIGGS
MARCELA ALFARO CÓRDOBA
LOUISE BEZUIDENHOUT
RAPHAEL COBE
SARA EL JADID
BIANCA PETERSON
ROBERT QUICK
HUGH SHANAHAN
SHANMUGASUNDARAM VENKATARAMAN
EKPE OKORAFOR
VEERLE VAN DEN EYNDEN

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

ABSTRACT

In today's data-centric research environment, effective data literacy is essential for ensuring data usability, integrity, and reproducibility. The Schools of Research Data Science (SoRDS), launched in 2016 in partnership with the Committee on Data (CODATA) and the Research Data Alliance (RDA), marks a decade of impactful training and capacity-building for early-career researchers (ECRs) from low- and middle-income countries (LMICs) with its tenth anniversary reached in August 2025. This paper examines its distinctive, holistic approach to equipping researchers with core competencies in data science, open science, and research data management (RDM). Unlike traditional programmes focused solely on technical skills, SoRDS integrates principles of data ethics, reproducibility, data stewardship, and interdisciplinary collaboration into its curriculum.

Central to its mission is the 'Train-the-Trainer' model, which empowers participants to become instructors and regional leaders, creating a sustainable and scalable community of practice. SoRDS not only provides technical training but also fosters a culture of openness and inclusivity, ensuring that the benefits of the data revolution reach underserved research communities. Over the past decade, the schools have been hosted in diverse regions, adapting content to local contexts and creating a strong global network of alumni, mentors, and institutions.

Crucially, SoRDS advances an 'RDM in context' approach, prioritizing what is practical, relevant, and achievable in low-resource settings. SoRDS tailors its training to the realities of LMICs, making its content more applicable, sustainable, and impactful for these communities. Drawing on a decade of documentation, this paper provides a retrospective synthesis of SoRDS' development, global expansion, alumni impact, and lessons learned, situating these within the broader training landscape. In particular, it

CORRESPONDING AUTHOR:
Shaily Gandhi
Interdisciplinary Transformation University Austria, Austria
shaily.gandhi@gmail.com

KEYWORDS:
research data management; FAIR; artificial intelligence in RDM; data literacy; open science; reproducible research

TO CITE THIS ARTICLE:
Gandhi, S., Diggs, S., Córdoba, M.A., Bezuidenhout, L., Cobe, R., El Jadid, S., Peterson, B., Quick, R., Shanahan, H., Venkataraman, S., Okorafor, E. and Van den Eynden, V. 2026 Building Responsible and Sustainable Open Data Literacy Skills for Early Career Researchers: A Decade of the SoRDS Programme. *Data Science Journal*, 25: 12, pp. 1–13. DOI: https://doi.org/10.5334/dsj-2026-012

draws comparisons with other training activities, noting the specific niche SoRDS has in this landscape. Finally, it outlines priorities for the programme's next stage.

The primary focus of this paper is to provide a reflective, evidence-based account of SoRDS: its historical development, its unique pedagogical model, its global expansion, and the lessons learned over a decade of implementation in low- and middle-income research environments.

INTRODUCTION

In the evolving landscape of data-centric research, effective data skills development is paramount for ensuring data integrity, accessibility, and compliance with the FAIR principles (Findable, Accessible, Interoperable, and Reusable) (Wilkinson *et al.*, 2016). The Carpentries, beginning with Software Carpentry in 1998, pioneered global data skills training through hands-on workshops, evidence-based teaching methods, and openly licensed materials. This community-driven approach to teaching foundational coding and computational skills established influential practices that shaped data education worldwide.

The Schools of Research Data Science (SoRDS) builds directly on The Carpentries' (Software Carpentry, Data Carpentry, and Library Carpentry—hereafter referred to as 'The Carpentries') materials and methods while extending this foundation in critical ways (Jordan, Michonneau and Weaver, 2018). SoRDS integrates technical skills training with research data management (RDM) principles, open science practices, and contextualized approaches designed specifically for early-career researchers (ECRs) in low- and middle-income countries (LMICs). This combination of proven pedagogical methods with expanded scope and targeted context has proven instrumental in equipping researchers with essential data skills while building sustainable capacity in underserved research communities.

The Venn diagram in Figure 1 illustrates the intersection of four foundational domains in data skills education: Open Science, Research Data Management, Data Science, and Research Integrity. At the core of this intersection is the SoRDS, which functions as a central integrative hub combining principles, practices, and training methods from all four areas.

This paper adopts a descriptive and reflective methodology that synthesizes a decade of SoRDS programme documentation, internal planning materials, openly available instructional resources, annual reports, alumni newsletters, and publicly accessible GitHub repositories that track the curriculum's evolution. The historical timeline, curriculum changes, and examples of alumni impact are drawn exclusively from verifiable, publicly documented SoRDS activities.

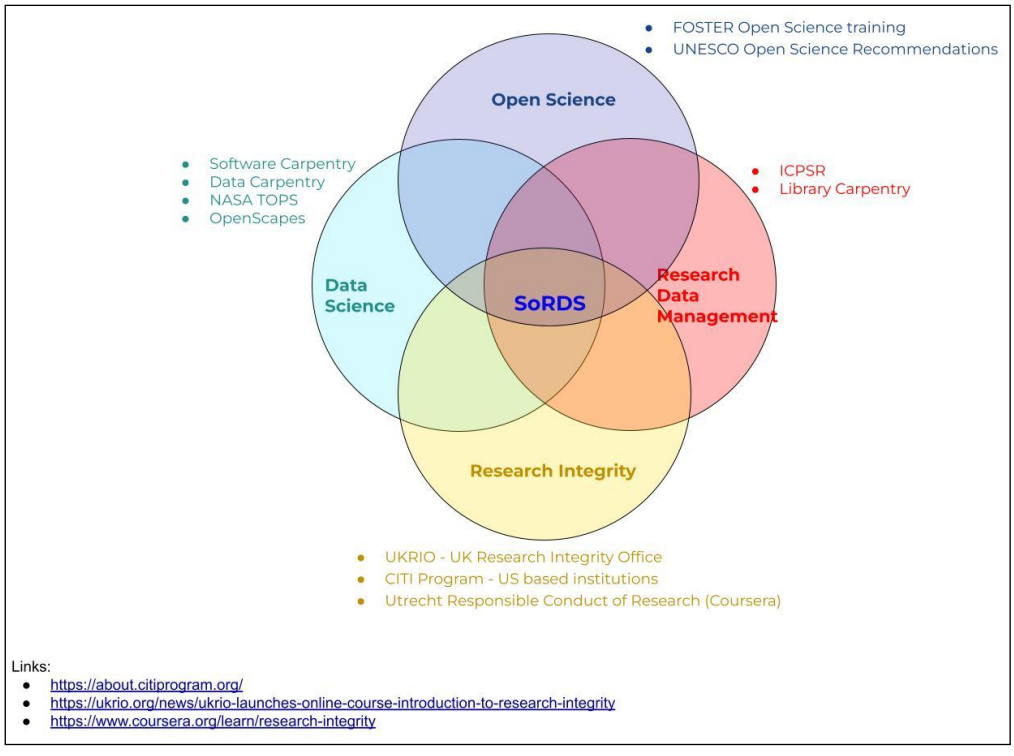

Figure 1 Curriculum design of SoRDS.

OVERVIEW OF SCHOOLS OF RESEARCH DATA SCIENCE (SoRDS)

SoRDS is an initiative by the Committee on Data (CODATA) of the International Science Council (ISC) and the Research Data Alliance (RDA), established to provide ECRs from LMICs with foundational data science skills essential for 21st-century research. The curriculum is developed entirely based on open source tools and materials. As a structured and comprehensive programme, SoRDS incorporates foundational Carpentries-style skills while also extending into deeper FAIR-aligned RDM concepts and advanced computational topics (Shanahan, Hoebelheinrich and Whyte, 2021). Central to the SoRDS curriculum are Open and Responsible Research Principles for promoting transparency and accessibility in research. It also emphasizes that data should be appropriately managed, preserved, and annotated for future use. This includes analysis techniques that cover statistical methods, machine learning, and data visualization.

The central concept of covering a broad but shallow range of topics in Data Science, RDM, Open Science, and Research Integrity was built in from the first instances of SoRDS. Initially, the schools covered core Carpentries material (Shell, Git, R, and SQL), Data Visualization, Machine Learning, and Computational Infrastructures. Author Carpentry was taught as an optional module in the evenings. In 2017, there was a substantial readjustment of the Research Integrity and RDM materials for better alignment. Based on participant feedback and needs assessment, the curriculum team replaced SQL with Author Carpentry as a core module, recognizing its greater relevance for ECRs.

Ethics exercises for each module were introduced to ensure that the modules always had a reflective component on Research Ethics (Bezuidenhout, Quick and Shanahan, 2020). In 2019, an applied RDM Lab, with a hands-on module in which participants practice core RDM tasks such as metadata creation, data documentation, FAIR assessment, and repository preparation, was introduced alongside an Information Security module. Beginning in 2024, SoRDS also launched a Python-based version of its computational modules, offered in parallel to the original R-focused track. This allowed participants to choose the programming language most relevant to their research context.

Each annual iteration of the school is jointly designed through a structured planning cycle involving curriculum review, participant needs assessment, and evaluation of previous years' feedback collected through International Centre for Theoretical Physics (ICTP) post-event surveys. Major shifts, such as the integration of ethics exercises in 2017, the launch of the RDM Lab in 2019, and the Python version in 2024, were implemented in response to documented participant feedback and instructor debrief reports.

Participants apply SoRDS-acquired skills in academic instruction, research design, institutional data policy, and regional training. Many advance to leadership roles, acting as multipliers by embedding RDM in curricula, promoting FAIR data, and organizing workshops modeled on SoRDS (Cobe *et al.*, 2023). Institutions have launched certificate programmes or integrated data skills into academic offerings as direct outcomes. For example, University College Cork has developed a standalone module in Data Stewardship that was developed by the team who attended the Data Stewardship course developed by SoRDS (University College Cork, n.d.). These curricula often take the form of multi-module certificate programmes or part-time courses, such as the University of Vienna's two-semester Data Steward certificate (University of Vienna, 2022), and include competency frameworks recommended by the European Open Science Cloud (EOSC) to professionalize data steward roles (Basalti *et al.*, 2024; Demchenko *et al.*, 2021). Early initiatives highlighted the importance of teaching data science and computational skills in low- and middle-income countries (Shanahan *et al.*, 2015). Also initiatives have explored integrating data literacy and research data management competencies directly into disciplinary curricula, particularly within physics laboratory education (Bode, Jaeger & Schneidewind, 2023a, 2023b). Two-tiered training models have also been proposed to improve researchers' data management practices (Read *et al.*, 2019). The programme's adaptable content and delivery suit varied linguistic, infrastructural, and disciplinary contexts, ensuring ongoing relevance. This scalability strengthens individual competencies while building institutional and national capacity for responsible RDM.

Figure 2 illustrates the chronological progression of the SoRDS from its inception in 2016 through to its 10th anniversary in 2025. Each year marked in the timeline represents continued efforts to deliver foundational and increasingly advanced data science training to ECRs.

From as early as 2017, SoRDS curricula began integrating advanced topics such as machine learning, computational workflows, data visualization, and reproducible research practices. Based on the programme's internal archives and the annual planning documents maintained by CODATA and ICTP, which record each year's curricular updates and structural changes. These topics were taught using real-world datasets and hands-on exercises to ensure deep, practical understanding. Evidence of this progression can be found in the openly available course materials on the CODATA-RDA GitHub repository (CODATA-RDA-DataScienceSchools, 2025), which documents the evolving scope and sophistication of the training content over the years (Quick, Córdoba, Cobe *et al.*, 2023).

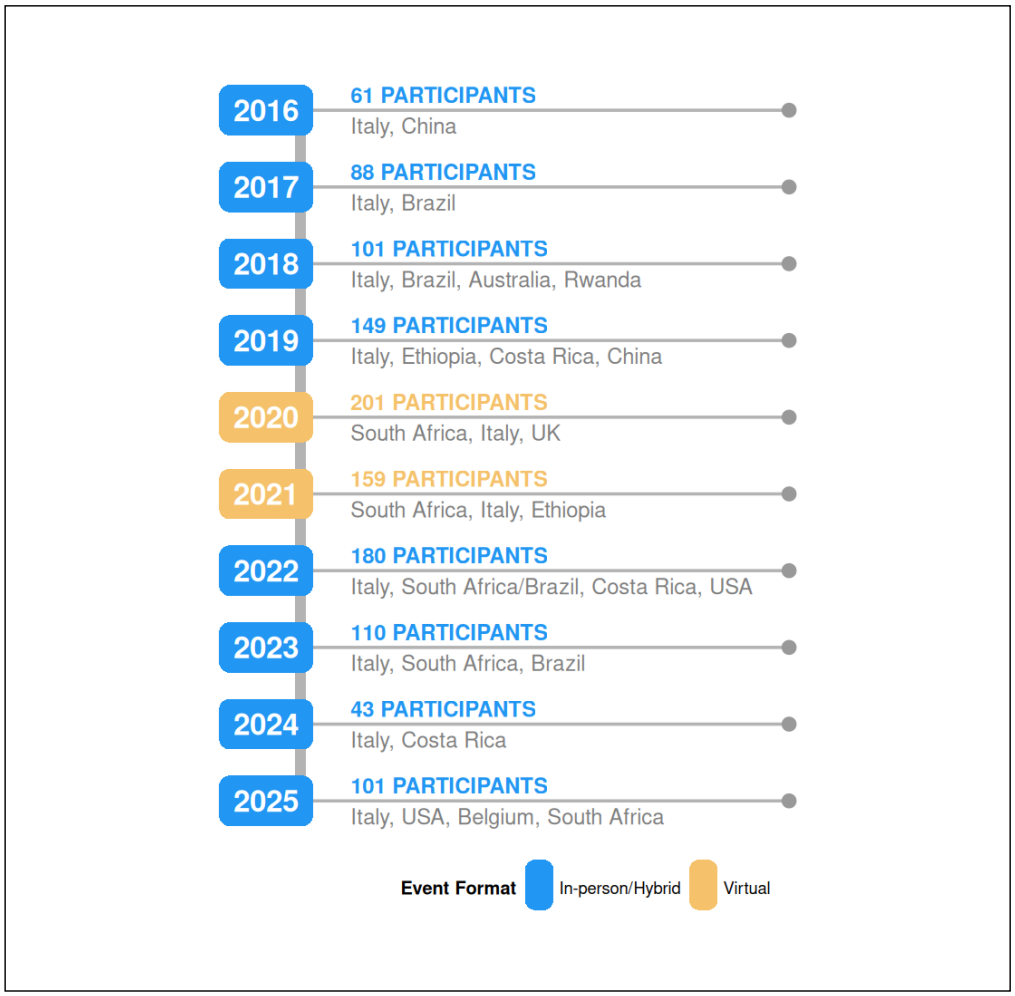

Figure 2 Timeline for SoRDS.

The timeline in Figure 2 underscores number of participants trained each year and the geographic locations where the schools were delivered. It also reflects the programme's adaptation during the COVID-19 period, when delivery shifted temporarily to virtual formats before returning to in-person and hybrid events. By 2021, over 400 participants from 40 countries had been trained across 10 annual schools (Bezuidenhout *et al.*, 2021). By 2023, this number of participants had grown to over 1,000 ECRs trained in 24 events held in 10 countries worldwide (Quick, Córdoba, Diggs *et al.*, 2023), reflecting both the programme's expansion and diversification into regional and domain-focused events. Notably, the COVID-era events were more heavily attended due to remote participation, as shown in Figure 2.

The map in Figure 3 highlights SoRDS' global engagement across Africa, Asia, Latin America, Oceania, Europe, and North America. Countries shown have hosted training, sent participants, or provided instructors, reflecting a commitment to inclusion. Africa's involvement spans the western (Ghana, Nigeria), eastern (Kenya, Ethiopia, Tanzania), and southern regions (South Africa, Botswana). Asia features India, Bangladesh, China, Indonesia, and the Philippines, meeting the demand for data science and RDM training. In Latin America, Brazil and Colombia are active hosts, with Costa Rica as a key coordination hub, enhancing accessibility. Europe's Italy and Belgium lead in organization and expertise, frequently hosting in-person schools and contributing experienced trainers.

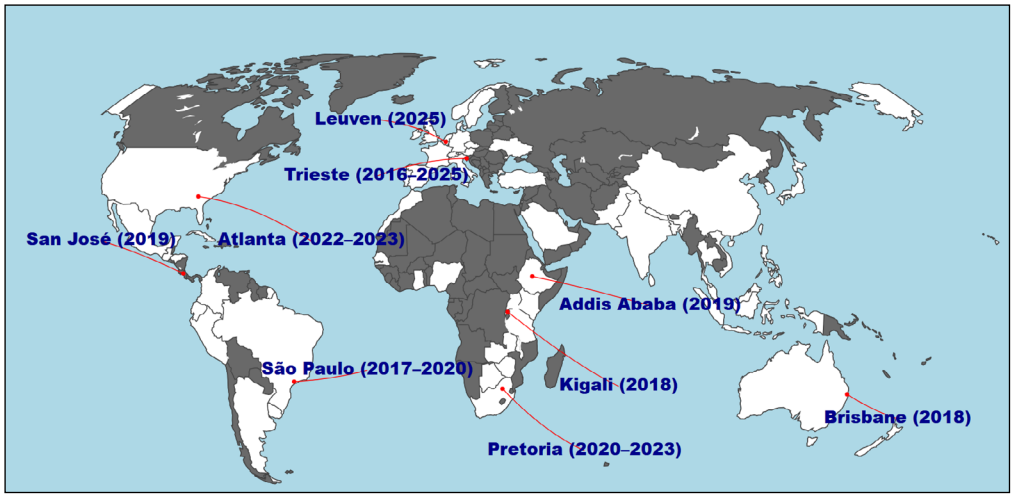

Figure 3 Global Footprint of Schools of Research Data Science (SoRDS): countries in white are countries where SoRDS has held events or whose early-career researchers have participated.

RDM TRAINING CONTEXT AND ITS INFLUENCE ON SoRDS

As data-driven research continues to grow, robust RDM practices are essential for maintaining data integrity, enhancing accessibility, and aligning with FAIR principles. While SoRDS incorporates several foundational Carpentries-style technical skills (e.g., shell, version control, R/Python basics), it extends beyond them by integrating broader FAIR-aligned RDM concepts, data ethics, research integrity, and contextualized training tailored for LMIC settings (Jordan, 2018; Shanahan, Hoebelheinrich and Whyte, 2021).

Despite the growing recognition of RDM as essential to good scientific practice and increasingly mandated by funders and publishers (Biernacka, Helbig and Buchholz, 2021; Kanza and Knight, 2022; He *et al.*, 2023; Majid *et al.*, 2018; Wilkinson *et al.*, 2016;), RDM training remains challenging to deliver effectively. The inherent complexity of RDM, shaped by diverse data practices, multiple stakeholders, and rapidly evolving standards and technologies (Oo *et al.*, 2021), has resulted in persistent skills gaps, particularly among ECRs, many of whom continue to rely on ad hoc solutions due to limited formal training (Goben and Griffin, 2019; Krahe *et al.*, 2020; Maienschein, MacCord and Elliott, 2019; Wiley and Kerby, 2018). These challenges were SoRDS' emphasis on broad-but-shallow coverage, integrated ethics, and applied, hands-on learning.

SoRDS CURRICULUM

While SoRDS is designed primarily for ECRs, a separate and more specialized set of professional data-steward curricula has emerged to support librarians, data stewards, and research support staff. The curriculum covers policy interpretation, infrastructure design, and teaching techniques.

RDM CURRICULUM CHARACTERISTICS AS REFLECTED IN SoRDS

COMPLIANCE AND BEST PRACTICES

Throughout its development, SoRDS placed a strong emphasis on adherence to the FAIR principles and other relevant policies, legal and ethical requirements, licensing, documentation, and metadata practices (Rantasaari, 2022; Kanza & Knight, 2021; LIBER RDM Working Group, 2020). These elements became increasingly prominent across successive programme iterations as external mandates and community standards evolved.

Data Stewardship and Governance: SoRDS consistently framed RDM as a component of long-term stewardship, emphasizing data integrity, reliability, transparency, and reproducibility to support discovery and reuse (Wilkinson *et al.*, 2016; Rantasaari, 2022). This perspective informed both curriculum content and the practical exercises used across multiple cohorts.

Customization and Contextualization: In response to the diversity of disciplinary practices and participant backgrounds, SoRDS progressively incorporated contextualized and discipline-aware training approaches, reflecting broader trends toward user-centered RDM education

(Oo *et al.*, 2021). This was particularly important for supporting ECRs working in heterogeneous and resource-constrained research environments.

Practical and Interactive Learning: Hands-on, practice-based learning has been a core component of SoRDS since its early iterations, linking RDM principles to real research workflows through applied exercises such as Data Management Plan development (Yu, Deuble and Morgan, 2017; Rantasaari, 2022). Over time, interactive and participatory teaching approaches using red and green sticky notes and having alumni as helpers in the class were further refined to enhance engagement.

Data Engineering Shift: More recently, SoRDS has adapted to emerging shifts in RDM practice that move beyond bibliographic descriptions of datasets toward data engineering approaches requiring richer, machine-actionable metadata (CODATA & RDA, 2024). This shift influenced later curriculum updates, ensuring that SoRDS remained aligned with evolving expectations for data interoperability and reuse.

Taken together, these curriculum characteristics illustrate how SoRDS evolved in principle with global RDM training priorities over a decade of delivery, demonstrating a sustained process of adaptation rather than a static curriculum model.

SoRDS IN A GLOBAL TRAINING LANDSCAPE

In contrast to the broad data lifecycle approach curricula, The Carpentries focus on developing technical skills for data analysis and software development through practical hands-on tool-based instruction, often delivered in short, self-contained workshops (Biernacka, Helbig and Buchholz, 2021). The Carpentries' emphasis on data literacy, reproducible research, hands-on tool-based instruction, open and collaborative curriculum development, and Train-the-Trainer models (Biernacka *et al.*, 2020; Doehle, Bjornen and Chartier, 2019; Kanza and Knight, 2022; Diggs, 2025) informed several pedagogical elements later adapted within the SoRDS programme.

HOW RDM TRAINING LITERATURE INFORMED SoRDS DEVELOPMENT: KEY DISTINCTIONS AND SYNERGIES

RDM curricula are generally broader in scope, encompassing the full data lifecycle, compliance, governance, and policy considerations with domain-specific and multi-stakeholder perspectives. The Carpentries emphasize the acquisition of hands-on, technical skills for data handling, analysis, and reproducible workflows. These two educational approaches, although distinct in focus, are highly complementary and reinforce each other in practice.

Within this training landscape, SoRDS was intentionally designed to bridge these approaches by ensuring that the acquisition of 'technical and complementary non-technical skills' centered around the theory behind best practices in RDM, Open Science, and Open and Responsible Research is coupled with the 'hard-skills' acquired through practical application.

Crucially, the SoRDS programme is designed to make these competencies usable and useful in low-resourced settings as we spend significant time discussing context, common challenges, and adaptive strategies to ensure relevance and applicability. Moreover, this approach enables the development of transferable skills and the potential to expand research beyond traditional domains into new interdisciplinary areas, as well as repurpose data from one domain into another, potentially a completely different one.

This comparison demonstrates why SoRDS occupies a unique middle ground: it leverages the strengths of technical training traditions (e.g., Carpentries pedagogy) while expanding them into a holistic RDM-focused capacity-building programme tailored for LMIC contexts.

UNIQUE CONTRIBUTIONS OF SoRDS

Foundational skills for RDM implementation

The Carpentries teach essential skills like scripting, version control, and data cleaning, which provide a foundation for RDM. SoRDS builds on this foundation taking a holistic approach,

combining technical and theoretical RDM competencies for ECRs, ideal for passing on knowledge to drive change. Building on The Carpentries' live coding, peer instruction, and volunteer-led teaching, SoRDS adds RDM principles, open research, security, ethics, and policy. Through practical exercises, it helps ECRs master both tools and values underpinning responsible data science.

Integration with artificial intelligence (AI) and data engineering

Emerging AI-assisted approaches to metadata management and data quality assessment highlight new directions for RDM training that SoRDS focuses on to engage with in the future. AI technologies can automate the detection of data quality issues, while foundational technical training provides the context needed to understand and resolve these outputs (Diggs, 2025).

Adaptability to policy shifts

SoRDS responds to evolving data-sharing mandates by integrating policy awareness, compliance, and practical implementation strategies within its curriculum. While RDM curricula are often theoretical and technical training programmes mostly focus on tools, SoRDS integrates these perspectives to foster a data-literate, resilient research culture, advancing a sustainable, inclusive ecosystem grounded in best practices, reproducibility, and openness.

Train-the-Trainer model

The Train-the-Trainer approach at SoRDS builds institutional capacity by training librarians, data stewards, and research support staff in RDM skills and pedagogy. SoRDS has run a pilot in collaboration with KU Leuven and a consortium of Ecuadorian Universities to formalize its on-boarding of instructors, using a variant of the principle of 'observe one, do one, teach one' (Kotsis and Chung, 2013), where trainee instructors shadow and assist teaching at one training event, then train others at the next event. It promotes cascading knowledge through programmes, combining technical content and instructional design (Schmidt *et al*., 2017).

Complementarity and integration

In summary, the SoRDS model integrates technical skill development, theoretical RDM foundations, and capacity-building approaches into a cohesive training programme. This comparison is presented to contextualize the pedagogical choices underpinning SoRDS, which integrates and adapts several of these elements into a comprehensive, LMIC-responsive model.

TRIESTE'S 10-YEAR CELEBRATION

Over the past decade, the annual SoRDS schools in Trieste, Italy, in collaboration with the ICTP, have played a pivotal role in data science education, contributing significantly to the advancement of data practices. Since its inception, SoRDS has expanded globally, by conducting regional programmes in multiple countries and tailoring content to local needs. This global growth directly supports the programme's goal of enhancing data skills among researchers worldwide, as documented in the SoRDS newsletters (Cobe *et al*., 2023). These events have facilitated collaborations, knowledge exchange, and the development of advanced schools focusing on domain-specific areas, thereby strengthening the global data community.

For instance, the Urban Data Science summer school hosted in India in 2018 and 2019 was a result of collaborations initiated at the foundational school in Trieste in July 2017 (Gandhi and Anyiam, 2022). Several peer-reviewed publications highlight this alumni-driven impact. Tachie *et al*. (2024) report that Alberta Aryee and Nii Adjetey Tawiah, students at the 2022 Atlanta school, co-authored a study with former student Christabel Tachie on classifying oils and margarines using FTIR spectroscopy and machine learning. Bezuidenhout *et al*. (2019) include Ola Karrar, a Trieste 2018 alumna, among the authors of a PLOS One article examining the overlooked effects of economic sanctions on academia. Quick, Córdoba, Diggs *et al*. (2023) feature co-chairs and alumni of SoRDS in an IEEE conference paper detailing a SoRDS event for health equity researchers at Minority Serving Institutions. Bezuidenhout, Quick and Shanahan (2020) have designed modular data ethics instruction, reflecting collaborative curriculum development across alumni and hosts. Alzate-Cardona *et al*. (2018) describe Oscar

Arbeláez-Echeverri's alumni of SoRDS work on the 'Vegas' Monte Carlo simulation software for magnetic materials; Oscar acknowledges Quick's mentorship from the 2017 Trieste school in utilizing OpenGrid. These examples underscore a thriving ecosystem where SoRDS alumni transition into contributors by publishing collaboratively, leading new events, and enriching the data skill landscape through sustained engagement. These alumni continue to contribute actively to the SoRDS network, frequently returning as organizers and support staff at later events.

In August 2025, SoRDS celebrated its 10th anniversary at the ICTP in Trieste. This milestone event celebrated a decade of SoRDS' commitment to advancing RDM and data science education globally (CODATA, 2025; ICTP, 2025a). Over the past 10 years, SoRDS has played an instrumental role in building the capacity of ECRs, particularly from LMICs, with essential skills in data stewardship, open science, and computational methods. Bezuidenhout *et al.* (2021) report, 'the results of the survey strongly support the SoRDS' long-term goals of facilitating data science training/capacity building within LMICs, and to foster communities of ECRs conducting responsible and open data science research,' with 90% of alumni continuing to apply these skills in their work. The anniversary highlighted the programme's impact on fostering international collaboration, enhancing data literacy, and promoting equitable access to data science training. By bringing together alumni, instructors, and stakeholders, the celebration aimed to reflect on SoRDS' achievements and chart a course for its future contributions to the global community.

In 2025, the advanced workshops at SoRDS were uniquely curated by its alumni, celebrating a decade of impact and growth. As part of the school's 10th anniversary, these workshops showcased how past participants applied their data expertise to their fields, returning to lead domain-specific sessions. The three parallel workshops, ranging from Big-Data Analytics, Computational Infrastructures, and Urban Data Science, demonstrated the diverse application of data skills at the (ICTP, 2025b). This approach not only highlighted the practical impact of the school's training but also reinforced the ongoing contribution of its alumni to the global research community. These alumni continue to contribute actively to the SoRDS network, frequently returning as organizers and support staff at later events. The planning for the 10th anniversary followed the established SoRDS multi-stage design cycle, including instructor nomination, curriculum mapping, and alumni-contributed workshop proposals.

ARTIFICIAL INTELLIGENCE IN RESEARCH DATA MANAGEMENT (RDM)

Artificial intelligence (AI) is emerging as a transformative contributor to RDM. Automated machine learning (AutoML) approaches now support end-to-end data pipeline tasks, such as cleaning, missing-value imputation, feature generation, and preprocessing. AI helps in reducing manual burden and accelerating the preparation of large, complex datasets (Mumuni and Mumuni, 2024).

AI-driven tools can also be used to automate data quality validation, detecting anomalies and inconsistencies by learning from historical patterns to proactively flag potential errors (Tamm and Nikiforova, 2025). On the metadata front, emerging AI-assisted metadata management frameworks can facilitate automated metadata generation and curation, improving data discoverability and supporting FAIR data practices (Davenport & Redman, 2022; Yang *et al.*, 2025). While AI-based security mechanisms are still evolving in RDM, analogous applications in cybersecurity, where AI models detect unusual access behaviors, suggest promising potential for safeguarding sensitive research data.

As the SoRDS curriculum continues to evolve, integrating AI skills has been identified as an important area for future development. In the context of SoRDS, AI-related content is beginning to be incorporated into both foundational and advanced modules. Discussions during the 2023 and 2025 planning cycles identified AI-assisted metadata generation, automated data quality validation, and reproducible machine-learning workflows as priority skill areas for future SoRDS iterations. Concepts of Machine learning and Artificial Neural Networks were introduced early on from 2017 in the basic schools. Several alumni-led advanced workshops, like Urban Data Science in 2025 and Research Data Management lab in 2024 and 2025, have already piloted

introductory materials on AI-enabled data analysis, and SoRDS intends to formalize these components as part of its evolving curriculum which will be implemented in 2026 schools. This integration aligns with the programme's long-term strategy to ensure that ECRs are prepared to navigate emerging technologies that directly affect RDM practice.

In the longer term, the planned updated curriculum for AI will develop a workflow, namely cleaning and annotating data using best RDM practices, followed by analysis using Machine Learning methods and ensuring that this is done in an ethical fashion. This workflow will then inform the detailed updates to the curriculum.

CONCLUSION

Integrating the broad range of competencies that SoRDS has identified into educational programmes is essential for preparing researchers to navigate data-intensive research. Initiatives like SoRDS play pivotal roles in enhancing data literacy and promoting best practices in data management. The Train-the-Trainer model effectively scales data skills education, fostering communities of practice. Trieste's contributions over the past decade of Data Schools have been instrumental in advancing data skill enhancement globally. Furthermore, the incorporation of AI into RDM processes offers promising avenues for automating routine tasks, improving data quality, and enabling sophisticated analyses. Ultimately, while the wider landscape of RDM and data literacy initiatives continues to evolve, the decade-long experience of SoRDS offers a distinctive, empirically grounded model for scalable, context-sensitive capacity building. The programme's history, alumni-driven growth, and integration of ethics, stewardship, and emerging technologies represent a coherent contribution to the field and a practical framework for future global RDM training efforts.

Looking ahead, SoRDS plans to expand its regional implementations, increase the number of alumni-led advanced schools, formalize its Train-the-Trainer certification, and integrate additional modules addressing emerging topics such as AI-assisted metadata generation and reproducible computational workflows. These next steps reflect SoRDS' commitment to continued evolution based on documented community needs.

SUPPLEMENTARY FILES

Due to the volunteer-driven nature of the initiative over the past 10 years, all reported figures are estimates. The primary focus has been on operations and event execution rather than formal bookkeeping. As the schools have relied heavily on volunteer support, maintaining detailed attendance records has not been a priority.

REPRODUCIBILITY

Materials from the first workshop are available online (Quick, 2016). Beginning in Kigali in 2018, a GitHub repository was created to gather materials for workshops, which includes presentations and exercises. Occasional snapshots of these repositories were extracted and assigned persistent identifiers to supplement publication submissions. The most recent snapshot of materials was created after Trieste in 2023 (Quick, Córdoba, Cobe *et al*., 2023).

SoRDS encourages the adoption and reuse of all materials and methods used during events, and publishes all materials openly with CC-by license on the GitHub Repository (CODATA-RDA-DataScienceSchools, 2025).

ETHICS AND CONSENT

All the data quoted in this paper is publicly available via the SoRDS website and newsletter.

ACKNOWLEDGEMENTS

We acknowledge the support, collaboration, and contributions from SoRDS, CODATA, RDA, and ICTP. Their continued partnership has been invaluable in advancing our work.

FUNDING INFORMATION

We appreciate the long-term, consistent support provided by ICTP, CODATA, and RDA.

We also recognize the generous support for individual events from RDA-US, FAIRsFAIR, Springer Nature, RStudio, Microsoft Research, ICTP-SAIFR, University of Pretoria, Indiana University, DANS, UOS VLIRS, EOSC, EGI, UNA (Costa Rica), CONARE, and EAIFR.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Shaily Gandhi orcid.org/0000-0002-8893-2130
Interdisciplinary Transformation University Austria, Austria

Steve Diggs orcid.org/0000-0003-3814-6104
University of California Office of the President, United States

Marcela Alfaro Córdoba orcid.org/0000-0002-7703-3578
University of California, Santa Cruz, United States

Louise Bezuidenhout orcid.org/0000-0003-4328-3963
CWTS Leiden University, The Netherlands

Raphael Cobe orcid.org/0000-0002-0852-2183
São Paulo State University, Brazil

Sara El Jadid orcid.org/0000-0001-9793-5061
Queens University, Belfast, United Kingdom

Bianca Peterson orcid.org/0000-0001-6927-9159
Fathom Data, South Africa

Robert Quick orcid.org/0000-0002-0994-728X
UITS Research Technologies, Indiana University, United States

Hugh Shanahan orcid.org/0000-0003-1374-6015
Department of Computer Science, Royal Holloway, University of London, United Kingdom

Shanmugasundaram Venkataraman orcid.org/0000-0002-3200-2698
Data Archiving and Networked Services (DANS), The Hague, The Netherlands

Ekpe Okorafor orcid.org/0000-0003-1615-8964
Accenture Applied Intelligence/Nigerian British University, Nigeria

Veerle Van den Eynden orcid.org/0000-0003-2542-2747
UHasselt, Belgium

REFERENCES

Alzate-Cardona, J.D., Sabogal-Suárez, D., Arbeláez-Echeverri, O.D. and **Restrepo-Parra, E.** (2018) 'VEGAS: Software package for the atomistic simulation of magnetic materials', *Revista Mexicana de Física*, 64(5), pp. 490–497. Available at: https://doi.org/10.31349/RevMexFis.64.490

Basalti, C., Fazekas-Paragh, J., Forni, M., van Gelder, C., Hasani-Mavriqi, I., Janik, J., Kalová, T., Kuchma, I., Lindroos, H., Lütcke, H., Pinnick, J., Raga, N., Thorpe, D. and **Wildgaard, L.** (2024) *Recommendations for Data Stewardship Skills, Training and Curricula with Implementation Examples from European Countries and Universities* (Version v1) [Report]. EOSC Task Force on Data Stewardship Curricula and Career Paths. Zenodo. Available at: https://doi.org/10.5281/zenodo.10573892

Bezuidenhout, L., Drummond-Curtis, S., Walker, B., Shanahan, H. and **Alfaro-Córdoba, M.** (2021) 'A school and a network: CODATA-RDA Data Science Summer Schools alumni survey', *Data Science Journal*, 20(10). Available at: https://doi.org/10.5334/dsj-2021-010

Bezuidenhout, L., Karrar, O., Lezaun, J. and **Nobes, A.** (2019) 'Economic sanctions and academia: Overlooked impact and long-term consequences', *PLOS One*, 14(10), p. e0222669. Available at: https://doi.org/10.1371/journal.pone.0222669

Bezuidenhout, L., Quick, R. and **Shanahan, H.** (2020) '"Ethics when you least expect it": A modular approach to short course data ethics instruction', *Science and Engineering Ethics*, 26(4), pp. 2189–2213. Available at: https://doi.org/10.1007/s11948-020-00197-2

Biernacka, K., Bierwirth, M., Buchholz, P., Dolzycka, D., Helbig, K., Neumann, J., Odebrecht, C., Wiljes, C. and **Wuttke, U.** (2020) *Train-the-trainer concept on research data management* (Version 3.0). Zenodo. Available at: https://doi.org/10.5281/zenodo.4071471

Biernacka, K., Helbig, K. and **Buchholz, P.** (2021) 'Adaptable methods for training in research data management', *Data Science Journal*, 20(1), p. 14. Available at: https://doi.org/10.5334/dsj-2021-014

Bode, J., Jaeger, P. and **Schneidewind, S.** (2023a) 'Datenkompetenz im Physikstudium — ein Erfahrungsbericht', *arXiv*. Available at: https://doi.org/10.48550/arxiv.2301.03455

Bode, J., Jaeger, P. and **Schneidewind, S.** (2023b) 'Integrating data literacy into university curricula student centred learning in undergraduate physics lab courses', *Proceedings of the Conference on Research Data Infrastructure*, 1. Available at: https://doi.org/10.52825/CoRDI.v1i.349

Cobe, R. (Ed.), Shanahan, H., Bezuidenhout, L., Quick, R., Peterson, B., Okorafor, E., Alfaro Córdoba, M., El Jadid, S., Venkataraman, S., Van den Eynden, V. and **Gandhi, S.** (2023) *CODATA-RDA Schools of Research Data Science* (October 1, 2020 – November 30, 2023; Version v1) [Newsletter]. Zenodo. Available at: https://doi.org/10.5281/zenodo.16875357

CODATA. (2025) *Research data science summer schools*. Available at: https://codata.org/initiatives/data-skills/research-data-science-summer-schools/ (Accessed: 14 February 2025).

CODATA, and **Research Data Alliance.** (2024) *Enabling global FAIR data: WorldFAIR policy recommendations for research infrastructures* [Policy brief].

CODATA-RDA-DataScienceSchools. (2025) *Materials for schools of research data science. GitHub*. Available at: https://github.com/CODATA-RDA-DataScienceSchools/Materials

Davenport, T.H. and **Redman, T.C.** (2022) 'How AI is improving data management', *MIT Sloan Management Review*, 63(4), pp. 101–105. Available at: https://sloanreview.mit.edu/article/how-ai-is-improving-data-management/

Demchenko, Y. and **Stoy, H.** (2021) Research data management and data stewardship competences in university curriculum. *IEEE Global Engineering Education Conference (EDUCON)*. Available at: https://www.uazone.org/demch/papers/educon2021-data-stewardship-competence-fw-v02.pdf

Diggs, S. (2025) *UC3 New Year Series: Data publishing at CDL in 2025*. UC3 – California Digital Library. Available at: https://uc3.cdlib.org/2025/03/05/uc3-new-year-series-data-publishing-at-cdl-in-2025/

Doehle, P., Bjornen, K. and **Chartier, M.** (2019) *Promoting data literacy across campus with Carpentries: The experience of three librarians* [PowerPoint slides]. Edmon Low Library, Oklahoma State University. Available at: https://www.okacrl.org/wp-content/uploads/002-112_OSU_Carpentries.pptx

Gandhi, S.R. and **Anyiam, F.E.** (2022) 'Urban data science education: A key actor towards improving data-driven policy-making for solving urban problems', *Journal of Education, Society and Behavioural Science*, 35(5), pp. 1–14. Available at: https://doi.org/10.9734/jesbs/2022/v35i530421

Goben, A. and **Griffin, T.M.** (2019) 'In aggregate: Trends, needs, and opportunities from research data management surveys', *College & Research Libraries*, 80(5), pp. 643–663. Available at: https://doi.org/10.5860/crl.80.7.903

He, D. and **Wang, L.** (2023) 'Job analyses of earth science data managers: A survey validation of competencies to inform curricula in research data management education', *Journal of Education for Library and Information Science*, 64(2), pp. 104–119. Available at: https://doi.org/10.3138/jelis-2021-0023

International Centre for Theoretical Physics. (2025a) *The CODATA-RDA School for Research Data Science* (smr 4092). ICTP. Available at: https://indico.ictp.it/event/10857/

International Centre for Theoretical Physics. (2025b) *The CODATA-RDA Advanced Workshops for Research Data Science* (smr 4168). ICTP. Available at: https://indico.ictp.it/event/10990/

Jordan, K.L. (2018) *Evidence of Carpentries' impact on learners*. The Carpentries. Available at: https://carpentries.org/blog/2018/07/evidence-impact/

Jordan, K.L., Michonneau, F. and **Weaver, B.** (2018) *Analysis of Software and Data Carpentry's pre- and post-workshop surveys* (Version 1) [Assessment report]. The Carpentries. Available at: https://doi.org/10.5281/zenodo.1325464

Kanza, S. and **Knight, N.** (2021) *Failed it to nailed it: Responsible data management: Legal & ethical aspects* (AI3SD-Event-Series:Report-21). University of Southampton. Available at: https://doi.org/10.5258/SOTON/P0034

Kanza, S. and **Knight, N.** (2022) 'Behind every great research project is great data management', *BMC Research Notes*, 15, p. 20. Available at: https://doi.org/10.1186/s13104-022-05908-5

Kotsis, S.V. and **Chung, K.C.** (2013) 'Application of the "see one, do one, teach one" concept in surgical training', *Plastic and Reconstructive Surgery*, 131(5), pp. 1194–1201. Available at: https://doi.org/10.1097/PRS.0b013e318287a0b3

Krahe, M.A., Toohey, J., Wolski, M., Scuffham, P.A. and **Reilly, S.** (2020) 'Research data management in practice: Results from a cross-sectional survey of health and medical researchers', *Health Information Management Journal*, 49(2–3), pp. 108–116. Available at: https://doi.org/10.1177/1833358319831318

LIBER Research Data Management (RDM) Working Group. (2020) 'The 6 pillars of engaging researchers in research data management (RDM)', *LIBER (Ligue des Bibliothèques Européennes de Recherche – Association of European Research Libraries)*. Available at: https://libereurope.eu/wp-content/uploads/2020/12/The-6-Pillars-of-Engaging-Researchers-in-Research-Data-Management-RDM.pdf

Maienschein, J., MacCord, K. and **Elliott, S.** (2019) 'Help with data management for the novice and experienced alike', in G. Ramsey and A. De Block (eds.) *The dynamics of science*. Pittsburgh: University of Pittsburgh Press, pp. 123–140.

Majid, S., Foo, S. and **Zhang, X.** (2018) 'Research data management by academics and researchers: Perceptions, knowledge and practices', in L. Chen, Y. Liu and T.M.K.K.P. Ma (eds.) *Digital libraries and knowledge organization*. ICADL 2018. Lecture Notes in Computer Science, Vol 11282. Springer, pp. 160–175. Available at: https://doi.org/10.1007/978-3-030-04257-8_16

Mumuni, A.G. and **Mumuni, F.** (2024) 'Automated data processing and feature engineering for deep learning and big data applications: A survey', *Journal of Information and Intelligence*, 3(2), pp. 113–153. Available at: https://doi.org/10.1016/j.jiixd.2024.01.002

Oo, C.Z., Chew, A.W., Wong, A.L.H., Gladding, J. and **Stenstrom, C.** (2021) 'Delineating the successful features of research data management training: A systematic review', *International Journal for Academic Development*, 27(3), pp. 249–264. Available at: https://doi.org/10.1080/1360144X.2021.1898399

Quick, R. (2016) *Computational Infrastructures at CODATA-RDA Summer School in Research Data Science Aug 1–12 2016* (Version v1) [Lesson]. Zenodo. Available at: https://doi.org/10.5281/zenodo.154430

Quick, R., Córdoba, M.A., Cobe, R., Peterson, B., Shanahan, H., Costantini, A., EL-Sara, EL Jadid, S., sv1uk, abellew and **Bezuidenhout, L.** (2023) *CODATA-RDA-DataScienceSchools/Materials: Treiest2023* (Version v2023) [Software]. Zenodo. Available at: https://doi.org/10.5281/zenodo.8350033

Quick, R., Córdoba, M.A., Diggs, S., Cobe, R., Bezuidenhout, L., Shannahan, H. and **Peterson, B.** (2023) 'Foundational data science training for health equity researchers at minority serving institutions: A SoRDS event', *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*. Houston, TX, USA, 26–29 June 2023. IEEE, pp. 663–667. Available at: https://doi.org/10.1109/ICHI57859.2023.00115

Rantasaari, J. (2022) 'Multi-stakeholder research data management training as a tool to improve the quality, integrity, reliability and reproducibility of research', *LIBER Quarterly*, 32(1), pp. 1–54. Available at: https://doi.org/10.53377/lq.11726

Read, K., Larson, C., Gillespie, C., Oh, S.Y. and **Surkis, A.** (2019) 'A two-tiered curriculum to improve data management practices for researchers', *PLOS One*, 14(5), p. e0215509. Available at: https://doi.org/10.1371/JOURNAL.PONE.0215509

Schmidt, B. and **Shearer, K.** (2017) *The WHAT and the HOW of research data management: Towards a unified view of train-the-trainer competencies*. Digital Curation Centre. Available at: https://www.dcc.ac.uk/sites/default/files/documents/IDCC17~/80_How_Why_RDM.pdf

Shanahan, H., Harrison, A. and **May, S.T.** (2015) 'Teaching data science and cloud computing in low and middle income countries', *Advanced Techniques in Biology & Medicine*, 3(3), p. 150. Available at: https://doi.org/10.4172/2379-1764.1000150

Shanahan, H., Hoebelheinrich, N. and **Whyte, A.** (2021) 'Progress toward a comprehensive teaching approach to the FAIR data principles', *Patterns (N Y)*, 2(10), p. 100324. Available at: https://doi.org/10.1016/j.patter.2021.100324

Tachie, C.Y.E., Obiri-Ananey, D., Alfaro-Cordoba, M., Tawiah, N.A. and **Aryee, A.N.A.** (2024) 'Classification of oils and margarines by FTIR spectroscopy in tandem with machine learning', *Food Chemistry*, 431, p. 137077. Available at: https://doi.org/10.1016/j.foodchem.2023.137077

Tamm, H.C. and **Nikiforova, A.** (2025) From Data Quality for AI to AI for Data Quality: A Systematic Review of Tools for AI-Augmented Data Quality Management in Data Warehouses. Available at: https://arxiv.org/abs/2406.10940

University College Cork. (n.d.) *DH5001 – Digital Humanities Programme Overview*. Available at: https://www.ucc.ie/en/dh5001/

University of Vienna. (2022) *"Data Steward" certificate programme*. Research Data Management. University of Vienna. Available at: https://rdm.univie.ac.at/data-stewards-at-the-university/become-a-data-steward/

Wiley, C.A. and **Kerby, E.E.** (2018) 'Managing research data: Graduate student and postdoctoral researcher perspectives', *Issues in Science and Technology Librarianship*, 89, pp. 1–15. Available at: https://doi.org/10.29173/istl1725

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and **Mons, B.** (2016) 'The FAIR guiding principles for scientific data management and stewardship', *Scientific Data*, 3, p. 160018. Available at: https://doi.org/10.1038/sdata.2016.18

Yang, W., Fu, R., Amin, M.B. and **Kang, B.H.** (2025) 'The impact of modern AI in metadata management', *Human-Centric Intelligent Systems*, 5, pp. 323–350. Available at: https://doi.org/10.1007/s44230-025-00106-5

Yu, F., Deuble, R. and **Morgan, H.** (2017) 'Designing research data management services based on the research lifecycle – a consultative leadership approach', *Journal of the Australian Library and Information Association*, 66(3), pp. 287–298. Available at: https://doi.org/10.1080/24750158.2017.1364835

TO CITE THIS ARTICLE:
Gandhi, S., Diggs, S., Córdoba, M.A., Bezuidenhout, L., Cobe, R., El Jadid, S., Peterson, B., Quick, R., Shanahan, H., Venkataraman, S., Okorafor, E. and Van den Eynden, V. 2026 Building Responsible and Sustainable Open Data Literacy Skills for Early Career Researchers: A Decade of the SoRDS Programme. *Data Science Journal*, 25: 12, pp. 1–13. DOI: https://doi.org/10.5334/dsj-2026-012

Submitted: 15 August 2025
Accepted: 03 March 2026
Published: 19 March 2026

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