Article

ROLE OF LIBRARIES IN RESEARCH DATA CURATION

Arun Kumar

Jaipuria Institute of Management, Noida, Uttar Pradesh,
E-mail: arun.kumar@jaipuria.ac.in

ABSTRACT:

As the landscape of academic research becomes increasingly data-centric, the function of libraries is rapidly evolving from traditional archival roles to active stewardship of digital research data. This study explores how libraries are taking on critical responsibilities in the curation, preservation, and governance of research data throughout its lifecycle. Drawing on qualitative methods—literature reviews, case studies, and institutional surveys—this research investigates the practical, technological, ethical, and policy-related dimensions of research data curation in libraries. The findings emphasize how libraries support FAIR (Findable, Accessible, Interoperable, and Reusable) data practices through infrastructure, metadata management, legal compliance, and training. By integrating both manual and AI-assisted curation tools, libraries enhance data reuse, research integrity, and interdisciplinary collaboration. However, challenges such as scalability, privacy concerns, and resource limitations persist. The study highlights global and Indian frameworks guiding data stewardship and outlines best practices and future directions for libraries to fortify their role in a data-driven academic ecosystem.

Keywords: *Academic Libraries, Data Ethics, Data Governance, Data Stewardship, Digital Preservation, FAIR Principles, Library Technology, Metadata Standards, Open Access, Research Data Curation*

1. INTRODUCTION

The word "curation," comes from the Latin word "cure," which means "care for." The process of creating, arranging, and overseeing data collections so that users can access and utilise them is known as data curation. It comprises collecting, organising, indexing, and cataloguing data for a range of purposes, such as scientific research, academic study, and business decision-making. Data curation can be incorporated into data preparation tasks for Business Intelligence and analytics applications and is a component of the larger data management process. Data curation guarantees that datasets are comprehensive, thoroughly documented, and in a structure and format that facilitates long-term access, discovery, and reuse.

To guarantee research data's long-term accessibility, usability, and integrity, curation actively manages, maintains, and disseminates it throughout its whole lifecycle. Libraries are essential for maintaining the security, reusability, and accessibility of research data. Academic and research libraries are now major players in data stewardship, having broadened their duties beyond traditional cataloguing and archiving as data-driven research increases. In the digital age, research data has become crucial for scientific and educational advancement. Libraries have historically served as repositories of knowledge, but they now also manage research data, enabling researchers to effectively manage and

share datasets. Research data must be discoverable, accessible, interoperable, and reusable (FAIR principles), and libraries are being held more and more responsible for this.

Research data curation actively manages, preserves, and distributes research data over its entire lifecycle to ensure its long-term accessibility, usability, and integrity. Libraries are critical in keeping research data accessible, reusable, and secure. As data-driven research grows, academic and research libraries have expanded their responsibilities beyond traditional cataloging and archiving, becoming major players in data stewardship. Research data has become invaluable for educational and scientific growth in the digital age. Libraries have traditionally acted as knowledge archives, but their role has evolved to encompass research data management, allowing researchers to manage and exchange datasets successfully. Libraries are increasingly accountable for ensuring that research data is discoverable, accessible, interoperable, and reusable (FAIR principles). Curating research data has benefits beyond preservation and storage. Maintaining data properly ensures compliance with funding agencies and institutional standards, promotes interdisciplinary collaboration, and increases research repeatability. To optimise the advantages of research data curation, however, challenges like exorbitant costs, data privacy issues, technological hurdles, and resistance to data sharing must be resolved.

The purpose of this paper is to examine and discuss how libraries' roles in organising and conserving research data are changing. Libraries now actively curate research data in addition to their traditional function of archiving information, as the significance of data-driven research grows. Important facets of research data curation are examined in this study, such as data management techniques, technology developments, ethical issues, and policy frameworks. In order to shed light on how libraries can improve their data stewardship initiatives, it also looks at national and international best practices.

2. LITERATURE REVIEW

The changing role of libraries in research data curation is highlighted in recent literature, which also highlights enduring difficulties as well as new trends. Recent academic papers and reports from 2022 to 2025 are summarised in this review.

AI Multiple Research (2025) investigates the difficulties and solutions associated with AI data governance with an emphasis on scalability concerns, data quality, and compliance. It highlights how crucial strong governance frameworks are to managing AI data in an ethical manner.

IBM Insights (2025) examines how AI ethics and governance are developing, emphasising the value of AI literacy and the necessity of all-encompassing frameworks to handle moral dilemmas in AI implementation.

Norman Mooradian, Patricia C. Franks, and Amitabh Srivastav (2025) discuss how AI intensifies data privacy issues due to its dependency on large datasets and ability to generate new personal information. It also looks at new privacy-enhancing techniques and how well they fit with current information governance and privacy guidelines.

ACRP White Paper (2025) assists the clinical research community comprehend the implications and use of AI in a rapidly changing environment. It directs ethical considerations and responsible oversight when applying AI to clinical research.

Samuel Schuster et al. (2024) discuss the ethical implications of integrating artificial intelligence (AI) into scientific research. They address new and complicated ethical problems that come up while highlighting the advantages of AI and stressing the necessity for the scientific community to create new standards for its proper application.

Manish Kumar Singh and Gireesh Kumar T.K. (2023) examine the the difficulties academic institutions have in gathering research data, emphasising problems like the lack of infrastructure, the creation of policies, and the requirement for qualified staff. The authors also go over the expectations of stakeholders and offer ways to improve data curation procedures.

Samuel Schuster et al. (2023) delve into the ethical implications of integrating artificial intelligence (AI) into scientific research. It talks about new ethical issues and possible advantages, highlighting the necessity for the scientific community to create new standards for the proper application of AI in research.

Frontiers in Artificial Intelligence (2023) explores the special ethical problems that AI has brought to research are examined, including questions of accountability, responsibility, and the suitability of current ethical standards. To properly address these issues, ethical training and public education are crucial.

The OCLC Research Library Partnership and LIBER (2023) report highlights the opportunities and challenges faced by academic libraries while facilitating a conversation on research data management. It highlights how crucial capacity building and strategic cooperation are to improving RDM services

Rebecca Bryant, Brian Lavoie, and Amanda K. Rinehart (2023). investigates the potential for data sharing services to be expanded through multi-institutional collaborations. The authors analyse cooperative research data management initiatives and offer insights into workable tactics for increasing RDM capacity through in-depth interviews.

Lisa R. Johnston (2016). The need for data curation services in academic and disciplinary data repositories is examined in this volume, which also offers helpful implementation tips for efficient data curation procedures.

Sheila Corrall (2012) highlights the changing roles of information professionals in data curation and sharing while reviewing the opportunities and difficulties that libraries and librarians encounter when managing research data.

Xu et al. (2019) examine academic libraries' practices, challenges, and perspectives on managing research data. They emphasise training, policy formation, sustainability, and researcher involvement.

Samuel Schuster et al. (2023) discuss the ethical implications of integrating AI into scientific research, including privacy, data ownership, and the potential biases introduced by AI systems.

Yuhao Jiang et al. (2022) provide a comprehensive overview that addresses the ethical issues associated with AI, including data privacy, algorithmic bias, and the societal impacts of AI deployment.

Whittlestone et al. (2019) report provides a roadmap for research into the ethical and societal challenges posed by algorithms, data practices, and AI technologies.

Lisa R. Johnston (2016) explores the variety of reasons, motivations, and drivers for why data curation services are needed in the context of academic and disciplinary data repository efforts.

Despite technological advancements, limited research exists on the long-term sustainability of AI-driven data curation. Future studies should explore cost-effective automation solutions, methods for mitigating bias in AI-generated metadata, and practical machine-learning applications in large-scale repositories. While existing literature addresses AI governance frameworks, few studies examine how institutions implement these frameworks in real-world research data curation. Further research should assess institutional compliance with AI ethics policies and investigate case studies of bias detection in AI-managed research repositories.

Additionally, the current literature lacks empirical studies on policy-driven incentives that encourage data sharing. Future research should evaluate the impact of data-sharing mandates issued by funding agencies (e.g., NIH, NSF, Horizon Europe) and assess the effectiveness of institutional strategies in fostering a culture of data sharing. Comparative research on global data governance frameworks is also limited. Future studies should examine how various countries balance data privacy, security, and open science principles to inform the development of harmonized international research data policies.

These research directions highlight the dynamic and evolving landscape of Research Data Curation in academic libraries. They emphasize the importance of ongoing professional development, the adoption of standardized practices, and the implementation of robust support services to meet researchers' changing needs.

3. OBJECTIVES

1. To investigate the function of libraries in the curation of research data
2. To examine the main facets of curating research data
3. To investigate the advantages (pros and cons) and difficulties of curating research data.
4. To determine the technologies and tools utilised in the curation of research data
5. To investigate legal and ethical issues in the curation of research data

4. RESEARCH METHODOLOGY

This study employs a qualitative, exploratory research design to examine libraries' evolving role in research data curation. Such an approach is suitable for investigating complex, emerging practices and capturing rich contextual detail. The methodology integrates multiple qualitative techniques. First, a systematic literature review was conducted following established scholarly protocols. This review collected peer-reviewed articles, professional reports, and policy documents from key bibliographic databases. The

extracted sources were coded thematically to map prevailing concepts (for example, data management policies, repository services, and metadata practices) and to establish a conceptual foundation for understanding library data curation. In parallel, case study analysis was undertaken on selected academic libraries and consortia that exemplify data curation efforts. Cases were chosen purposively to represent diverse institutional contexts. Data for each case were gathered through analysis of institutional documents (such as service plans and technical reports) and, where possible, through semi-structured interviews with library staff involved in data services. This qualitative material was used to construct detailed narratives of how libraries implemented curation services in context. The case study approach is justified by its ability to capture "real world" collaboration and innovation in a new and rapidly growing area such as research data management. Additionally, the study incorporated institutional surveys and interview data to gauge broader trends. For example, existing survey results from academic library networks were examined, and new questionnaires were distributed to data management librarians. These instruments solicited information on service offerings, challenges, and attitudes toward FAIR and open-data practices. Qualitative responses from surveys and interviews were analyzed using thematic content analysis to identify common patterns and divergent practices. All collected data (literature, case documentation, survey responses) were analyzed iteratively. Thematic coding was applied to identify key dimensions of data curation (e.g. infrastructure, skills, policy), and findings were triangulated across sources to enhance validity. Quantitative elements (such as frequencies from survey questions) were summarized descriptively to supplement the qualitative picture. Ethical research practices were observed (informed consent and confidentiality in interviews), and the analysis was aligned with the study's objectives of understanding how libraries contribute to data stewardship. Together, these methods provide a comprehensive, context-rich basis for answering the research questions

5. SCOPE AND LIMITATIONS OF THE STUDY

The study explores the role of academic libraries in stewarding research data, focusing on data curation practices. It examines how libraries organize, preserve, and make research data accessible, focusing on FAIR principles, metadata standards, and repository infrastructures. The research is based on institutional case examples from India and selected international contexts, offering comparative insights into national and global approaches to data curation. Key areas addressed include institutional repositories, metadata creation, legal compliance, technological tools, and librarians' role in data literacy. However, the study has limitations, including limited access to internal documentation, potential regional applicability due to disparities in infrastructure, funding, and policy enforcement, and the use of a qualitative lens.

6. ROLE OF LIBRARIES IN RESEARCH DATA CURATION:

Libraries are critical in supporting research data curation, ensuring that research outputs are well-organized, preserved, and accessible. They help researchers to develop data management plans, use metadata standards, and manage institutional repositories. They also support open data and FAIR principles to ensure compliance with open-access mandates. Libraries manage research data repositories, provide legal and ethical guidance,

and offer data literacy training. They act as intermediaries between the different aspects of research- funders, researchers, and publishers- helping to promote interdisciplinary collaborations and ensure that data is interoperable across different areas. As research increasingly becomes data-driven, the role of libraries in data management and digital preservation is increasing.

7. KEY ASPECTS OF RESEARCH DATA CURATION:

Data management planning is the process of organizing, storing, and sharing data, generating metadata, providing long-term storage, sharing and accessing data, following legal and ethical compliance, and quality-assuring data to preserve scientific integrity. Table 1 contains a list and description of the key components.

Table 1: Key Aspects and Description

Aspect	Description
Data Management Planning (DMP)	Developing a Data Management Plan (DMP) to outline data collection, storage, sharing, and compliance with funding policies.
Data Organization and Structuring	Implementing standardized file naming conventions and folder structures, and using consistent data formats for easy retrieval.
Metadata Creation and Documentation	Applying metadata standards (e.g., Dublin Core, DataCite) improves data discoverability and provides contextual information.
Data Storage and Security	Using secure and scalable storage solutions, implementing backup strategies, and ensuring compliance with privacy regulations.
Data Sharing and Access	Depositing data in repositories (e.g., Zenodo, Dryad), assigning persistent identifiers (DOIs), and applying open access principles.
Data Preservation and Sustainability	Using long-term preservation strategies, ensuring data remains accessible despite technological changes, and following FAIR principles.
Data Quality Control and Validation	Implementing data cleaning and validation processes to ensure accuracy, consistency, and completeness.
Legal and Ethical Considerations	Addressing copyright, intellectual property rights, licensing (e.g., Creative Commons), and ethically managing sensitive data.
Data Reuse and Interoperability	Ensuring data integration across disciplines, making data machine-readable, and promoting citation and attribution practices.
Training and Support	Providing data literacy training, assisting with data management tools, and supporting compliance with institutional and funding policies.

8. KEY BENEFITS OF RESEARCH DATA CURATION:

Curation of Research Data is a crucial practice that ensures the accessibility, trustworthiness, and impact of research data. It helps researchers maintain the integrity of their data, protect against data loss and corruption, and encourage reproducibility of research results. Curation also entails sustainable stewardship of data, legal and ethical responsibilities, increased visibility and citation impact, and addressing duplication of efforts. It also supports interdisciplinary research and advances in scientific knowledge using data mining techniques, AI, and machine learning practices. By harnessing the benefits of data curation, researchers can overcome financial, technical, and strategic planning hurdles, fostering innovation and discoveries in research.

9. MERITS AND DEMERITS OF RESEARCH DATA CURATION

Research Data Curation effectively utilizes data curation benefits while overcoming financial, technical, and strategic planning challenges. Overcoming these through policy frameworks, funding, and advanced technology can maximize its potential. The merits and Demerits of Research Data Curation are given below in Table 2.

Table 2: Merits and Demerits of Research Data Curation

Merits (Advantages)	Demerits (Challenges/Limitations)
Enhanced Accessibility & Reusability – Ensures data is findable, accessible, and reusable for future research.	High Costs & Resource Demands—This requires significant financial investment in storage, infrastructure, and skilled personnel.
Improved Research Integrity & Reproducibility – Facilitates validation and verification of research findings.	Time-Consuming Process – Metadata creation and quality checks require additional time and effort from researchers.
Long-Term Data Preservation – Protects research data from loss, degradation, or obsolescence.	Data Privacy & Security Risks – Sensitive data may pose legal and ethical concerns (GDPR, HIPAA compliance).
Compliance with Legal & Ethical Standards – Helps researchers meet funding agency requirements (NIH, NSF, Horizon Europe).	Technical & Standardization Challenges – A Lack of uniform metadata standards can create interoperability issues.
Increased Research Visibility & Citation Impact – Well-curated data attracts citations, enhancing academic reputation.	Resistance to Data Sharing—Some researchers hesitate to share data due to concerns about competition, misuse, or lack of attribution.
Cost-Effective & Efficient Data Management – Reduces redundancy and duplicate research efforts, saving time and resources.	Scalability Issues – Managing large datasets (big data) requires advanced storage and processing capabilities.
Facilitates Innovation & Interdisciplinary Research – Supports AI, machine learning, and cross-disciplinary collaborations.	Sustainability & Long-Term Commitment – Requires continuous funding and institutional support to maintain data repositories.

10. TOOLS & TECHNOLOGIES FOR RESEARCH DATA CURATION

Research data curation involves specialized tools and technologies, such as repository platforms, metadata management, AI-driven analytics, and long-term digital preservation, to ensure data integrity, accessibility, and interoperability across disciplines. **Data Repositories & Management Platforms** are described in Table 3.

11. DATA REPOSITORIES & MANAGEMENT PLATFORMS

Research data storage, management, and dissemination platforms like Dataverse, Figshare, Zenodo, DSpace, Hydra/Fedora, and CKAN are essential for preserving and sharing data. These platforms support metadata tagging, version control, access rights, and long-term preservation, making them popular among academic institutions, governments, and research consortia.

Table 3: Data Repositories & Management Platforms

Tool/ Technology	Description	Key Features
Harvard Dataverse	Open-source repository for storing and sharing research data.	Metadata integration
		Version control & access permissions
Dryad	A curated repository for publishing and preserving research data.	DOI assignment
		Supports data sharing in open science
Zenodo	Open-access data repository developed by CERN.	GitHub integration
		Long-term archiving
figshare	Cloud-based repository for researchers to store, share, and manage datasets.	DOI generation
		Supports multimedia research data
DANS Easy (Netherlands)	Dutch repository for long-term access to research data.	Complies with FAIR principles
		Multi-disciplinary data storage
National Digital Library of India (NDLI)	A central repository for academic and research data in India.	Provides access to Indian scholarly content
		Supports multilingual search
Shodhganga	Digital repository for Indian Ph.D. theses and research data.	Open-access thesis storage
		Supports Indian research institutions

12. METADATA MANAGEMENT & ANNOTATION TOOLS

Metadata is crucial for research data discovery and reuse, and tools like Open Refine, Mediata, and ISA Tools Suite help adhere to discipline-specific standards. Details given in Table 4.

Table 4: Metadata Management & Annotation Tools

Tool/Technology	Description	Key Features
Dublin Core Metadata Initiative (DCMI)	A widely used metadata standard for organizing research data.	Standardized metadata elements
		Enhances data discoverability
DataCite	Provides DOI registration and metadata schema for datasets.	Persistent identifier assignment
		Enables dataset citations
RO-Crate	FAIR-compliant metadata packaging for research data.	Machine-readable metadata
		Portable and reusable data format
Open AIRE Research Graph	A metadata management tool for linking research outputs.	Provides interoperability between repositories
		Supports open-access compliance

13. LONG-TERM DIGITAL PRESERVATION TECHNOLOGIES

Long-term access to data entails Robust preservation systems, such as Preservica, Archivematica, and LOCKSS, which are required for long-term data access. These systems provide format migration, integrity checks, and archive standards. Details are given in Table 5.

Table 5: Long-Term Digital Preservation Technologies

Tool/Technology	Description	Key Features
LOCKSS (Lots of Copies Keep Stuff Safe)	A digital preservation system ensures multiple copies of research data.	Distributed preservation
		Self-healing technology
Portico	A digital archive for preserving scholarly literature and datasets.	Long-term accessibility
		Secure cloud storage
Archivematica	Open-source digital preservation system.	Supports multiple file formats
		Ensures metadata integrity
DSpace	Open-source software for managing research data repositories.	Used by many universities
		Customizable metadata schemas
EPrints	Repository management software for long-term archiving.	User-friendly interface
		Supports research paper and data submission

14. DATA CLEANING & PROCESSING TOOLS

Data cleaning and processing are key procedures in research data curation that ensure accuracy, consistency, completeness, and dependability. These programs help find and fix errors, remove duplicates, standardise formats, and prepare data for analysis, sharing, and archiving. Table 6 provides the full description.

Table 6: Data Cleaning & Processing Tools

Tool/Technology	Description	Key Features
OpenRefine	A powerful data-cleaning tool used for structuring messy data.	Data transformation & normalization
		Supports multiple formats (CSV, JSON, XML)
Trifacta Wrangler	AI-powered data preparation tool.	Detects anomalies in datasets
		Supports large-scale data wrangling
KNIME	Open-source data analytics and processing platform.	Graphical workflow interface
		Supports machine learning integration
Pandas (Python Library)	Data analysis and cleaning tool for structured data.	Large-scale data processing
		Works with CSV, Excel, and SQL databases

15. DATA SHARING & INTEROPERABILITY TOOLS

Data sharing and interoperability tools are an important part of open science. The FAIR data principles regard data sharing and interoperability as a key focus of researchers, making datasets openly available while ensuring accessibility, integration, and reuse across platforms and disciplines. Table 7 provides the full description.

Table 7: Data Sharing & Interoperability Tools

Tool/Technology	Description	Key Features
CKAN	Open-source data portal for publishing open-access research data.	API for data sharing
		Data visualization support
Research Data Management Toolkit (RDMKit)	Supports European Open Science Cloud (EOSC) standards for research data sharing.	Provides guidelines for data governance
		Enhances interoperability
Global Biodiversity Information Facility (GBIF)	A worldwide network for sharing biodiversity research data.	Enables cross-institutional data integration
		Open-access biodiversity datasets
ISRO's Bhuvan Geoportal	A platform for accessing and sharing Indian geospatial and satellite data.	Disaster management support
		Multi-resolution data access

16. AI & MACHINE LEARNING FOR DATA CURATION

Artificial intelligence and machine learning are transforming Research Data Curation with automated operations such as information extraction and classification, anomaly detection, and semantic enrichment. These automated techniques scale operational improvements while ensuring our management of vast and heterogeneous datasets becomes more accurate, reliable, and efficient. They can also intelligently suggest the next course of action regarding sharing and archiving. Table 8 provides the full description.

Table 8: AI & Machine Learning for Data Curation

Tool/Technology	Description	Key Features
IBM Watson Knowledge Catalog	An AI-driven metadata management and data governance tool.	Automates metadata classification
		Ensures compliance with data policies
Google Cloud AutoML	An AI-powered tool for automated research data annotation.	Supports NLP & image classification
		Enables scalable data curation
TensorFlow Data Validation (TFDV)	A machine learning tool for automated data quality checks.	Detects anomalies in research data
		Supports big data applications

17. ETHICAL AND LEGAL CONSIDERATIONS

Data curation has ethical and legal implications that differ according to regional laws, research fields, and data sensitivity. Ethical and legal considerations allow for privacy, equity, transparency, and informed consent. See Table 9 for details.

Table 9: Ethical and Legal Considerations

Consideration	Ethical Issues	Legal Issues	Indian Context	Global Context
Informed Consent	Ensuring individuals are fully informed about data usage, storage, and sharing.	Legal requirement to obtain explicit consent for data collection and sharing.	ICMR and IRBs emphasize informed consent for health and biomedical research.	The GDPR requires explicit consent for personal data collection.
	Participants should voluntarily consent to data collection and use.	Must comply with national and international privacy laws.	PDPB will reinforce this.	HIPAA also demands informed consent for health data.
Data Anonymization	Data should be anonymized to	Legal obligations to	PDPB requires	The GDPR mandates the

& Privacy	protect individual privacy.	anonymize personal data to protect privacy.	anonymization to protect privacy.	anonymization of personal data to ensure privacy.
	Ethical responsibility to prevent misuse of personal data.	Violations may lead to fines or penalties.	Focus on protecting sensitive health and biometric data.	HIPAA focuses on the privacy of health data.
Fairness & Bias in Data	Data curation should avoid bias and ensure diversity in datasets.	Legal consequences for discriminatory data practices (e.g., biases in algorithms).	Ethical concerns over biased representation in demographic and social research.	OECD Guidelines on AI emphasize fairness in data use.
	Must not perpetuate discrimination or exclusion of specific groups.	To prevent inequality, fairness in data collection and processing must be ensured.	Ensuring diverse population representation.	GDPR and CCPA stress non-discriminatory practices.
Accountability & Transparency	Curators should be transparent about how data is collected, stored, and used.	Legal requirements for transparency in data handling and processing.	PDPB encourages transparency in how personal data is handled.	The GDPR mandates transparency in data processing, including the purpose and scope of data collection.
	Ethical responsibility to ensure data practices are traceable and transparent to stakeholders.	Data curators must disclose data use and ensure public access to information on data handling.	India's evolving framework for data governance.	
Data Protection & Privacy Laws	Researchers must ensure that data is kept secure and handled carefully,	Strict laws for data protection include GDPR, HIPAA, and	PDPB will set guidelines for handling personal data.	The GDPR sets stringent guidelines for the protection of personal data.

	especially sensitive data.	PDPB.		
	Ensuring privacy for vulnerable populations (e.g., children, marginalized groups).	Breaches can result in severe fines and penalties.	Data security requirements for sensitive personal information.	HIPAA regulates medical data security.
Cross-Border Data Transfers	Ethical obligation to ensure that personal data is securely handled when shared across borders.	Legal restrictions on data transfer to countries without adequate data protection laws.	PDPB restricts cross-border transfer of sensitive data unless the destination country has adequate protection laws.	The GDPR restricts data transfer outside the EU to countries with similar protection laws.
	Protecting privacy rights when transferring data internationally.	International agreements (e.g., EU-U.S. Privacy Shield) govern cross-border data transfers.		CCPA governs such transfers.
Intellectual Property Rights	Ethical responsibility to respect intellectual property rights in data creation and use.	Legal protections for intellectual property rights, including research data ownership and licensing.	The Indian Copyright Act regulates the ownership of research data.	Global use of Creative Commons licenses to protect intellectual property.
	Proper attribution and licensing of datasets.	Data creators must ensure appropriate attribution and licensing agreements.	Use of Creative Commons licenses for open data.	Copyright laws for data protection.

18. FRAMEWORKS & GUIDELINES FOR RESEARCH DATA CURATION

International and Indian frameworks focus on research data curation by establishing an overarching role for standardizing research data management, open access provision, metadata facilitation, and ethical data sharing. Frameworks strengthen their impact on research, facilitate interdisciplinary science, and improve data-driven decisions.

19.INTERNATIONAL FRAMEWORKS & GUIDELINES

Research Data curation is based on global standards and frameworks, ensuring the data is Findable, Accessible, Interoperable, and Reusable (FAIR) and exhibits ethical, secure, and long-term data management practices. Frameworks establish organized ways to curate data stewardship throughout the research lifecycle. See Table 10 for details.

Table 10: International Frameworks & Guidelines

Framework / Guideline	Description	Key Principles & Impact
FAIR Data Principles (Findable, Accessible, Interoperable, Reusable)	Developed by the GO FAIR Initiative, this global standard ensures that research data is structured for easy access, reuse, and interoperability.	Data should be machine-readable and well-documented.
		Encourages metadata standards & open-access repositories.
OECD Principles & Guidelines for Access to Research Data	Established by the Organisation for Economic Co-operation and Development (OECD) to promote global data sharing in research.	Advocates for open data access while balancing privacy and security concerns.
		Encourages international collaboration.
DataCite Metadata Schema	A metadata framework designed to facilitate the citation, discovery, and reuse of research datasets.	Ensures persistent identifiers (DOI) for datasets.
		Enables cross-disciplinary research.
Dublin Core Metadata Initiative (DCMI)	A widely used metadata standard for data classification and retrieval in research institutions and digital libraries.	Provides standardized metadata elements for categorizing research data.
		Improves data discoverability.
ISO 16363 – Audit and Certification of Trustworthy Digital Repositories	A framework for evaluating digital repositories that store research data to ensure long-term preservation and security.	Ensures data authenticity & integrity.
		Establish best practices for repository management.
WDS (World Data System) Certification	Developed by the International Science Council (ISC) to certify trusted data repositories.	Promotes sustainable data management & ethical sharing.
		Encourages high-quality, reliable data storage.
CODATA Data Science Framework	The Committee on Data (CODATA) under ISC provides guidelines for data-intensive research.	Supports big data analytics & AI-driven research curation.
		Encourages global research data interoperability.

Research Data Alliance (RDA) Recommendations	The RDA develops best practices and recommendations for data sharing and infrastructure across research fields.	Promotes interoperable and standardized data-sharing frameworks.
		Connects global research communities.

20. NATIONAL FRAMEWORKS & GUIDELINES

India has acknowledged the significance of data curation as part of a bigger scientific endeavor, policy development, and innovation. Numerous regulations and institutional initiatives have been initiated to assist in managing and sharing data, along with additional details in Table 11.

Table 11: National Frameworks & Guidelines

Framework / Guideline	Description	Key Principles & Impact
National Data Sharing and Accessibility Policy (NDSAP)	Launched by the Government of India to promote open-access government data.	Encourages public access to non-sensitive datasets.
		Supports evidence-based policymaking.
Open Government Data (OGD) Platform	A government initiative that provides curated research and administrative data for public use.	Facilitates transparent governance & research collaboration.
		Enables AI & analytics applications.
ICAR Research Data Management Policy	The framework of the Indian Council of Agricultural Research (ICAR) is to standardize agricultural research data curation.	Supports climate-resilient crop development.
		Ensures metadata compliance for agri-data repositories.
CSIR Open Access Policy	A policy by the Council of Scientific & Industrial Research (CSIR) to provide free access to publicly funded research data.	Promotes open-access publication and data repositories.
		Strengthens global collaborations in science & technology.
IISER Data Management & Preservation Policy	Research data guidelines by the Indian Institutes of Science Education and Research (IISERs).	Ensures long-term storage and accessibility of research datasets.
		Encourages inter-institutional research collaborations.
Indian Space Research Organisation (ISRO) Data Sharing Policy	Defines protocols for sharing satellite imagery & space research data via platforms like Bhuvan & NRSC.	Supports climate monitoring, disaster response, and geospatial analysis.
		Enables multi-institutional access to space research data.
UGC Guidelines on Research Data	Issued by the University Grants Commission (UGC)	Establishes data-sharing ethics and intellectual property

Management	to ensure structured research data curation in Indian universities.	guidelines.
		Supports theses & dissertations curation in digital repositories.
Digital Library Initiatives (NDLI & Shodhganga)	National Digital Library of India (NDLI) and Shodhganga manage academic research data and theses.	Preserves Indian research outputs for global access.
		Promotes open-access education.

21. COMPARISON BETWEEN MANUAL VS. AUTOMATED DATA CURATION METHODS

Data curation can be undertaken manually or with automated tools, and the best method will depend on the volume, complexity, depth of quality, and researcher resources.

Key Takeaways from the Comparative Analysis:

1. Manual Data Curation is more suited for less complex or smaller-scale data, requiring an individual to bring pre-existing knowledge and context to interpret nuanced judgment or interpretation, or when working with unstructured data. Although it delivers a high degree of case-by-case accuracy and flexibility, it is slower, costly, and cannot be duplicated at scale.

2. Automated Data Curation is suited for extensive data and repeated cyclical tasks needing speed, consistency, and scalability. Automated Curation struggles with complexity and unstructured data and lacks judgment/contextual knowledge, which may be key to the curation decisions made by the human curator. Automated solutions deliver cost-effective options to large research teams on projects needing large amounts of data processing.

When to Use Which?

- Manual Curation: Choose this method when the dataset is small, complex, or qualitative, or when you need the human touch for interpretation and context.

- Automated Curation: Suitable for the continued processing of large amounts of data or the efficiency of tasks requiring duplicated processes over time, particularly with speed and accuracy.

Table 12: Manual & Automated Data Curation Techniques

Aspect	Manual Data Curation	Automated Data Curation
Process	Human-centric, involving manual data selection, cleaning, organization, and storage.	Software-driven, utilizing scripts, machine learning models, and algorithms for automatic data processing, cleaning, and storage.
Speed	Slow, especially with large datasets or complex operations.	Fast, can process large datasets in a fraction of the time.

Accuracy	Highly accurate with correct human oversight. Susceptible to human errors like inconsistency or bias.	Accuracy varies based on the algorithm used. More prone to errors in complex or unstructured data unless properly tuned.
Cost	High cost in terms of labor, as it requires skilled personnel for data management.	Lower cost in terms of labor, but may require investment in tools and technologies.
Scalability	Difficult to scale, as the process needs to be repeated for each dataset manually.	Highly scalable, able to handle large datasets and complex workflows without significant human intervention.
Flexibility	Highly flexible, human curators can easily adapt to contextual nuances in the data and change processes.	Limited flexibility unless manually adjusted or trained for new conditions or data types.
Consistency	Inconsistently, the quality may vary depending on the curator's experience, knowledge, and time.	Consistent, as automated systems follow pre-defined rules or models without variation.
Data Complexity	Suitable for complex and highly contextual data, which requires human interpretation, especially in qualitative research.	Struggles with highly complex or unstructured data unless AI and advanced algorithms are specifically trained.
Error Detection	Manual error detection requires attention to detail, which can lead to oversight in large datasets.	Automated error detection is faster but might overlook context or misclassify errors without perfect algorithms.
Metadata & Documentation	Manual metadata entry can be labor-intensive and inconsistent, but can add rich contextual details.	Automated metadata generation is faster but may lack contextual richness or be incorrect if the system is not sophisticated enough.
Data Provenance & Traceability	The curator documents excellent traceability at each step in the curation process.	Traceability depends on the system's logging and documentation capabilities, but can be lacking in complex workflows.
Maintenance	Requires continuous oversight to keep the curation process aligned with evolving data standards.	Once systems are set up, minimal human oversight is required, but occasional updates and re-training may be needed.
Data Access & Sharing	Data sharing may be slower, as data often must be manually reviewed before sharing.	Faster sharing, as data can be processed and uploaded into repositories automatically, especially for open-access

	platforms.	
Quality Control	High-quality control, as the curator can manually check and ensure data meets specific standards.	Quality control is algorithm-dependent and may require periodic validation and adjustment of the system.
Use Cases	Ideal for small datasets with highly complex, sensitive, or nuanced data. Common in fields like qualitative research and art history.	Best suited for large-scale datasets, routine processing tasks, or datasets with structured data (e.g., quantitative research, large-scale surveys).

Case Studies and Examples of Research Data Curation: National and International Context

Several universities, research institutes, and government agencies worldwide have exemplified research data curation to enhance scientific collaboration, accessibility, and preservation.

Table 13: Case Studies & Examples of Research Data Curation – Indian & Global Context

S.No	Institution / Initiative	Region	Overview	Impact & Success
1	ICAR – National Bureau of Plant Genetic Resources (NBPGR)	National	Curates and preserves agri-biodiversity & genetic resources for sustainable agriculture.	Improved food security by developing climate-resilient crops.
				Facilitates collaboration among agricultural scientists.
2	ISRO – National Remote Sensing Centre (NRSC) & Bhuvan Geoportal	National	Stores and curates satellite imagery and geospatial data for environmental and urban research.	Supports disaster management and climate research.
				Aids in thoughtful city planning and rural development.
3	Indian Statistical Institute (ISI) – Census & Socio-Economic Data	National	Manages and curates large-scale economic and social datasets (Census, NSS, Economic	Enables data-driven policymaking in healthcare, employment, and social welfare.
				Supports

			Surveys).	interdisciplinary research.
4	National Digital Library of India (NDLI) – Academic Data Repository	National	An open-access digital library developed by IIT Kharagpur provides access to Indian and global research data.	Ensures long-term preservation of Indian scholarly work.
				Enhances education accessibility.
5	CSIR – Open Access Research Data in Science & Technology	National	Maintains scientific research data across chemistry, biotechnology, and pharmaceuticals.	Boosts collaborative research between scientists and industries.
				Supports drug discovery & healthcare innovation.
6	IMD – Meteorological Data Curation	National	Curates climate & weather data for monsoon forecasting and disaster management.	Supports agriculture planning and disaster preparedness.
				Contributes to global climate change research.
7	National Data Sharing and Accessibility Policy (NDSAP) – Open Government Data (OGD) Platform	National	Provides open-access government data across various sectors.	Promotes transparent governance & evidence-based decision-making.
				Encourages AI & startup innovation.
8	Harvard Dataverse (USA)	International	Open-source repository for storing and sharing research data across disciplines.	Facilitates global data sharing & research collaboration.
				Used by universities worldwide.
9	UK Data Service (UK)	International	Curates social and economic datasets, including census and policy	Supports government policymaking & academic research.
				Encourages

			research data.	interdisciplinary collaboration.
10	NASA Earth Science Data Systems (USA)	International	Manages climate & satellite data for scientific research.	Enables climate change analysis & disaster response.
				Supports global environmental monitoring.
11	European Open Science Cloud (EOSC – EU)	International	Federated cloud infrastructure ensures FAIR data principles (Findable, Accessible, Interoperable, Reusable).	Encourages open science & data sharing.
				Connects researchers across Europe.
12	Human Genome Project (Global)	International	Curated genomic data to map the entire human genome.	Transformed genomic research & precision medicine.
				Facilitates global biomedical research collaborations.
13	CERN's Open Data Portal (Switzerland)	International	Stores Large Hadron Collider (LHC) experimental data for physics research.	Contributed to the discovery of the Higgs Boson.
				Enables open-access physics research worldwide.
14	Australian Research Data Commons (ARDC – Australia)	International	National research data infrastructure for scientific data management.	Boosts research efficiency & interdisciplinary collaboration.
				Supports big data & AI applications.
15	Borealis Data Repository (Canada)	International	National open-access research data repository connecting universities.	Supports cross-institutional collaboration.
				Ensures long-term research data preservation.

22. CONCLUSION

Research data curation is essential for ensuring the accessibility, integrity, and use of research outputs in the long term. Libraries have transitioned from static knowledge

holders to active research data management participants. Libraries are well-positioned to aid researchers with data curation practices through proper tools such as metadata, digital preservation, and legal obligations about research data. Libraries can align with researchers on the practices they already have in place to support the goals of FAIR (Findable, Accessible, Interoperable, and Reusable) outputs. There may be costs associated with user requests, the potential for data privacy issues, and technological challenges with using data curation practices. However, libraries can position themselves to enhance data stewardship and knowledge sharing using policies, funding, and institutional support.

23. SUGGESTIONS & FURTHER STUDIES:

To increase libraries' roles in research data curation, institutions should invest in capacity building by training librarians in data management and emerging technologies. All disciplines should adopt standardised procedures for maintaining data curation consistency. Libraries should establish enduring joint efforts with research funding bodies and technology platform owners to develop sustainable data bank solutions enabling information preservation and knowledge sharing. Furthermore, artificial intelligence and automation approaches can improve the efficiency and accuracy of data curation processes. Researchers need policy incentives with awareness campaigns to develop habits of open data sharing for establishing a collaborative academic environment.

Future studies should investigate the role of artificial intelligence and machine learning in automating data curation and improving metadata management. Comparative studies examining the efficacy of various research data curation techniques across disciplines would yield significant insights into best practices. Research on the ethical and legal issues surrounding cross-border data sharing can assist organisations in navigating complex regulatory regimes. Case studies of successful national and international research data curation programs provide valuable insights for libraries wishing to improve their data management methods. Finally, additional studies on user involvement and academics' perspectives on data curation services can assist libraries in tailoring their support mechanisms to meet the demands of the academic community better.

REFERENCES

Agarwal, A. & Singh, N. (2016) 'Ethical implications of data sharing in research', *International Journal of Ethics in Data Science*, 4(1), pp. 34–46.

Akers, K.G. & Doty, J. (2013) *A study of faculty data curation behaviors and attitudes at a U.S. research university*, *College & Research Libraries*, 74(5), pp. 526–541. https://doi.org/10.5860/crl-374

Akers, K.G. & Doty, J. (2013) *Review of research data curation practices and attitudes of organizations.* Available at: https://trace.tennessee.edu/cgi/viewcontent.cgi?article=1126&context=utk_dataone

Bennett, L. & Booth, P. (2017) 'Research data curation: How libraries are changing the data landscape', *Journal of Librarianship and Information Science*, 49(1), pp. 37–48.

Binns, R. (2018) 'Ethics and data curation', in *Data Management and Data Curation for Researchers*. Springer.

Binns, R. (2019) 'Ethical data curation: Ensuring privacy, security, and trust in digital archives', *Journal of Data Ethics*, 1(2), pp. 58–70.

Bryant, R., Lavoie, B. & Rinehart, A.K. (2023) *Building research data management capacity: Case studies in strategic library collaboration*. OCLC Research. Available at: https://doi.org/10.25333/6dgm-9w15

Bunn, J. (2014) *Research data management and curation*. Facet Publishing. ISBN: 978-1856048694.

Christen, K. (2018) *Data curation in the digital era: Perspectives on ethics and the future of data repositories*. Springer. ISBN: 978-3319746886.

Creative Commons (2021) *About the licenses*. Creative Commons. Available at: https://creativecommons.org/licenses/

Davis, K. (2018) *Data ethics: The new competitive advantage*. Wiley. ISBN: 978-1119474774.

Davis, K. & Patterson, D. (2012) *The ethics of big data: Balancing risk and innovation*. O'Reilly Media. ISBN: 978-1449313094.

Data Ethics Framework by the UK Government (2018) *Ethical principles for data curation and usage in public service*. UK Government Digital Service. Available at: https://www.gov.uk/

Draude, C. & Lam, A. (2017) 'Ethics of data management and sharing in academic research', *Journal of Library and Information Science*, 19(1), pp. 12–25.

Eubanks, V. (2018) *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press. ISBN: 978-1250112632.

European Union (2018) *General Data Protection Regulation (GDPR). Official Journal of the European Union*. Available at: https://gdpr.eu/

Government of India (2019) *Personal Data Protection Bill, 2019*. Ministry of Electronics and Information Technology. Available at: https://www.meity.gov.in/content/personal-data-protection-bill-2019

Gurstein, M. (2015) *Open data: Empowering the people?* Springer. ISBN: 978-3319207327.

Harvard University Data Science Initiative (2020) *Principles of data curation*. Harvard University Press.

Harris, J.M. & Kramer, R. (2018) 'Understanding the role of ethics in the data curation lifecycle', *Data Management Journal*, 9(3), pp. 121–137.

Haustein, S., Wyatt, S. & Cox, A. (2017) *The ethics of research data curation*. Springer. ISBN: 978-3319263799.

Hernandez, G. & Lopez, C. (2019) 'The impact of GDPR on data curation in European research institutions', *Data Protection Review*, 12(4), pp. 105–119.

Indian Council of Medical Research (ICMR) (2017) *National ethical guidelines for biomedical research involving human participants*. Available at: https://www.icmr.nic.in/

International Association for Social Science Information Services and Technology (IASSIST) (2019) *Data management and ethical considerations in research*. Available at: https://iassistdata.org/

Kord Davis, D. & Patterson, K. (2017) *Data governance: The real business of data*. O'Reilly Media. ISBN: 978-1491941283.

Liu, Y. & Cabrera, D. (2020) 'Legal and ethical challenges in data sharing: The global perspective', *International Journal of Data Science and Technology*, 6(2), pp. 45–59.

Liu, Y. & Griffiths, T. (2020) 'Data privacy and its impact on global research', *Journal of Information Privacy and Security*, 16(2), pp. 89–105.

MacDonald, C. & Hughes, D. (2019) 'Data privacy and security: Navigating the ethics of data sharing', *International Journal of Information Ethics*, 18(3), pp. 120–134.

Murillo, A.P. (2024) 'Trends in data curation: Competencies and tools', *Proceedings of the Association for Information Science and Technology*, 61(1), pp. 1–4. https://doi.org/10.1002/pra2.1181

National Institutes of Health (NIH) (2021) *Best practices for ethical data sharing in biomedical research*. Available at: https://www.nih.gov/

Open Data Institute (ODI) (2020) *The ODI data ethics framework*. Open Data Institute. Available at: https://theodi.org/

Organization for Economic Co-operation and Development (OECD) (2019) *OECD principles on artificial intelligence*. Available at: https://www.oecd.org/going-digital/ai/principles/

Raji, R. & Shankar, V. (2021) 'Data protection laws in India: Challenges and opportunities', *Indian Journal of Law and Technology*, 7(1), pp. 45–67.

Rao, P.S. & Varma, M. (2020) 'Ethical challenges in handling sensitive data: A review of current practices in India', *Journal of Information Ethics*, 9(2), pp. 74–92.

Rao, V.S. (2018) 'Data protection laws in India: The road ahead', *International Journal of Technology Law*, 8(2), pp. 125–139.

Sharma, P. & Singh, S. (2020) 'Ethical considerations in research data curation', *Indian Journal of Research*, 34(4), pp. 200–210.

Solove, D.J. (2021) *Understanding privacy*. Stanford Law Books. ISBN: 978-0804752703.

Thielen, J., Marsolek, W. & Narlock, M. (2023) 'Conceptualizing slow curation', *Journal of eScience Librarianship*, 12(2), e740. https://doi.org/10.7191/jeslib.740

Timmerman, M. & de Man, M. (2018) 'Ethical and legal issues in open data: A comparison between European and U.S. perspectives', *Journal of International Data Ethics*, 10(3), pp. 153–166.

U.S. Department of Health and Human Services (1996) *Health Insurance Portability and Accountability Act of 1996 (HIPAA)*. Available at: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html

Webster, J. & Jacobson, T. (2016) 'The role of data curation in research: Ethics and legal frameworks', *Data Science & Technology Journal*, 3(2), pp. 75–88.

Wright, D. & De Hert, P. (2017) *Privacy impact assessment: Legal and ethical challenges in data sharing*. Springer.

Zook, M. & Boyd, D. (2020) 'Ethical dimensions of algorithmic data curation', *Computational Social Science*, 3(2), pp. 67–80.

Zuboff, S. (2019) *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Public Affairs.