Submitted on: 23.08.2023

Significance of Principles of Research Data Management in Curating Government Official Data: Perspectives in Indian Subcontinent

Ch. Ibohal Singh, *Ph.D.*
Professor & Head
Department of Library and Information Science, Manipur University, India
ORCiD: https://orcid.org/0000-0001-7768-885X; email: ibohal68@gmail.com

Rajkumari Sofia Devi, *Ph.D.*
Assistant Professor
Department of Library and Information Science, Manipur University, India
ORCiD: https://orcid.org/0000-0003-3615-6472; email: rksofia.02@gmail.com

Gyanajeet Yumnam
ICSSR Doctoral Fellow (Ph.D. Research Scholar)
Department of Library and Information Science, Manipur University, India
ORCiD: https://orcid.org/0000-0002-6332-1261; email: ygyanajeet@gmail.com

Abstract

Effective research data management (RDM) has been playing a crucial role in ensuring the reliability, accessibility, and longevity of valuable information, particularly within the context of curating official government data. In this paper, the significance of the principles of RDM in the Indian Subcontinent has been explored by examining the perspectives about the curation of official government data. The region has witnessed a rapid expansion in data-driven decision-making processes across the various governmental sectors in recent years. However, the research data management in this context has often been overlooked, leading to data loss, inaccessibility, and poor data quality. Therefore, implementing robust principles of RDM becomes imperative to address these challenges effectively. This paper aims to highlight the importance of RDM principles in curating official government data. It sheds light on the benefits of adopting such principles, including enhanced data integrity, improved data sharing and collaboration, and increased reproducibility of research findings. The same also explores the specific perspectives considering the region's unique socio-cultural, economic, and technological factors. It

addressed the challenges faced in implementing RDM principles, such as limited awareness, inadequate infrastructure, and varying levels of digital literacy among data custodians and researchers. To effectively promote the adoption of RDM principles in the curation of government official data, the study outlines essential strategies and recommendations. These mainly include capacity-building programs, training workshops, the development of RDM guidelines, and the establishment of dedicated data repositories or platforms for data sharing and preservation. By implementing these principles, governments can ensure the preservation and accessibility of valuable data, leading to evidence-based policy decisions, enhanced research outcomes, and improved governance. This study seeks to contribute towardsadvancing RDM practices and encourage their widespread adoption in the subcontinent.

***Keywords**: Research Data management, Data Curation, Preservation, Government Data, Indian Sub-Continent*

---

1. Introduction

Data has emerged as a cornerstone of decision-making, policy formulation, and strategic planning across various sectors in the modern era. Government bodies are pivotal in shaping the socio-economic landscape as vast and diverse datasets custodians. The effective management of government official data is imperative to ensure transparency, accountability, and informed governance and facilitate robust research and evidence-based policy-making (Tam & Clarke, 2015). The sheer volume and complexity of government data, encompassing fields such as demographics, economics, public health, education, and more, necessitate meticulous handling and organization (Wang & Lo, 2016). The advent of digital technologies and the proliferation of data collection mechanisms have led to exponential growth in data production, thereby underscoring the urgency of implementing robust research data management practices (Janssen et al., 2020).

Research data management (RDM) involves a comprehensive set of strategies, processes, and tools to systematically collect, organize, preserve, and share data generated during research endeavors (Surkis & Read, 2015; Patel, 2016). When applied to government official data, data management refers to the management of data collected by government agencies. This data can be used to make decisions, improve services, and hold government accountable. These practices acquire even greater significance due to their potential to influence critical decisions affecting millions of lives. Effective research data management ensures data integrity, security, accessibility, and longevity, mitigating the risks of data loss, unauthorized access, and obsolescence (Mittal et al., 2023). Data governance is the set of policies and procedures that govern how data is managed and ensures that data is accessible, reliable, and secure for ensuring government data is used effectively and ethically (Thompson et al., 2015).

Moreover, using official government data in scholarly research can drive innovation, inform public discourse, and address complex societal challenges. From analyzing historical trends to

projecting future scenarios, researchers rely on access to accurate and reliable data to draw meaningful conclusions and provide insights that guide policy direction (Charalabidis et al., 2016). Hence, a well-structured research data management framework bridges the government's data-producing machinery and the research community, fostering a symbiotic relationship that benefits both stakeholders. This study investigates the multifaceted realm of research data management, specifically focusing on official government data of the Indian Sub-Continent. It explores the challenges and opportunities inherent in managing such data, the evolving landscape of data governance, and the pivotal role of collaboration between government agencies and researchers. By dissecting the intricacies of this intersection, we aim to shed light on the significance of effective research data management in harnessing the potential of official government data for the betterment of society.

2. Key Aspects of Government Data Management

- **Data collection**: Government agencies collect data from various sources, including surveys, censuses, and administrative records. The data collected is used to track government programs, make policy decisions, and provide services to the public.
- **Data storage**: Government data is stored in various ways, including on-premises servers, cloud-based platforms, and data warehouses. The storage method used depends on the size and sensitivity of the data.
- **Data security**: Government data must be protected from unauthorized access, use, disclosure, modification, or destruction. Data security measures include encryption, access control, and audit logging.
- **Data access**: Government data must be accessible to authorized users in a timely manner. Data access is typically controlled by a policy defining who can access which data and for what purpose.
- **Data quality**: Government data must be accurate, complete, and up-to-date. Data quality is ensured through validation, cleaning, and auditing.
- **Data sharing**: Government data can be shared with other government agencies, the public, and private businesses. Data sharing is governed by data sharing policies, which define the terms and conditions under which data can be shared.

3. Methodology and Scope

Data and information were sourced from official government websites, ministry publications, documents, and directories. The research primarily centers on the management of research data within the government of the Indian Subcontinent, with a specific focus on India. The study's scope encompasses data accessible upto 1st June, 2023.

4. Indian Sub-Continent

The Indian subcontinent is a physiographic region in Southern Asia, situated on the Indian Plate, projecting southwards into the Indian Ocean from the Himalayas. Geopolitically, it spans major landmasses from Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka. Although "Indian subcontinent" and "South Asia" are often used interchangeably to denote the region, the geopolitical term of South Asia frequently includes Afghanistan, which is not considered part of the subcontinent. The Indian subcontinent is home to over 1.7 billion people, making it the second most populous region after East Asia. A diverse range of cultures, languages, and religions characterizes the region. Most of the population is Hindu, followed by Muslims, Christians, and Sikhs.

4.1. Scenario on RDM in the Sub-Continent

Governments within the Indian Subcontinent oversee curating and releasing the aforementioned Open Government Data (OGD) portals (Table 1). These platforms generously provide an array of data and information to the public free of charge, without any usage restrictions. The core purpose of these portals is to uphold transparency, nurture collaboration, and ignite innovation. The specific functionalities of OGD may differ across countries or regions, but their overarching objectives remain consistent. These encompass enhancing government transparency, accountability, and engagement, fueling creative advancements, and enhancing public services. The dataset predominantly covers Census and Survey, Governance and Administration, Home Affairs and Enforcement, Education, Health and Family Welfare, Youth and Sports, Judiciary, Biotechnology, Water and Sanitation, Population, and Agriculture.

Table 1: Government RDM Platforms in Indian Sub-Continent

Sl. No.	Country	No. of Ministries	Portal Name	Links to the Data Portal
1.	India	54	Open Government Data (OGD) Platform of India	https://data.gov.in/
2.	Bangladesh	40	Bangladesh Open Data	http://data.gov.bd/
3.	Nepal	28	Open Data Pakistan	http://nationaldata.gov.np/
4.	Bhutan	10	National Statistics Bureau	https://www.pbs.gov.pk/
5.	Pakistan	31	Pakistan Bureau of Statistics	https://opendata.com.pk/
6.	Sri Lanka	28	Open Data Portal of Sri Lanka	https://data.gov.lk/
7.	Maldives	16	Maldives Bureau of Statistics	https://statisticsmaldives.gov.mv/

Open Government Data holds multifaceted utility, catering to various applications. This comprehensive resource can serve various purposes, encompassing:

(a) Informed Decision-Making: OGD empowers citizens, researchers, and policymakers to make well-informed decisions by accessing reliable, up-to-date data. This facilitates evidence-based strategies in both the public and private sectors.

(b) **Accountability and Transparency**: By sharing government-generated data openly, the accountability of public institutions is reinforced. This encourages a culture of transparency, as citizens can scrutinize activities and hold authorities accountable.

(c) **Innovation and Research**: OGD serves as a wellspring of inspiration for innovators and researchers. It enables the development of novel technologies, solutions, and services that can address societal challenges effectively.

(d) **Economic Growth**: Entrepreneurs and businesses can leverage OGD to identify market trends, consumer behavior, and emerging opportunities. This drives economic growth by fostering new enterprises and optimizing existing ones.

(e) **Social Empowerment**: OGD democratizes access to information, enabling citizens to engage actively in civic matters. Informed participation enhances the public's ability to contribute meaningfully to policy discussions and societal progress.

(f) **Urban Planning and Development**: Local governments can utilize OGD to enhance urban planning and infrastructure development. Accurate data aids in designing sustainable cities that cater to the needs of their residents.

(g) **Healthcare and Education**: OGD supports the healthcare and education sectors by providing insights into public health trends, educational performance, and resource allocation. This aids in devising targeted interventions and enhancing services.

5. Indian Scenario

Government official data in India encompasses a wide range of information, including demographic data, censuses, administrative records, economic indicators, health statistics, educational data, and research studies. Multiple ministries, departments, and agencies are involved in data collection, each responsible for specific sectors or areas of governance. These data sets are crucial for understanding societal trends, identifying gaps, and formulating evidence-based policies and programs. However, the availability and usability of this data depend on effective research data management practices. Proper data documentation, storage, preservation, and dissemination are essential to maintain data integrity, facilitate reproducibility, and enable data sharing within the research community. To ensure the quality and reliability of data, the government has established standardized methodologies, protocols, and tools for data collection. These protocols adhere to international standards and best practices, enabling comparability and compatibility with global datasets. Emphasis is placed on data standardization, harmonization, and validation processes to minimize errors and inconsistencies.

Research Data Management (RDM) is crucial to curating the Indian government's official data. Effective research data management practices are paramount due to the vast amount of data generated and the need for accurate and reliable information to inform policy formulation, planning, and evaluation. This article provides an elaborate overview of the research data management practices employed by the Indian government, highlighting key aspects such as data collection, organization, storage, security, accessibility, utilization and preservation, and sharing of research data to ensure its long-term usability and accessibility. Effective RDM practices are essential for promoting transparency, reproducibility, and accountability in

research, enabling data-driven decision-making, and fostering collaboration among researchers. In the Indian context, several initiatives and guidelines have been implemented to improve RDM and enhance the quality and impact of government data.

5.1. Steps Taken by the India Government

The Indian government has recognized the importance of effective data management and has taken several steps to promote RDM practices. One notable initiative is the National Data Sharing and Accessibility Policy (NDSAP) introduced by the Department of Science and Technology (DST) in March 2012. The NDSAP aims to facilitate data sharing, accessibility, and reuse among various stakeholders, including government agencies, researchers, and the public. It provides guidelines for data publication, metadata standards, and licensing frameworks, promoting transparency, accountability, and data preservation for long-term usability.

Storage and preservation of research data are critical for its long-term accessibility. The National Data Sharing and Accessibility Policy (NDSAP) in India mandates the creation of a National Data Archive to preserve and provide access to data generated by various government organizations. This archive ensures the longevity of data by employing appropriate storage infrastructure and preservation techniques, such as backup systems, data replication, and disaster recovery plans. By safeguarding data against loss, corruption, and unauthorized access, RDM practices contribute to the integrity and reliability of government data.

To promote data sharing and collaboration, the Indian government encourages the use of open data licenses. The National Data Sharing and Accessibility Policy defines guidelines for licensing datasets, enabling researchers to reuse and build upon existing data. Open licenses such as the Creative Commons licenses are recommended, which allow users to freely access, distribute, and modify the data while respecting appropriate attribution and usage conditions. By adopting open data licenses, the Indian government facilitates the creation of a vibrant ecosystem of data-driven research and innovation.

Table 2: RDM Data Repository of India under GOI

Sl. No.	Name	Ministry (GOI)	Links to the Data Portal
1.	CSIR – Knowledge Resource Centre	Ministry of Science and Technology	https://csirhrdc.res.in/knowledge-resource
2.	ICAR-Krishi (Agricultural Knowledge Resources and Information System Hub for Innovations)	Ministry of Agriculture and Farmers Welfare	https://krishi.icar.gov.in/
3.	National Informatics Centre (NIC)	Ministry of Electronics and Information Technology (MeitY)	https://www.nic.in/
4.	ICSSR Data Service	Ministry of Education	http://www.icssrdataservice.in/
5.	Indian Space Science Data Center (ISSDC)	ISRO, Ministry of Electronics and Information Technology (MeitY)	https://www.issdc.gov.in/
6.	National Digital Library of India (NDLI)	Ministry of Education	https://ndl.iitkgp.ac.in/
7.	Open Government Data Platform (OGD)	Ministry of Electronics and Information Technology (MeitY)	https://data.gov.in/

5.2. Open Government Data (OGD) Platform India

To support the implementation of the NDSAP, the National Informatics Centre (NIC) has developed the Open Government Data (OGD) Platform in India (Fig. 1). The OGD Platform is a centralized repository and access point for various government datasets, including official documents, reports, and statistical data. The platform promotes data interoperability by adopting international standards such as the Data Catalog Vocabulary (DCAT). It facilitates the creation of metadata to describe datasets. By ensuring consistent and comprehensive metadata, researchers, policymakers, and citizens can quickly locate and assess the suitability of available government data for their research in various formats, fostering transparency and data-driven decision-making. Additionally, various ministries and departments have data portals, allowing easy access to sector-specific data.

Fig. 1: Screenshot of Open Government Data Portal (OGD) of Government of India (Source: https://data.gov.in/)

Furthermore, the Indian government has established the National Data Repository (NDR) as a comprehensive data repository for all exploration and production activities in the oil and gas sector. The NDR facilitates the secure storage, sharing, and dissemination of geoscientific data and documents generated by government bodies, public sector undertakings, and private companies operating in the sector.

6. Role of the National Informatics Centre (NIC)

As citizens' demands for online services continue to rise and governments launch numerous eGovernance Projects, the demand for Data Centres is skyrocketing. Establishing a strategic infrastructure that enables high availability, rapid scalability, efficient management, and optimal resource utilization is imperative. This ongoing requirement emphasizes the need for robust data centers to meet the growing expectations and demands of citizens and the government.

The National Informatics Centre (NIC) has established cutting-edge National Data Centres at NIC Headquarters in Delhi, Pune, Hyderabad, and Bhubaneswar to cater to the increasing demand. Additionally, they have set up 37 smaller Data Centres in different State Capitals, ensuring comprehensive services to governments at all levels. As the technology partner of the Government of India, NIC operates under the Ministry of Electronics and Information Technology (MeitY). Since its inception in 1976, NIC's primary goal has been to deliver technology-based solutions to the Central and State Governments. Established by NIC, these Data Centers amalgamate 24/7 system operations and management with skilled on-site personnel. They serve as the nucleus of India's e-Governance infrastructure, offering services to various e-Governance initiatives implemented by the Government of India.

In the realm of data infrastructure, India has made significant strides in recent years. The journey began in 2008 with establishment of the first Data Centre in Hyderabad, setting the stage for a technological revolution. This was followed by the inauguration of the NDC Pune in 2010, NDC Delhi in 2011, and NDC Bhubaneshwar in 2018, each contributing to the growing digital landscape of the country. To further expand the reach and capabilities of data services, Prime Minister Narendra Modi, in February 2021, laid the foundation stone of the inaugural National Data Centre for the North Eastern Region (NEDC) in Guwahati, Assam, using the power of video conferencing. The NEDC is poised to become a symbol of digital empowerment for the region, equipped with cutting-edge technology, including a state-of-the-art network and Security Operating Centres.

At the core of the NEDC lies the Network Operating Centre, designed to diligently monitor and manage the region's critical ICT infrastructure, ensuring uninterrupted availability of services around the clock. This progressive endeavor aims to catalyze the region's growth by leveraging the transformative potential of digital innovation.

In 2014, the National Informatics Centre (NIC) introduced the National Cloud Services as part of the MeghRaj Government of India Cloud Initiative. This initiative aimed to revolutionize the government's approach to cloud computing. NIC Cloud Services have been established across several key locations, including Bhubaneswar, Delhi, Hyderabad, and Pune, utilizing the National Data Centres. The primary objective of these cloud services is to meet the demands of the Digital India Programme and the expanding requirements of existing projects. Over 18,000 Virtual Servers have been allocated to more than 1100 Ministries/Departments, dedicated explicitly to e-Governance Projects to achieve this.

By leveraging the power of cloud computing, the NIC aims to enhance the efficiency and effectiveness of various government initiatives. Virtual servers allow for scalable and flexible infrastructure, enabling Ministries/Departments to implement their e-Governance Projects seamlessly. This concerted effort signifies the government's commitment to embracing technology-driven solutions for governance, promoting digital inclusiveness, and accelerating the nation's progress toward a digitally empowered society.

7. Data Curation and Preservation

7.1. Role of ICSSR Data Service (INFLIBNET)

Curating and preserving government data is vital for its long-term usability. The National Data Repository for Social Sciences (NDRFSS) was launched by the Ministry of Statistics and Programme Implementation (MoSPI) as a central repository for social science data. The "ICSSR Data Service" is the culmination of the signing of the Memorandum of Understanding (MoU) between the Indian Council of Social Science Research (ICSSR) and the Ministry of Statistics and Programme Implementation (MoSPI). The MoU provides for setting up "ICSSR Data Service: Social Science Data Repository" and hosting NSS and ASI datasets generated by MoSPI. Under the initiative, social science research institutes, NGOs, individuals, and others dealing with social science research are also being approached to deposit/provide their research datasets for hosting in the repository of ICSSR Data Service. The ICSSR Data Service includes social science and statistical datasets of various national-level surveys on debt & investment, domestic tourism, enterprise survey, employment and unemployment, housing condition, household consumer expenditure, health care, etc., in its repository. ICSSR Data Service aims to facilitate data sharing, preservation, accessibility, and reuse of social science research data collected from the entire social science community in India & abroad. The Information and Library Network (INFLIBNET) Centre, Gandhinagar, has been assigned to set up the data repository.

7.2. Indian Space Science Data Center (ISSDC)

The Indian Space Science Data Center (ISSDC) is a vital hub within the Indian Deep Space Network (IDSN) Byalalu campus of ISTRAC/ISRO. Its overarching objective is to furnish an extensive range of services to the global science community regarding the science missions of the Indian Space Research Organization (ISRO). ISSDC is a beacon with cutting-edge technology to facilitate data ingestion, processing, archival, and dissemination. Its infrastructure is designed to offer formidable computation power, expansive storage capacity, and a robust high-bandwidth network, ensuring the utmost security for hosting diverse applications necessary to support ISRO's planetary, lunar, and space science missions.

The ISSDC boasts a multi-layered architecture meticulously crafted to accommodate the myriad requirements of these missions. Each layer exhibits scalability, resilience, and flexibility, effortlessly adapting to the demands of present and future planetary and space science endeavors. The primary beneficiaries of this remarkable facility are the principal investigators overseeing the science payloads. Additionally, the data is made accessible to scientists from

other institutions and even the general public, fostering an atmosphere of openness and collaboration in space exploration. The ISRO Science Data Archive (ISDA) is a comprehensive repository housing all the scientific data collected from Indian science missions, beginning with Chandrayaan-1. Each mission's archive contains a wide range of data, including raw and processed data, calibration data, supplementary data, derived data products, documentation, and software. ISDA adopts the established archive standards of the Planetary Data System (PDS) and adheres to the guidelines set by the International Planetary Data Alliance (IPDA). This ensures compliance with global standards for long-term data preservation, maintaining data usability, and facilitating the scientific community's access to high-quality data for analysis. Initially, the principal investigators of the science payloads have exclusive access, but eventually, and the data will be made available to scientists from other institutions and the general public.

8. Education Sector

The **Ministry of Education, Government of India,** has recognized the importance of research data management in promoting scientific inquiry and innovation. Several initiatives and programs have been implemented to facilitate curating and managing research documents and data. Here are some key initiatives:

a) **National Digital Library of India (NDLI)**: NDLI is an ambitious project to build a comprehensive digital repository of educational resources, including research papers, theses, and datasets. It provides free access to various resources to students, researchers, and the general public. NDLI incorporates various technologies, such as cloud computing and machine learning, to enable efficient search, retrieval, and preservation of digital content.

b) **Open Access Policy**: The Ministry of Education has implemented an open access policy to promote the dissemination of research findings. Under this policy, publicly funded research must be published in open-access journals or repositories, making it freely available. This initiative ensures that research outputs are easily accessible, fostering collaboration and knowledge sharing.

c) **Research Data Management (RDM) Guidelines**: The Ministry of Education has developed guidelines for research data management to provide a framework for researchers to manage their data effectively. These guidelines cover various aspects, including data documentation, storage, sharing, and long-term preservation. By following these guidelines, researchers can ensure the integrity, accessibility, and reusability of their data.

d) **National Knowledge Network (NKN)**: NKN is an advanced high-speed network infrastructure connecting educational and research institutions nationwide. It enables seamless communication and collaboration among researchers, facilitating the sharing and exchange of research documents and data. NKN also supports data-intensive research through its high-performance computing and storage facilities.

e) **Data Sharing and Collaboration Platforms**: Also, the platforms of the National Data Sharing and Accessibility Policy (NDSAP) facilitate data sharing and collaboration among researchers by providing a legal framework for data sharing across various sectors, including research. Platforms like **Data.gov.in** and **Open Government Data Platform (OGD)** enable researchers to access and utilize government data for their research projects.
f) **Research Data Repositories**: The Ministry of Education encourages the establishment of research data repositories to promote data sharing and preservation. These repositories provide secure storage and access to research data, ensuring its long-term availability. Examples of such repositories include Shodhganga for ETDs, ShodhGangotri, Indian Genome Variation Consortium (IGVC), which hosts genomic data, and the Indian Space Science Data Centre (ISSDC), which curates data from space research missions.
g) **Capacity Building and Training Programs**: Capacity building and training programs are instrumental in promoting effective RDM practices among researchers and government officials. The Indian government, in collaboration with academic institutions and research organizations, conducts workshops, seminars, and training sessions to enhance data management skills. These initiatives focus on data documentation, sharing, citation, and security. Organizations like the Indian Council of Social Science Research (ICSSR) and the National Council of Educational Research and Training (NCERT) conduct capacity-building activities to promote good RDM practices among researchers. By equipping researchers and government officials with the necessary knowledge and skills, RDM capacity-building programs contribute to improving data quality and accessibility.
h) **Research Funding and Grants**: The government provides research funding and grants to support projects focusing on data-driven research and innovation. These funds can be used for data collection, analysis, and management activities. The government encourages researchers to adopt robust data management practices and produce high-quality research outputs by providing financial support.
i) **Collaboration and Partnerships**: Collaboration between government agencies, research institutions, and the private sector is essential for effective RDM. The Indian government collaborates with international organizations such as the World Bank, United Nations, and Open Government Partnership (OGP) to promote best practices in data management. Public-private partnerships are encouraged to leverage data curation and preservation expertise and resources.

9. The Challenges

Despite of these initiatives, several challenges persist in curating Indian government official documents and data.

9.1. Data breaching and hacking incident

India was the second-most affected country by data breaches in 2022, accounting for 20% of all records exposed. This is according to a report by Tenable, a cybersecurity company based in Maryland, US. The report analyzed over 1,300 data breaches between November 2021 and October 2022. The report found that the most common type of data breach in India was a ransomware attack, which accounted for 33% of all breaches. Ransomware attacks are when hackers encrypt a victim's data and demand a ransom payment in exchange for decryption. Other common data breaches in India included unsecured databases (17%) and phishing attacks (14%) (CNBCTV18.com, 2023). As outlined in the report presented by the Ministry of Electronics and Information Technology, Government of India, to the Indian Parliament, 47 instances of data leaks and 142 cases of data breaches transpired over the preceding five years spanning from 2017 to 2022. Moreover, according to data monitored by CERT-In (Indian Computer Emergency Response Team), the years 2020, 2021, and 2022 witnessed 10, 5, and 7 incidents of data leaks specifically concerning governmental entities, respectively (The Wire, 2023). These data leaks and breaches involved citizens' sensitive personal information, such as their names, addresses, phone numbers, and Aadhaar numbers. Some notable examples are sidelined below:

- **Aadhaar data breach (2018)**: One of India's most significant data breaches, affecting over 1.1 billion people. The Aadhaar card is a 12-digit unique identification number issued by the government of India to its citizens. The data breach exposed the personal details of Aadhaar cardholders, including their names, addresses, fingerprints, and iris scans.
- **SBI data breach (2019)**: This breach affected over 3.2 million customers of the State Bank of India (SBI), the largest bank in India. The data breach exposed the personal details of SBI customers, including their names, account numbers, and contact information.
- **Justdial data breach (2019)**: This breach affected over 100 million users of Justdial, a popular online directory service in India. The data breach exposed the personal details of Justdial users, including their names, phone numbers, and email addresses.
- **Kudankulam nuclear power plant data breach (2019)**: This data breach affected the Kudankulam nuclear power plant, one of India's largest nuclear power plants. The data breach exposed the employees' personal details of the Kudankulam nuclear power plant, including their names, addresses, and contact information.
- **BigBasket data breach (2020)**: This breach affected over 20 million customers of BigBasket, an online grocery store in India. The data breach exposed the personal details of BigBasket customers, including their names, addresses, payment information, and order history.
- **Unacademy data breach (2020)**: This breach affected over 20 million users of Unacademy, an online education platform in India. The data breach exposed the personal details of Unacademy users, including their names, email addresses, and phone numbers.
- **Air India data breach (2021)**: This data breach affected over 4.5 million customers of Air India, the national airline of India. The data breach exposed the personal details of

Air India customers, including their names, addresses, travel information, and credit card data.

- **Dominos India data breach (2021):** This breach affected over 1.2 million customers of Dominos India, a popular pizza chain in India. The data breach exposed the personal details of Dominos India customers, including their names, addresses, and payment information.

These few data breaches that have happened in India in recent years. These breaches have exposed the personal details of millions of Indians and have raised concerns about the security of personal data in India. Data breaches are a serious threat to individuals and businesses. They can lead to identity theft, financial fraud, and other crimes. It is important for individuals to be aware of the risks of data breaches and to take steps to protect their data. Businesses should also protect their data by implementing strong security measures and educating their employees about data security. The Indian government has taken steps to address the issue of data breaches. In 2018, the government passed the Personal Data Protection Bill to protect individuals' personal data. The bill is still being debated in the Parliament, but it is a step in the right direction. The government should also educate businesses and individuals about data security. This includes providing information about the risks of data breaches and how to protect personal data. Businesses should also be required to report data breaches to the government. By taking these steps, the government can help to protect the personal data of Indians and prevent future data breaches.

Data Quality, authenticity, and standardization: One of the primary challenges in managing government official data is ensuring its quality, accuracy, reliability, and standardization of data formats and metadata across different sources. It is crucial for research and policy-making. However, data manipulation, incomplete records, and outdated information have been reported. Implementing robust quality control measures and establishing mechanisms for data verification is essential for building trust in government data. Data collected by various government agencies may have variations in definitions, formats, and collection methodologies, making it difficult to integrate and analyze them comprehensively.

Data Accessibility and Privacy: While data openness and accessibility are crucial for transparency and public accountability, balancing them with data privacy concerns is also necessary. Government official data often contains sensitive personal or classified data, and ensuring data security and privacy is complex. Striking the right balance between openness and privacy is a critical challenge. Developing secure data access protocols, anonymization techniques, and data protection frameworks can help mitigate these concerns.

Infrastructure and Technical Capabilities: Effective research data management requires robust technical infrastructure and data storage, processing, and analysis capabilities. Many government departments in India still lack adequate technological resources and expertise, which makes it challenging implementing sophisticated data management practices.

Interagency Collaboration and Coordination: Government official data is collected and maintained by multiple agencies and departments at both the central and state levels. Ensuring coordination and collaboration among these entities for effective data management is a

significant challenge. Harmonizing data collection methods, sharing protocols, and avoiding duplication of efforts is essential for streamlining the data management process.

10. Opportunities and Solutions

Looking ahead, several strategies can enhance RDM practices for curating official Indian government documents and data. First, promoting a culture of data sharing and collaboration among government agencies, researchers, and the public is essential. Encouraging data publication, providing incentives for data sharing, and fostering data literacy can facilitate knowledge exchange and reuse.

a) **Data Governance Framework**: Developing a comprehensive data governance framework is essential to address the challenges in research data management in India. The framework should define roles, responsibilities, data collection, storage, sharing, and usage standards. It should also incorporate data protection and privacy measures, ethical considerations, and mechanisms for monitoring and enforcement.
b) **Capacity Building and Training**: To overcome the technical and skill-related challenges, it is imperative to invest in capacity building and training programs for government officials, researchers, and data custodians involved in data management. These programs can focus on data management best practices, data curation techniques, and metadata standards, equipping stakeholders with the necessary knowledge and skills to manage and curate official documents and data effectively.
c) **Metadata and Documentation**: Metadata provides information about the data, is crucial for understanding the data context, and facilitates data discovery. Government agencies should adopt standardized metadata schemas and ensure comprehensive documentation for all datasets. This will enable researchers to locate and utilize the data effectively.
d) **Data Sharing Platforms and Infrastructure**: Developing centralized data-sharing platforms and infrastructure can enhance the accessibility and interoperability of government official data. These platforms should include user-friendly interfaces, data visualization tools, and mechanisms to ensure data security and privacy. Open data initiatives and collaborations with academic institutions and research organizations can facilitate data sharing and utilization.

Lastly, continuous monitoring, evaluation, and feedback mechanisms should be established to assess the effectiveness of RDM initiatives and address emerging challenges. Regular audits, user surveys, and stakeholder consultations can provide valuable insights for refining RDM policies, improving data quality,

11. Conclusion

The significance of principles of research data management in curating government official data within the Indian subcontinent is paramount for fostering transparency, accountability, and informed decision-making. The meticulous application of these principles ensures the preservation of invaluable data. It enhances its accessibility and usability for researchers, policymakers, and the general public. The Indian subcontinent, with its diverse and extensive

data landscape, benefits significantly from a robust data management framework. The region can overcome data fragmentation and inconsistency challenges by adhering to internationally recognized standards, such as data categorization, documentation, and sharing protocols. This, in turn, will contribute to more accurate analyses, better policy formulation, and effective governance.

Moreover, the principles of research data management act as catalysts for interdisciplinary collaboration. As different sectors recognize the potential of shared data resources, silos are dismantled, allowing for holistic and comprehensive research endeavors. This collaborative approach promotes innovation and empowers stakeholders to address complex societal issues more effectively. In an era characterized by information overload, ensuring the authenticity and reliability of official government data is imperative. By adopting data management principles, the Indian subcontinent can mitigate data quality, security, and privacy concerns. This, in turn, enhances public trust and confidence in government institutions.

Finally, embracing research data management principles sets the stage for a data-driven transformation within the Indian subcontinent. As data becomes an increasingly vital resource, integrating these principles into data curation processes will undoubtedly contribute to the region's progress, sustainability, and socio-economic development. By doing so, the Indian subcontinent can harness the full potential of its government's official data to pave the way for a brighter and more informed future.

References

Tam, S. M., & Clarke, F. (2015). Big data, official statistics and some initiatives by the Australian Bureau of Statistics. *International Statistical Review*, *83*(3), 436-448. https://doi.org/10.1111/insr.12105

Wang, H. J., & Lo, J. (2016). Adoption of open government data among government agencies. *Government Information Quarterly*, *33*(1), 80-88. https://doi.org/10.1016/j.giq.2015.11.004

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. *Government Information Quarterly*, *37*(3), 101493. https://doi.org/10.1016/j.giq.2020.101493

Surkis, A., & Read, K. (2015). Research data management. *Journal of the Medical Library Association: JMLA*, *103*(3), 154. https://doi.org/10.3163/1536-5050.103.3.011

Patel, D. (2016). Research data management: a conceptual framework. *Library Review*, *65*(4/5), 226-241. https://doi.org/10.1108/LR-01-2016-0001

Mittal, D., Mease, R., Kuner, T., Flor, H., Kuner, R., & Andoh, J. (2023). Data management strategy for a collaborative research center. *GigaScience*, *12*. https://doi.org/10.1093/gigascience/giad049

Charalabidis, Y., Alexopoulos, C., & Loukis, E. (2016). A taxonomy of open government data research areas and topics. *Journal of Organizational Computing and Electronic Commerce*, *26*(1-2), 41-63. https://doi.org/10.1080/10919392.2015.1124720

Thompson, N., Ravindran, R., & Nicosia, S. (2015). Government data does not mean data governance: Lessons learned from a public sector application audit. *Government information quarterly*, *32*(3), 316-322. https://doi.org/10.1016/j.giq.2015.05.001

India sufferred second-highest data breaches in 2022 with 450 million records exposed: Report. cnbctv18.com. (2023, March 3). https://www.cnbctv18.com/technology/india-sufferred-second-highest-data-breaches-in-2022-with-450-million-records-exposed-report-16088781.htm

In past five year, 47 incidents of data leak and 142 data breaches: Meity. The Wire. (2023 March 16 2023). https://thewire.in/government/india-data-leak-breach-lok-sabha