Data Formats & Architecture

MARC21 (Machine-Readable Cataloging)

A standard digital format used by traditional ILMS platforms to encode bibliographic information so computers can process catalog records.

Dublin Core

A standardized set of basic metadata elements (e.g., Title, Creator, Subject) used to describe digital objects in institutional repositories.

JSON (JavaScript Object Notation)

A lightweight, text-based format for storing and exchanging data between web applications and library systems (e.g., exporting chat logs or API data).

LOD (Linked Open Data)

A method of publishing library metadata on the web so it can interconnect across global databases (e.g., linking a local library author catalog to Wikipedia or WorldCat).

Emerging Technologies & Privacy Laws

IoT (Internet of Things)

A network of physical devices (e.g., RFID gates, motion sensors, climate monitors) embedded with sensors and software that automatically collect and transfer data over the internet.

NLP (Natural Language Processing)

A branch of AI that enables computers to analyze, interpret, and extract meaning from human language (used to analyze unstructured reference chat transcripts).

DPDP Act, 2023 (India)

India's national data privacy framework. Mandates Indian institutions including academic libraries obtain clear patron consent, protect digital personal data, and erase records once purpose is served.

GDPR (EU - Global Benchmark)

The global benchmark privacy law enacted by EU. Sets strict rules on how organizations worldwide collect, process, anonymize, and erase personal data.

Privacy • Tech • Governance

What is Big Data?

Definition

Big Data is data that is too large, too fast, or too complex to be effectively handled by conventional data-processing tools. Big Data refers to massive, complex, and rapidly accumulating datasets that traditional library software or relational databases cannot store, process, or analyze efficiently.

Example in a Library

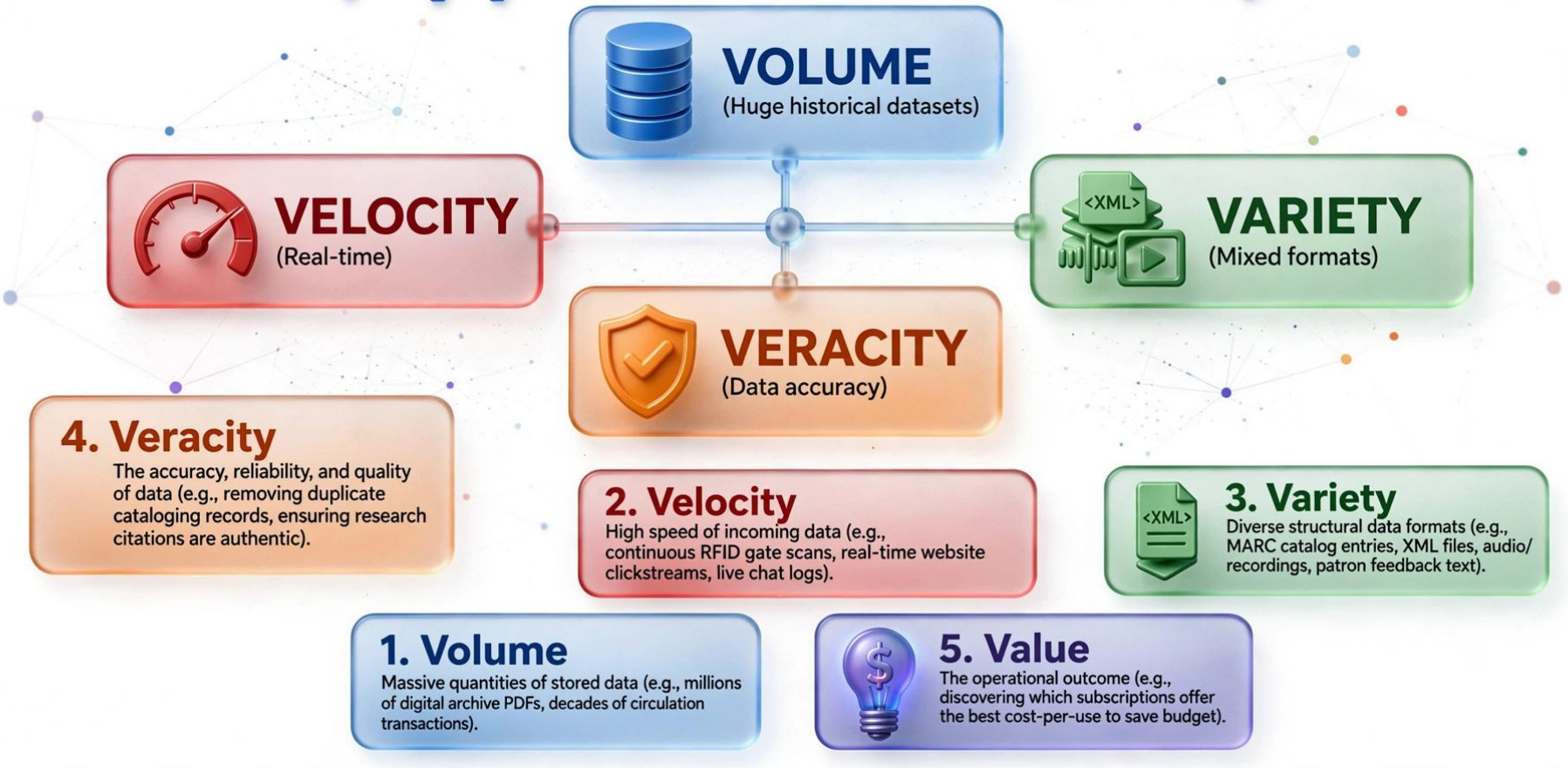
A digital library may generate search queries, downloads, circulation records, e-resources usage, website visits, user-generated content, and social-media interactions. When such data become very large, diverse, and continuously generated, they can be considered Big Data.

In One Line

Big Data = very large + rapidly generated + diverse + complex data that requires advanced methods for processing and analysis.

Big Data • Libraries • Analytics

The 5 V's of Big Data
(Applied to Libraries)



Main Characteristics of Big Data – 6 Vs

V 

1. Volume

Meaning • Huge amount of data

Example
Millions of library transaction records

V 

2. Velocity

Meaning • Data generated and processed rapidly

Example
Real-time website searches

V 

3. Variety

Meaning • Different types and formats of data

Example
• Text, images, video, audio, databases

V 

4. Veracity

Meaning • Accuracy and reliability of data

Example
• Removing duplicate or incorrect records

V 

5. Value

Meaning • Useful information obtained from data

Example
• Identifying user reading patterns

V 

6. Variability

Meaning • Data may change in meaning, form, or rate over time

Example
• Changing search behaviour during examinations

Volume • Velocity • Variety • Veracity • Value • Variability

Types of Big Data

1. Structured Data

Meaning: Data organised in a fixed, predefined format, usually rows and columns in a database.

Example:

Patron details, book/item records, and circulation transactions stored in an ILMS such as Koha or Alma.

2. Semi-Structured Data

Meaning: Data that does not follow a rigid table structure but contains tags, fields, keys, or other markers that organise the data.

Example:

MARC21/XML records, Dublin Core metadata encoded in XML, and JSON exports from digital repositories.

3. Unstructured Data

Meaning: Data without a predefined tabular or data model, making it more difficult to process using conventional database systems.

Example:

Virtual reference transcripts, scanned manuscript images, videos, audio recordings, full-text documents, and social-media comments.

Structured • Semi-Structured • Unstructured

Connecting Library Technologies to Big Data

Library Technology	How It Connects to Big Data & Analytics	Primary 5-V Focus
	**Integrated Library Management System (ILMS) (e.g., Koha, Alma)** Generates and stores large volumes of transactional data such as circulation, renewals, returns, acquisitions, and patron activity. Historical data can support collection-use analysis, acquisition planning, and weeding decisions.	Volume & Value
	**Radio Frequency Identification (RFID)** RFID readers and gates generate data about item movement, circulation, returns, and stock verification. When generated continuously across many items, these data can form high-volume data streams for analysis.	Velocity & Volume
	**Virtual Reference Services (e.g., Tawk.to, LibAnswers)** Generate chat transcripts and interaction records. Text analytics/NLP can identify frequently asked questions, topics, service patterns, and other characteristics of reference interactions.	Variety & Veracity
	**Smart Facility IoT (sensors and smart-card systems)** Sensors can continuously generate data about occupancy, foot traffic, environmental conditions, and use of library spaces. Analytics can identify patterns in facility utilisation.	Velocity & Volume

ILMS • RFID • Virtual Reference • Smart IoT | Volume • Velocity • Variety • Veracity • Value

Sources of Big Data in Libraries

ILMS / LMS

Circulation, acquisition, registration, and learning-activity data.

RFID

Item movement, circulation, returns, and inventory data.

Digital Libraries / Repositories

Metadata, downloads, searches, and full-text usage data.

Discovery Systems

Search queries, clicks, and resource-use patterns.

Virtual Reference

Chat transcrifts, queries, and interaction data.

IoT / Smart Library Systems

Occupancy, footfall, environmental, and equipment data.

Web & Social Media

Website visits, search behaviour, comments, shares, and engagement data.

ILMS • RFID • Digital Libraries • Discovery • Virtual Reference • IoT • Web & Social Media | Sources of Big Data

Benefits of Big Data for Libraries - Part 2

Part 2 of 2 • Advanced Insights for Libraries

6. Resource Optimisation

Optimise budgets, staff allocation, space, and equipment.

7. Research Support

Analyse scholarly communication, research trends, and citation patterns.

8. Performance Evaluation

Measure resource usage, service effectiveness, and institutional impact.

9. Operational Efficiency

Detect patterns that can improve circulation, inventory, and other routine operations.

Benefits Part 2: Resource Optimisation • Research Support • Performance Evaluation • Operational Efficiency

Benefits of Big Data for Libraries - Part 1

1. Evidence-Based Decision-Making

Supports decisions using data and usage patterns.

2. Improved Collection Development

Helps optimise selection, acquisition, and weeding.

3. Personalised Services

Enables more relevant resources and recommendations.

4. Better Resource Allocation

Supports efficient use of budgets, staff, space, and technology.

Part 1: Decision-Making • Collection • Personalised • Resource Allocation

Benefits of Big Data for Libraries - Part 2

5. Improved User Experience

Helps identify changing user needs and service gaps.

6. Predictive Planning

Helps anticipate resource demand and future service requirements.

7. Research Support

Provides insights into research trends, publications, citations, and collaboration.

8. Service Evaluation

Enables measurement of resource usage, service performance, and library impact.

• Part 2: User Experience • Predictive Planning • Research Support • Service Evaluation •

Role of the Library & Information Professional in Big Data Era

1. Data Literacy

Understand, evaluate, and use data correctly.

2. Data Management

Collect, organise, store, and maintain data properly.

3. Data Curation

Select, organise, preserve, and add value to data for future use.

4. Data Governence

Ensure data quality, privacy, security, access, and proper use.

5. Analytics & Evidence-Based Decision-Making

Analyse data and use the findings to improve library services and make informed decisions.

Data Literacy • Data Management • Data Curation • Data Governance • Analytics & Decision-Making

Challenges of Big Data Management (BDM)

1. Storage

Managing rapidly growing volumes of digital collections and system logs.

2. Processing

Processing and analysing millions of records efficiently.

3. Integration

Combining legacy data such as MARC21 with modern formats such as JSON and Linked Open Data.

4. Data Quality & Veracity

Handling incomplete, duplicate, inaccurate, and inconsistent data.

5. Security & Governance

Protecting user data and complying with privacy regulations such as the DPDP Act 2023 and GDPR.

Storage • Processing • Integration • Veracity • Security & Governance | BDM Challenges

Data Privacy Comparison: DPDP Act 2023 vs GDPR - Part 1

Feature
Scope

DPDP Act
Digital personal data processed in India, including digitised data.

GDPR
Personal data processed by organisations within its scope.

Library App
Protect patron borrowing history, and chat logs.

Feature
Consent

DPDP
Requires consent to be free, specific, informed, and unambiguous where consent is

GDPR
Consent must be freely given, specific, informed, and unambiguaus.

Library App
Clearly explain why personal data are collected.

Data Privacy Comparison: DPDP Act 2023 vs GDPR - Part 2

Feature • Data Retention

Data Retention

DPDP Act:
Personal data should not be retained when no longer necessary for stated purpose, subject to applicable requirements.

GDPR:
Storage limitation: retain data only as long as necessary for the purpose.

Library App:
Delete/anonymise unnecessary patron records.

Feature • Jurisdiction

Jurisdiction

DPDP Act:
Primarily applies to processing within India, with specified extra-territorial application.

GDPR:
Can apply to organisations outside EU when specified EU-related processing occurs.

Library App:
Important for libraries providing services across jurisdictions.

DPDP Act 2023 vs GDPR - Part 2: Data Retention • Jurisdiction

Data Privacy Comparison: DPDP Act 2023 vs GDPR

Feature | DPDP Act 2023 (India) | GDPR (EU) | Library Application

Feature	DPDP Act, 2023 (India)	GDPR (EU)	Library Application
Scope Feature	Digital personal data processed in India, including digitised data.	Personal data processed by organisations within its scope.	**Library App** Protect patron records, borrowing history, and chat logs.
Consent Feature	Requires consent to be free, specific, informed, and unambiguous where consent is the lawful basis.	Consent must be freely given, specific, informed, and unambiguous.	**Library App** Clearly explain why personal data are collected.
Data Retention Feature	Personal data should not be retained when no longer necessary for stated purpose, subject to applicable requirements.	Storage limitation: retain data only as long as necessary for the purpose.	**Library App** Delete/anonymise unnecessary patron records.
Jurisdiction Feature	Primarily applies to processing within India, with specified extra-territorial application.	Can apply to organisations outside EU when specified EU-related processing occurs.	**Library App** Important for libraries providing services across jurisdictions.

DPDP Act 2023 • GDPR • Privacy • Libraries • Consent • Retention • Jurisdiction

Big Data → Analytics → Insights → Library Decisions

A Simple Four-Step Process

1. Big Data → Collect

Large amounts of data are generated from library systems, RFID, websites, repositores, etc.

2. Analytics → Analyse

The data are examined to find patterns, trends, and relationships.

3. Insights → Understand

The analysis helps the library understand what is happening and what users need.

4. Library Decisions → Act

The library uses these findings to make better decisions and improve services.

Simple Example:

Circulation data → Analyse which books are frequently borrowed
→ Identify high-demand subjects
→ Buy more relevant books

In simple words:

Collect data → Analyse it → Understand what it tells us
→ Use it to make better library decisions.