Working with Data

Niamh Malin

2024-06-04

This guide explores the use of data within libraries through presenting questions to think through when undertaking your own data tasks, alongside practical examples from libraries around the world.

Introduction

Data is all around us, and it almost cannot be avoided. Libraries and universities are full of data, but staff often think they do not work with data. Have you ever been asked how many loans a book has had? What about how many visitors came in at the weekend? Or if a certain book is available? These are all data questions, answerable with quantifiable facts.

Data is often considered to be a number, like the number of times a book has been loaned, but what about who borrowed it, or what their review was? These are all different forms of data, and can provide all new insights to the popularity of a book - maybe a book was borrowed 10 times, but in fact got 10 1-star reviews, versus a book borrowed 5 times with 5 5-star reviews, using this data, which one would you recommend?

This guide will explore the use of data within libraries through proposing questions to think through when undertaking your own data tasks, alongside practical examples from libraries around the world.

What is Data?

Data is often overwhelming (the term 'Big Data' floats around a lot!), but that shouldn't stop you from exploring it, because it can greatly improve your work. Just to clarify, Big Data will not be discussed here, Big Data requires data to be high in volume and velocity, which is not the case for most library data (think about downloading a document every time the hashtag #photooftheday is used).

So first, let's establish some definitions (for this guide at least) while looking at one of our most central collections of data:

	Definition
Data	A collection of information.
Data Point	Single point of data (e.g. *a data point could be the title of a book, however alone that information tells us very little about the book. Therefore a 'record' containing relevant data points will be created.*).
Metadata	Data that provides information about other data (e.g. columns of data in a spreadsheet). (e.g. *in a library catalalogue the metadata about the book includes the price information, the isbn, author and title*
Dataset	A collection of data from a single source or for a single project. (e.g. *we might create a dataset of all catalogue records relating to a particular topic of interest to researchers*)
Database	A table with structured (organised) data.
Dataframe	A structured representation of data
Software	Program for a computer.
Hardware	The physical components of technology.
Back-end	The part of a system that is not usually visible or accessible to a user of that system.
Front-end	A software interface designed to enable user-friendly interaction with software.

Where is the data in a library?

Data can be qualitative (sentiment) or quantitative (numerical), therefore when we think about visitors to a library, do you know how they felt about their visit (qualitative data) or how many hours they spent at the library (quantitative data)?

Here are some examples to get you thinking about what data points stem from a library;

Institution	Staffing	Acquisitions
The Building / SpaceWider CommunityStakeholdersFacilitiesBusinessOpening Hours	Library StaffInstitution DepartmentsQualifications and Education	Subscriptions (ongoing)Purchasing (one time)SoftwareSuppliersSpending / Financial Information

Resources	Engagement	Patrons
Materials metadataCollection(s)Care and MaintenanceDiscovery ServicesAccessibility Services	Research Data ManagementTrainingOutreachToursCommunicationQuantity of Users	Helpdesk (in-person)Queries (email, online chat etc.)User SupportFeedbackInteractions

Considering the above, what do you interact with? Do you consider yourself to work with data?

Relevance to the Library Sector (Case Studies/Use Cases)

Data exists, so what?

Data should have a purpose. If data has no purpose, why waste your time managing it? The purpose of data could be any of the following;

- Provide insights to pertinent questions, and thus make informed decisions.
- Provide evidence of a decision or conclusion.
- Explore relationships / trends, to tell a story and provide insights.
- Explain (and display) complex information more effectively.

Data can provide additional power when making a decision (big or small) - knowing exactly how many users visit at the weekend can prove the library does require the funding and staffing to stay open on the weekend. However, if it proves no one is visiting, do not fudge the numbers. Use those numbers to question why, such as what else could attract users, what do patrons use it for currently, or was there a cause to the decline in weekend patrons?

Proof is in the Pudding

Knowing that data exists in a library is the first step, but what can be done with it? While this guide has touched on a few questions you can use data to answer, below will explore real world examples of libraries utilising their data to make powerful changes.

Data can inform decisions about:

- **Collection Management:** what is being used, when, by who, and why (if you have a review system).
  - Example: Within our collection, how often do users borrow material published before 2000? (But remember, borrowing is not the only form of usage!).
- **Outreach:** who forms our community, what matters to the community,

  - Example: Does the demographic of people attending our events reflect the demographic of people in our wider community? Which events could make the library more appealing to the wider community?

- **Funding / Expenditure:** Use data about patron usage, reader requests or patron feedback as part of an application for additional funding or staffing to prove your standing within the community.
  - Example: How much would it cost to buy an additional copy of a book, if it had more than three reservations at a given time?

- **Reviewing Success:** When starting a new project or approach, consider how it will be measured as a success. Knowing how to measure the impact of a project will enable you to actively reflect and produce solid facts about the success of your project.
  - Example: How can we measure patron satisfaction currently, so that at the end of the project a 10% increase can be quantified and thus the project can be considered a success?

Here are a few real world examples of big data projects within libraries, but remember, small or internal projects should be informed by data too.

- **Telling Stories with Library Data** July 2021

The Metropolitan Museum of Art (New York) Watson Library explores their production of data visualisations using Microsoft Power BI in regards to their library activities. Using the library management system (data which they already had), google analytics, digital collections and some manual tallies, they explored six years of library data. This exploration demonstrated the benefit of moving their blog from an external website, created an index of African American Artists,and ease in sharing library metrics for the previous six years publicly.

- **When is a Year Complete?** October 2023 It is well known that publications databases take time to update, but how long? Collection Analysis Librarian at Iowa State University, Eric Schares, wanted to know how long to wait before being able to analyse a calendar year of publications, and set about doing so by comparing data from three different publication services (Dimensions, Web of Science and Open Alex). The data (which continues to be updated), documented the rise of Open Alex, and demonstrated part of the effectiveness of each service for Iowa State University - however he also shares his code and thus you can replicate this data question at your own institution.

- **Use of Institution Data Analysis for Publisher Negotiations** July 2023 | **Utilising Data to Understand the Institutions Relationship with a Publisher** March 2024 Data Analyst at the University of Cambridge Library Niamh Malin, was responsible for supporting the many publisher negotiations. The first paper addresses the use of Microsoft Excel and Dimensions publications data to document the relationship of the university with Springer Nature. The second presentation documents the transition to

Microsoft Power BI to create data dashboards which ensure every publisher negotiation is informed by usage and publications data.

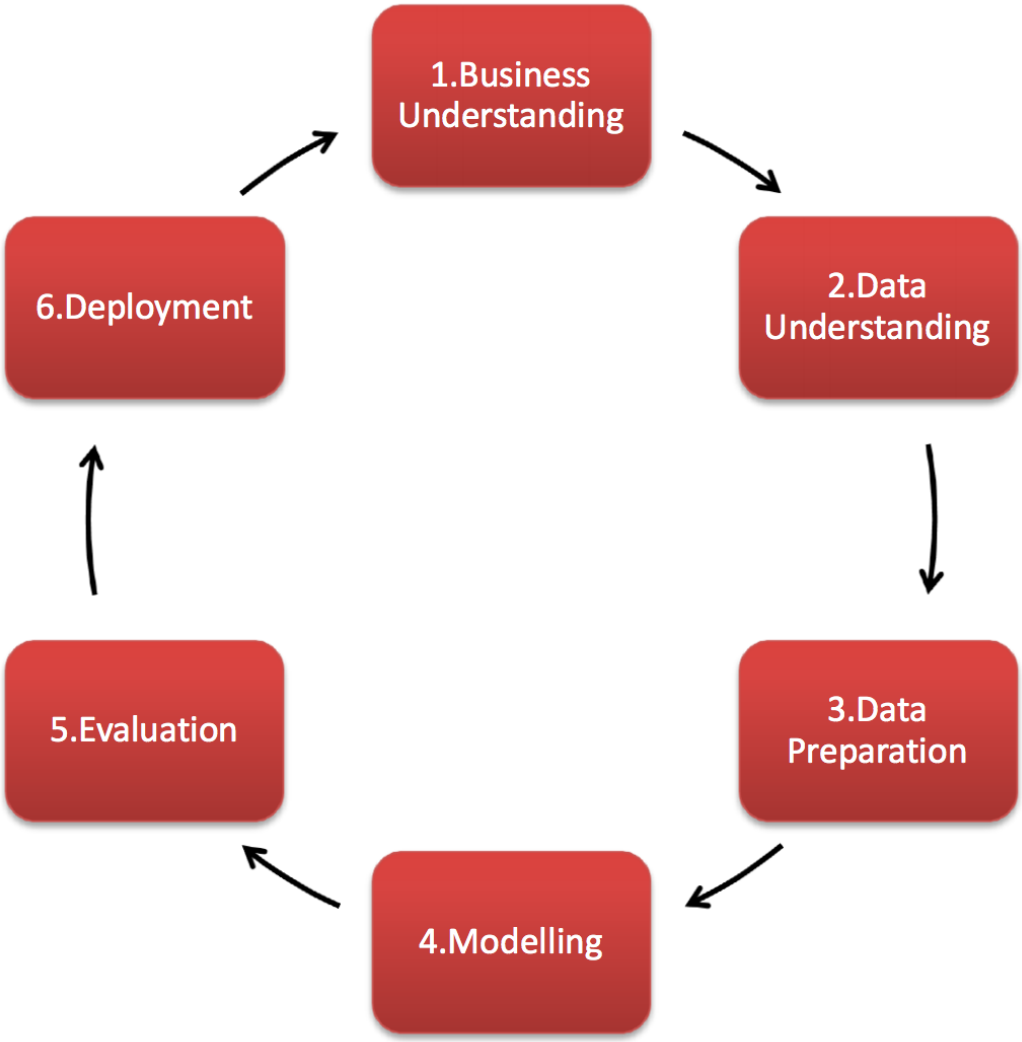
- **Using Analytics to Extract Value from the Library's Data** - Event Part One: Analytics Behind the Scenes and Part Two: Actionable Data Analysis September 2018. The National Information Standards Organization (NISO) hosted a two part webinar in 2018 about extracting value from library data. The slides for the six presentations are available, and cover topics such as; setting up a data analysis strategy, analysing metadata, utilising data visualisations, actioning data insights for utilising physical spaces and building confidence in your data analysis skills.
- **Libraries support data-driven decision making** February 2024. An OCLC-Liber blog post following their "Building for the future" program where librarians had gathered to discuss their understanding and practical interpretations of data driven libraries.
- **Utilising new modes of data to enhance research strategy and collaboration** November 2023. Digital science looks at how institutions can get high-quality insights into research by standardising their data systems.

Hands-on activity and other self-guided tutorial(s)

Now that you've found some data, and/or you've formulated some questions to answer, how do you know the data will be good enough? Let's take some time to assess the data, and ensure it is fit for purpose, because a key analogy in data analysis is;

```
**GIGO: Garbage in, garbage out!.**
```

The process of utilising data is a cycle, a standard process model that describes a common approach to data is CRISP-DM (Cross-industry standard process for data mining). It has six stages, listed below.



This section will enable you to assess your data, through a variety of leading questions, at each stage of the cycle. How long each phase takes will depend entirely on the project, there is no right or wrong, and it is cyclical, you will return to questions as you develop the project. These are structured as questions because one guide cannot have all the answers, and therefore giving a variety of questions which hopefully ensure you do not miss anything, and learn much more which is relevant to your project. This approach will enable you to assess if the data you have is relevant to the question, and if it has the potential to answer it correctly.

This guide hopes to calm your nerves about using data, and to see that you are capable of answering data questions confidently.

Business Understanding

The Business Understanding phase focuses on understanding the objectives and requirements of the project. Here you are planning your project

- What questions are you answering?
- Who are your stakeholders?
- What are the criteria and limitations to the project?
- What is the goal of using this data? Doing this project?
- What resources are (or are not) available?
- What is the timeline on the project? And the data input specifically?
- What stage of the project requires data analysis?
- What is the expected outcome of the data analysis?

Data Understanding

Data Understanding drives the focus to identify, collect, and analyse the relevant datasets.

- Can you access the data you require? What is the source?
- What data (and metadata) is missing?
- How reliable is the data?
- What relationships are relevant within the data?
- Who is involved in gathering and preparing the data?

Data Preparation

A common rule of thumb is that 80% of the project is data preparation.

- What data is not necessary from the dataset?
- How are errors or duplications handled?
- What new attributes or formulas are required?
- How do datasets interact with one another?
- Is the data formatted correctly?
- What acronyms are in use? Are they formatted correctly?
- Has the data been standardised?
- Are the column names useful and appropriate?
- Does every database have a unique identifier column?
- How do you handle missing metadata?
- Are you editing the master / only copy?
- Have you documented the process so that you can provide evidence if needed?
- Is the data in a clear and useable tabular format (rows and columns)?
- Have all merged cells been removed / updated?

Modelling

What is often regarded as the most exciting work is also often the shortest phase. Now is the time to build and assess various models or visualisations.

- What conclusions can be drawn from the data?
- What visualisations are appropriate to display the data?
- What comparisons and relationships should be highlighted to align with your initial goal?
- Does the data contradict the hypothesis? Why?
- Do the formulas need to be live? Tip: Live formula within a spreadsheet can cause it to be slow and large in size.

Evaluation

This phase looks more broadly at the data project and what to do next.

- Did the data fulfil the goal of this project?
- What was unable to be achieved?
- What is required for the next data project to run more efficiently?
- Has the project been completed successfully?

Deployment

Depending on the requirements, the deployment phase can be as simple as emailing a graph, or as complex as publishing a live dashboard of intricate data and visuals.

- Who requires the outcome of this project?
- Will the outcome need to be presented in multiple formats?
- Will this be repeated again? Has it been documented?
- Where will the data and outcome be stored?

Storing a datafile / spreadsheet

Once the data exists, it must be accessible. There are a variety of software which can store your data, and taking the time to assess the needs of your data is important.

- Is the file formatted correctly (CSV vs Microsoft Excel)?
- Will it need to be accessed online and/or offline?
- Does the file require a software licence to access?
- What backups are required?
- How long are files required to be stored? Use this webpage for guidance on different file formats.

Recommended Reading/Viewing

Defining Data Librarians

- Defining data librarianship: a survey of competencies, skills, and training (July 2018). Federer aims to define data librarianship by exploring the skills and knowledge that data librarians utilise and the training that they need to succeed.
- Introduction to Databrarianship: The Academic Data Librarian in Theory and Practice (2016). Thompson and Kellam explore the diverse field of data librarianship, highlighting its key commitment to accessible data.
- Data librarianship: a day in the life (2011). Interviews with data librarians across the world highlight the challenge and opportunities of data in libraries, including the creation of data services within an academic library.
- The future for numeric data services (2011). Exploring the future of data librarians, with trends in visualisation, mapping, standardisation, citation and data management plans.
- Librarian roles in the digital data-driven world (September 2021). Thai/English paper exploring the role of data librarians as someone who continues to acquire new knowledge and skills with technology and the digital era.
- **CILIP** (Chartered Institute of Library and Information Professionals) provides definitions on the data science roles available within a library.

Library Endorsed Resources

- **LIBER** members have created the Digital Scholarship & Data Science Essentials for Library Professionals, which this guide is a part of. Guidance also includes data visualisation and data as collections, and continues to grow.
- **JISC** has published resources to support data-driven decisions, including interactive insights on graduate outcomes and conducting online surveys as a form of data gathering.
- The **DSVIL** (Data Science and Visualization Institute for Librarians) provides great guidance and resources for finding, cleaning, analysing, visualising and managing data.
- The **Bodleian Library**, University of Oxford, hosts a variety of resources to support data analysis (including audio-visual), data mining, and visualisation tools.
- **Duke University Libraries** have a plethora of resources for data science, data management and data visualisation freely available.
- **ARL** (Associations of Research Libraries) demonstrate the impact data analytics can have within libraries and library communities.
- **IFLA** (International Federation of Library Associations and Institutions) has a variety of data analytics resources which cover using specific software, trends in the news, and practical applications of data within libraries.
- The **PLA** (Public Library Association) (which is part of ALA, the American Library Association), have developed data tools to enable public libraries to be compared across multiple metrics.

Learning Data Beyond Libraries

There are numerous platforms dedicated to data skills, listed below are some of the most popular for beginning your data journey.

- **The Carpentries**: Community-led coding and data science courses for researchers and librarians, with three specialties; Data Carpentry (data skills for conducting research), Library Carpentry (software and data skills for librarians), and Software Carpentry (lab skills for research computing).
- **Datacamp**: great for free cheat sheets (webpage or PDF) on data literacy, understanding data knowledge levels and data storytelling. As well as resources and courses for a variety of coding and data software.
- **LinkedIn Learning**: a wonderful resource of videos for anything from coding to people management to making the perfect Microsoft Excel graph.
- **Google Skillshop**: Training and certification in google analytics which could upport queries about library engagement.

Taking the next step

Ensuring Good Practice

Working with data is very exciting and rewarding, but can quickly become overwhelming. There are fears of being hacked and having data stolen, or even losing data due to bad organisation or not knowing it needed saving. Sadly this guide cannot have all of the answers on the practices and policies which should be in place. But it will instead ask a few provoking questions to ensure you can rest the worries of data loss and theft.

- Have you established a naming convention for files? Does it account for versions, dating and author? Tip: Have titles be meaningful, in a relevant order, and versions discernible.
- When will the file(s) be deleted or archived?
- Is there clear documentation of the processes undertaken? Therefore ensuring the task can be repeated, or to justify any conclusions. Tip: This should include the source of the data, any cleaning steps made, and all relevant metadata.
- Where are the files stored, are the folders easy to access? Is there a naming convention and folder hierarchy in place for the project, and the team?

What skills are you looking for?

If you have read through this guide, and still thirst for more, here are some questions to encourage your exploration in library data!

- What in-house data can you combine (such as the library management system and google analytics)?
- Think of a report that is repeated regularly but often time consuming, is there a software or data task that could make this quicker?
- What software do you regularly use, how could you expand those skills? Does the library management system provide specialist training?
- Is coding, AI or visualisation an area you would like to explore? What else do you want to do with data?
- Which elements of the data process do you want to focus? Cleaning, analysing, visualising etc.
- Can you attend a conference which discusses library data you are interested in?