# Chapter 1
# Exploring Data Science Initiatives Through an International Lens

**Nandita S. Mani**
https://orcid.org/0000-0002-0955-1066
*University of North Carolina at Chapel Hill, USA*

**Emily P. Jones**
https://orcid.org/0000-0002-4294-7564
*University of North Carolina at Chapel Hill, USA*

**Rebecca Carlson**
https://orcid.org/0000-0002-4380-8435
*University of North Carolina at Chapel Hill, USA*

**Fidan Limani**
*Leibniz Information Centre for Economics, Germany*

**Atif Latif**
*Leibniz Information Centre for Economics, Germany*

**Klaus Tochtermann**
*Leibniz Information Centre for Economics, Germany*

**Faten Hamad**
https://orcid.org/0000-0001-7548-0529
*The University of Jordan, Jordan*

**Christine J. Urquhart**
*Aberystwyth University, UK*

**Victoria Lemieux**
*The University of British Columbia, Canada*

**Sarah Ames**
*National Library of Scotland, UK*

**Jenna Bain**
*State Library of New South Wales, Australia*

**Justin M. Clark**
*Bond University, Australia*

## ABSTRACT

*Application of data science tools, techniques, and principles have increased within the field of library and information science. This trend is especially noticeable in academic libraries where strategic priori-*

DOI: 10.4018/978-1-7998-9702-6.ch001

*ties around data science initiatives have been created to further support and add value to the research enterprises at their institutions. This chapter seeks to highlight a global outlook on how data science has been addressed in library and information studies via case studies in areas including digital humanities, machine learning, and visual analytics. Increasing awareness of how partners across the globe are addressing data science needs at their institutions can help raise visibility of how data science can be infused and utilized within a variety of contexts.*

## INTRODUCTION

This chapter provides a global overview of current work in data science (DS) by library and information science (LIS) stakeholders. It includes case studies from the State Library of New South Wales in Australia, Leibniz Information Centre for Economics in Germany, National Library of Scotland, Institute of Evidence-Based Healthcare at Bond University in Australia, and the University of Jordan. These examples of DS work in LIS settings are contextualized within a review of the growth of this field, discussion of the challenges with global DS projects, and future recommendations to increase information on and access to DS projects around the world. The chapter objectives are to enable discussion on the current state of DS in LIS, identify potential next steps in furthering awareness of DS activities, and provide librarians and information professionals with examples of how others are applying DS in various environments for diverse aims.

### Background

Interest in application of DS principles within the field of LIS has increased significantly in recent years (Virkus & Garoufallou, 2020). The shared values, goals, and expertise found within DS and LIS has led to collaborations, research, educational programs, and knowledge sharing across institutions and disciplines. For example, data management, information organization, and information access are shared priorities of both disciplines (Cox et al., 2017; Semeler et al., 2019) and researchers in LIS are combining trends in big data and data-driven analyses with library collections, information systems, and data management (IFLA, 2018). The DS field is naturally collaborative and interprofessional, as researchers with computer science, software engineering, and statistical expertise partner to explore techniques and applications related to big data. LIS scholars are adding expertise to DS projects in the areas of knowledge management, data curation, data management, health sciences information, and education and training on tools and techniques (Virkus & Garoufallou, 2020).

As DS has grown within LIS scholarships, some academic libraries have created strategic priorities around DS initiatives to further support and add value to the research enterprises at their institutions (Burton et al., 2018; Mani et al., 2020). For example, a 2020 environmental scan of United States universities found that over 60 institutions with varying degrees of DS and academic library partnerships, including schools identified as leaders in library-DS collaborations, have created research centers, programs, divisions, schools, and services which leverage the expertise of data scientists and university libraries and librarians (Mani et al., 2020). Additionally, academic libraries have long been responsible for creating and maintaining institutional repositories; these repositories often contain datasets as well as records of scholarships, which are valuable to those engaging in data-related inquiry. As academic libraries partner on DS programs and initiatives, roles for LIS professionals specific to big data and DS are also being added, including data librarians, archivists, curators, and analysts (IFLA, 2018).

Both within and beyond libraries, LIS educators are contributing to DS initiatives by providing education and development opportunities for trainees and professionals interested in gaining knowledge and expertise with DS tools and techniques. Ample evidence of librarian-led classes and workshops specific to data management and science have been reported, including but not limited to introductory principles to hackathons and computer programming with open software (Briney, 2017; Burton et al., 2018; Erdmann, 2015; Miller, 2018). Additionally, in response to the further integration of DS and librarianship, many LIS schools are revising their curricula to include specializations or certifications in DS (Ortiz-Repiso et al., 2018). In their 2018 study, Ortiz-Repiso and colleagues found that over 50% of LIS schools, or iSchools, internationally offered data-related education, and some even offered data-focused degrees.

Despite many examples of projects centered around library resources and common interests between the LIS and DS communities, scholarly literature on real-world examples of DS applications produced by libraries is scant. Research on applications within libraries are often led by computer scientists or those trained in disciplines outside of LIS; however, there are more opportunities for collaboration with data librarians or other librarians who are DS experts (Al-Barashdi & Al-Karousi, 2019; Blummer & Kenton, 2018; Federer et al., 2020). Accordingly, not much is known about how librarians are using DS internally, in their own practice and research, as opposed to how librarians are using DS skills to further the work of their constituents and how researchers in information, computer science, and other disciplines are using DS to study library collections, resources, and services. Librarians are partnering with researchers, teaching students, and providing services to enable DS projects, as they enable many other types of research by their partners and communities, but these projects are not internal, or library-centric. Examples of academic libraries offering DS instructional workshops and researcher support are available on a global scale from institutions in Estonia (University of Tartu Library, n.d.), Ethiopia (EIFL in Ethiopia., n.d.) and across Europe (LIBER, n.d.) to the United States (University of California San Francisco, n.d.; University of Miami Libraries, 2021; University of Virginia Library, 2022)

Many published examples of library-based DS originate from the United States, from the institutions above as well as many others, despite the international uptake of interest. The predominance of scholarly discussion stemming from institutions based in the United States highlights a lack of diverse voices, perspectives, and uses of DS in the field. Thus, while there are some examples of ongoing DS research and practice being led by information professionals in various contexts from around the world (Bain, 2020; Chiware & Mathe, 2015; Hamad et al., 2020; Hutchinson, 2018; Limani et al., 2018), the extent to which this is occurring is unknown. It is also not currently known if there is a difference between the DS initiatives being published in different countries, in response to different institutions, information, data environments, and user needs. This chapter, while not able to answer questions of prevalence or provide a comprehensive picture of the scope of global DS research, provides global examples of how librarians are addressing and using DS at their respective institutions to improve their collections, resources, services, and advance research inquiry.

Case studies included provide context and perspective through real-world examples of the application, implementation, assessment, and barriers of DS projects in libraries or being undertaken by researchers at LIS schools. Case studies cover a range of topics, including digital humanities, machine learning and automation tools, visual analytics, and collections as data. Examples come from countries around the world, including Australia, Canada, Germany, Jordan, and the United Kingdom. Together, these case studies enable discussion on the current state of DS in LIS, potential next steps in furthering awareness of DS activities, and provide librarians and information professionals with examples of how others are applying DS in various environments for diverse aims. Following the case studies, a discussion related to the challenges to implementing and identifying LIS DS projects globally and future recommendations are provided.

## INDIVIDUAL CASE STUDIES

### Case 1: Taming the TIGER: Using Machine Learning Methods to Rediscover the State Library of New South Wales' Digital Collections

### Project Summary

The State Library of New South Wales (NSW) is Australia's oldest library established in 1826. It has over five million digital files in its collection and this number grows every day. With such an immense archive, it is impossible to catalog everything in detail, meaning that much of the rich information within the collections remains unexposed and difficult for library users to find. To help tackle this problem, the library began experimenting with the use of machine learning techniques to better describe its collections, starting with the use of existing cloud computing services in the market to create descriptive keyword tags for its digital materials. Through this experimentation, it became apparent that these services alone were not precise enough to accurately describe an archive with such a great variation in collection material and format types. As a result, the library developed its own custom algorithm called Tagging Images Generically for Exploration and Research (TIGER) that analyzes and synthesizes raw data created by multiple tagging services in order to create a shortlist of more accurate, relevant, and useful tags for every image in its collection. The library has also begun experimenting with ways to create more structure with the keyword tags identified through this process by applying additional machine learning techniques to build hierarchical taxonomy structures of similar and related keywords. The aim of this secondary phase is to build more meaningful relationships between machine-created metadata, and to improve the overall discoverability of the library's digital collections. These taxonomy structures are based on existing trusted taxonomy thesauri commonly used within the library sector, such as the library of Congress' *Thesaurus for Graphic Materials*.

### Project Highlights

- Development of custom algorithm that synthesizes keyword tagging results from commercial, off-the-shelf cloud computing services to create more accurate and useful tags for millions of digital files.
- Experimentation with building taxonomy structures to link synonyms and related tags created from the machine learning process to support greater browsability and discoverability of digital collections.

*Table 1. Data science skills, workflow, and program outcomes*

| Data Science Skills and Workflow | Program Outcomes |
|---|---|
| ● Use of commercial tagging services to tag digital images with keyword tags. Services used for project:<br>○ Amazon® Image Recognition<br>○ Google® Cloud Vision<br>○ Microsoft® Computer Vision<br>● Development of custom Python algorithm that synthesizes raw data from tagging services to create a shortlist of tags. The custom algorithm considers factors like:<br>○ Confidence level ratings for each tag<br>○ Repetition of tags (exact matching)<br>○ Average confidence level matching across services<br>● Use of Natural Language Processing techniques like lemming and stemming to identify synonyms, exact matches, and related terms within generated tags, which are then mapped to a hierarchical taxonomy structure. | ● Create keyword tags for digital files within the library's collections to expose greater level of detail within the archive, making digital content more discoverable.<br>● Identify detail within library's digital collections using alternative methods to traditional cataloging practices, creating greater detail in file level metadata across the collection, and ultimately exposing new and different layers of meaning and information about the library's important historical archives.<br>● Build greater confidence in the overall accuracy and relevancy of tags through process of analysis and synthesis.<br>● Provide access to synthesized tags to users through library's online digital collection interface. Tags are searchable and filterable.<br>● Create more structure for keyword tags by building taxonomy trees that group synonyms and similar tags together.<br>● Provide an interface that allows users to browse the created taxonomy trees as an alternative method for exploring the library's collections. |
| **Data Science Tools Used**: Amazon® Image Recognition, Google® Cloud Vision, Microsoft® Computer Vision | |

## Impact and Lessons Learned

This project has been highly experimental for the library and has exposed the benefits of metadata creation through machine learning techniques. It has also exposed some of the limitations of relying wholly on any commercial service to describe heritage collection material. By building a custom algorithm to complement these off-the-shelf tagging activities, the library has found a way to programmatically synthesize data and files on a large scale but, at the same time, build greater confidence within the organization in the quality of metadata being created about its collections. The scope of the original project, which began as a simple exploration of automated image tagging, saw significant organic growth as learnings were gained through practice. These learnings ultimately created opportunities to refine collection metadata, ensuring that data outputs would be useful, interesting, and relevant to library users and researchers. Most importantly, this project exposed levels of detail within the State Library's digital collection that had been previously difficult or impossible to find.

Work on this project is ongoing and machine-created metadata is continually being refined, added to, and evaluated to ensure its ongoing accuracy and relevancy.

The approaches to DS applied throughout the TIGER project demonstrate the potential for enormous value to all types of libraries and educational institutions. Rather than building and training custom tagging algorithms from the ground up, this hybrid methodology has allowed the State Library of NSW to benefit from well-tested and well-resourced industry expertise and, at the same time, still manipulate the end results to best suit its collections and audiences. Any library could follow a similar approach by exploiting existing machine learning services that are both accessible and affordable, and integrating them with specific thesauri or vocabularies relevant to the landscape in which they operate.

## Case 2: Exploring the Adoption of Knowledge Graphs by a National Economics Library in Germany

### Project Summary

With more open and collaborative scholarly practices, there are an ever-increasing number and variety of published research artifacts across domains. In addition, the scholarly communication process is constantly evolving to include these artifacts, such as research data, software, scientific blogs, citation links, and so on, that, in a way, present a part of a research. Correspondingly, requests to access these artifacts are to be expected and becoming the norm.

In the context of library communities, different stakeholders have expressed interest to explore and eventually integrate a broader set of research artifacts alongside publication collections, still the staple of library holdings. Either this is often from the individual resource's perspective (i.e., include a new type of artifacts as part of the catalogs), or from the perspective of a more complete or complementary view of research (i.e., provide as many facets of the research as possible as the different research artifacts provide (all) the elements to better understand a research contribution). Recently, Knowledge Graphs (KG) have been applied as a means to bring together data from heterogeneous sources, within or across multiple domains and, in many cases, has shown to be an effective approach.

Due to the variety of definitions and technologies for KGs, it is important to define the meaning and implications of a KG, as adopted in this work. Generally, a KG includes a a). structured definition (even a semantic description) of resources, and; b). the links that these resources (naturally) form between themselves. The scholarly communication domain is suitable to this technology for the following reasons:

1. Research artifacts are typically described via metadata, and a more structured description would enable machines, in addition to humans, to (re)use these resources.
2. There are natural links between research artifacts that form a (research artifacts) graph. Consider for example a paper that links to a dataset as its primary data source, a paper citing another paper, or a dataset reusing another dataset. All of these links, either as part of a collection, or identified and added later, help shape this graph.

As shown in Figure 1, source metadata is ingested from different artifact collections, pre-processing takes place (i.e., adding subject terms from a thesaurus), descriptions based on ontologies are enriched and stored as a Resource Description Framework (RDF) collection in a graph database. Then, the team queries this database to explore use cases that involve this variety of research artifacts.

At this experimental stage of research, this work focuses on providing more holistic access to multiple research on such artifact types. The point of reference being the Leibniz Information Centre for Economics (ZBW) in Germany, one of the largest libraries in the world for economics publications. Specifically, the team relies on the collection of publications from one of its publishing platforms, and the dataset collections from some of the projects conducted at this institution.

To this end, the team is exploring KGs as a means of integrating heterogeneous resources, including a common representation model for all of the artifacts, proposing its components, instantiation, and exploring the potential benefits (in terms of use cases) that it will bring to a library. The role of the KG is in both enabling a common (semantic) representation and storage of the different artifacts, and providing access to the resulting collection via different services, such as search, recommendations, etc.

The differences in artifact types bring challenges, including their harvesting, pre-processing, representation (ontology or vocabulary description), enrichment, and so forth. In addition, specifics of the work include bringing research artifacts that are not common in the existing public collections to the library and broader community. For further information about the Leibniz Information Centre for Economics, go to https://www.zbw.eu/.

DS skills and workflow in addition to program outcomes are provided in Table 2.

## Project Highlights

- Identification, harvesting, and pre-processing of scholarly artifacts that complement library holdings.
- Adoption of KG as a means to bringing these research artifacts into a common model and representation.
- Identification and exploration of use cases based on these artifacts' heterogeneity for the library domain.

*Table 2. Data science skills, workflow, and program outcomes*

| Data Science Skills and Workflow | Program Outcomes |
|---|---|
| ● Harvest, parse, and pre-process a variety of artifact collections.<br>● Artifacts semantic modeling: Select the appropriate ontologies/vocabularies to represent the scholarly artifacts.<br>● Convert the research artifacts to Resource Description Framework (RDF) based on the ontologies/vocabularies from the previous point.<br>● Organize the resulting RDF collection—based on provider, publisher, etc.—to easily maintain, track changes, and store it, including provenance information. Also, make the RDF metadata collection publicly available. | ● Structured different research artifacts and brought them in a common, machine-actionable representation—RDF.<br>● Converted more than 108,000 open access publications, 1.1 million dataset metadata, 126 million citation links, and 8,500 blog posts and included them as part of the KG undertaking.<br>● Applied enrichment to the collection—to the extent possible given the different artifacts. This included the assignment of terms from a thesaurus from the economic domain, harmonization of the language tags used across collections, geo-location information, etc.<br>● Explored several use cases and scenarios over the collection as a whole (cross-artifacts scenarios), as well as scenarios focusing on individual artifacts. |
| **Data Science Tools Used**: Semantic Web technology stack for modeling and representation (RDF, ontologies/vocabularies); The Apache® Jena™ Framework for Semantic Web application development | |

## Impact and Lessons Learned

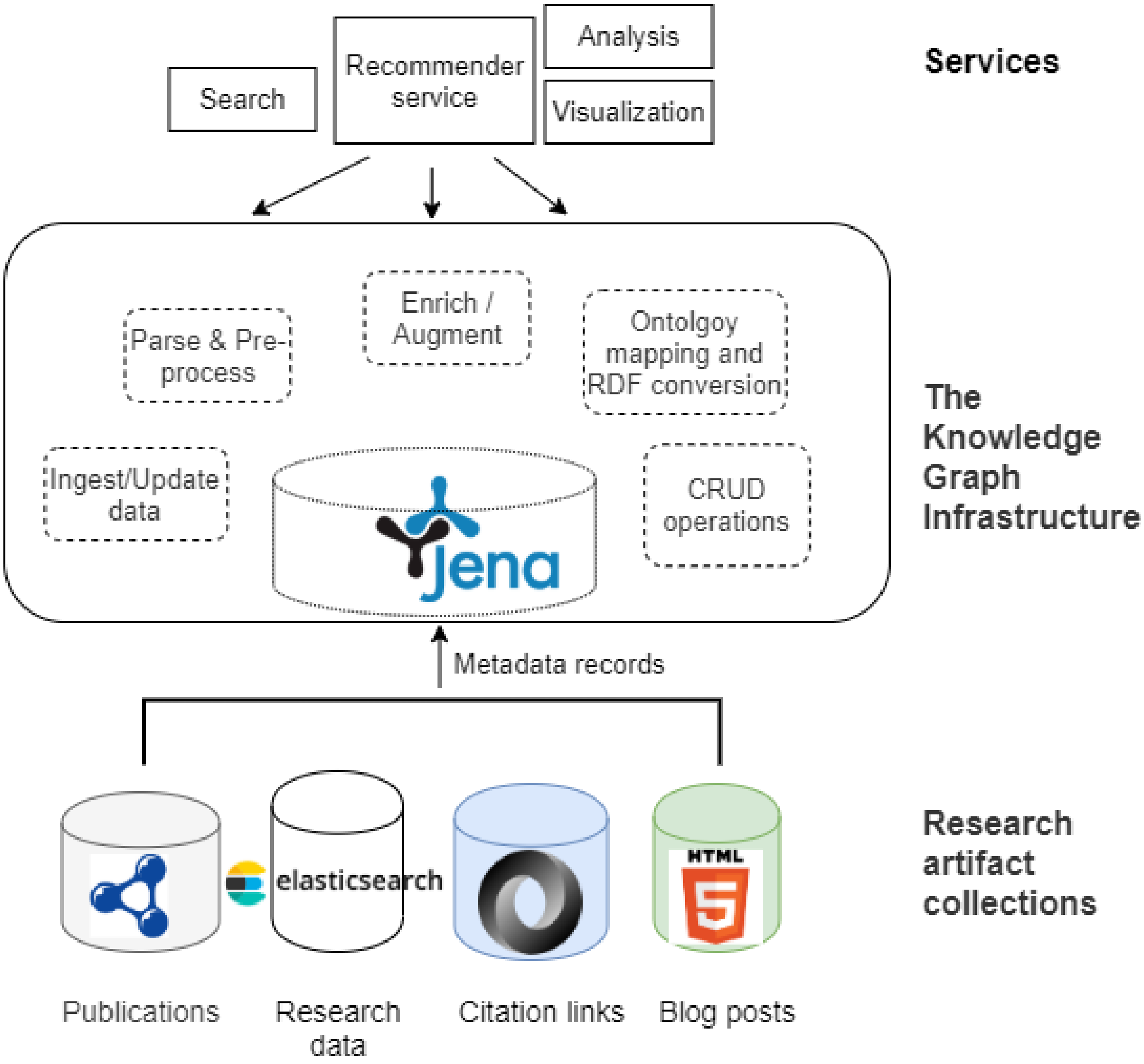
In this ongoing work, as a proof of concept, the KG ingests research artifacts of different types in a more compact research "whole" via the application of semantic technologies. This allowed exploration of use cases, such as a search that included cross-artifact scenarios, to the extent possible (due to the nature and metadata description of the different artifacts). To libraries, this highlights a potential value added for the users by including artifacts from the domain of choice that are complementary to their more established artifact collections. On the another hand, new research artifacts collections can also provide value on their own, without necessarily complementing another library collection.

As the variety and volume of new research artifacts increases, so does the requirement for using the appropriate tools to handle them. Exploring new or emerging technologies is always an opportunity to assess the possibilities to tackle the challenges that stem from this development. When it comes to bringing complementary resources into a common representation and enabling access to it as a whole, KGs could be one of the options to consider. Based on the team's experience in the library domain, KGs proved to be a suitable solution that scales relatively well (in terms of ease of modeling and schema definition) as new resources are targeted. This, in turn, enables a variety of services (e.g., analysis, search, recommendation) to grow at users' request to support data analysis needs at an institution.

One of the promising outcomes of this ongoing project is that the linking of scholarly resources through KGs can be a valuable driver in improving information infrastructures, such as Digital Libraries. However, engaging in such an endeavor requires experience and often faces challenges. The following is the description of lessons learned in terms of the team's experience, which drives research further.

- **Promising Research Direction**: After bringing together heterogeneous collections, KGs seem a promising venue for libraries to adopt as they offer the means for DS analysis for a larger and more diversified set of research artifacts. This will require conducting a user study to analyze this aspect, including obtaining feedback from a variety of library stakeholders.
- **Holistic Access to Research***:* Having access to a heterogeneous collection offers a more holistic approach to research. Namely, there are already artifacts that are or are becoming part of scholarly communication that could enhance access to or understanding of research. Given their variety, one has the possibility to better understand a research project, including its (re)use—full or partial, targeting only certain artifacts.
- **Extensibility:** Scholarly communication is constantly changing. As new artifacts reach a critical mass or are otherwise required (by funders, for academic credit, etc.), the adopted model should be able to accommodate them. The model does not require extensive planning in anticipation of potential artifact types in the future. Even before any new artifacts get selected to be included in the KG, the model should support eventual extensions whenever they happen relatively seamlessly. KGs do not require a fixed schema, planned for before a single instance is added to the KG. Rather, as the requirements change, new artifacts can be added (in RDF, in this case) to the same model without much difficulty.
- **Artifact Collections Often Require Dedicated "Attention":** Not all are published according to the same publication practices, thus one needs to investigate them before including them in the collection.
- **Scaling:** Developing dedicated "ingest" operations is required to deal with all sorts of aspects, starting from access (Web scraping, API; different formats, etc.), to modeling (identifying and applying appropriate semantic models), to enrichment (is the artifact metadata rich to support this operation? Are there relevant resources to match them with that are already published?), to conversion (often the larger collections require certain machine configurations to process), to publication (publish the resulting collections in a convenient way for the users to access), etc.
- **User Interface:** While querying the collection via SPARQL works, in this case, a user interface is the preferred means to access a collection, across the board. To this end, further work is needed to experiment and include appropriate user interfaces that can support the use cases enabled by the integration of all of the different artifact types.

*Figure 1. KG components to ingest, pre-process, enrich, describe based on ontologies, and be stored as an RDF collection. A set of services would be provided for the users to (use/re-use the RDF collection)*



## Case 3: The Data Foundry: Supporting Data Science Initiatives With Cultural Heritage Collections at the National Library of Scotland

### Project Summary

The Data Foundry is the National Library of Scotland's open data-delivery platform to support DS, digital scholarships, and other computational uses of the collections. Part of the library's Digital Scholarship Service, which was established in 2019, the Data Foundry hosts the library's collections as data—from digitized material to metadata, organizational data, and spatial data. It also enables users to analyze some datasets using Jupyter Notebooks, and features collaborations and external uses of the collections. The Data Foundry has three core principles—data collections on the platform are:

- **Open:** The National Library of Scotland publishes data openly and in re-useable formats.
- **Transparent:** The provenance of data will be taken seriously and explanations around how and why they have been produced will be provided.
- **Practical:** Datasets are presented in a variety of file formats to ensure that they are as accessible as possible.

Collections are published on the Data Foundry with the intention that users can repurpose and recast Scotland's national collections. These include datasets such as the first 100 years of Encyclopedia Britannica, LiDAR point-cloud data of the library's main building, and expenditure information from the library's accounts.

Establishing and maintaining the Data Foundry has involved culture change within the library, new collaborations both internally and externally, and new ways of thinking about the collections. Turning cultural heritage collections into data is a whole-library effort—particularly for datasets relating to digitized material. This spans the entire process from selection for digitization, digitization itself, Optical Character Recognition (OCR)/Handwritten Text Recognition (HTR), creation of a variety of file formats, packaging up the dataset, and assigning a Digital Object Identifier (DOI) to publishing online. Furthermore, it involves rethinking the idea of the "collection" itself and considering what constitutes a "dataset." Further information about the Data Foundry can be found at https://www.data.nls.uk/.

## Project Highlights

- Creating an open data platform for the National Library of Scotland's collections.
- Enabling collaborative projects with academicians, students, and creatives, which are featured on the platform.
- Delivering "collections as data" is a whole-library effort, involving cross-team collaborations and changes to workflows.

## Impact and Lessons Learned

The launch of the Data Foundry in 2019 received international interest, and the website and datasets have been in constant use in teaching, learning, and research around the world ever since, as well as being used by artists. This means that there are a number of case studies exploring external uses of the collections featured on the Data Foundry's "Projects" page, which is available at https://data.nls.uk/projects/. Internally, it has meant that the library has altered digitization workflows to enable the creation of datasets from newly digitized material. Teams from across the Library now support the Digital Scholarship Service and the Data Foundry, from developers to colleagues working in social media, who promote new datasets and projects to the library's followers.

Publicity work around the Data Foundry has also enabled the library to bring the idea of "collections as data" to audiences beyond academia, who may not expect libraries to engage in the world of DS, by publishing Twitter "threads" detailing how a dataset is created, or what HTR is. Recognizing that not all users have the skills to work with large cultural heritage datasets, the library's Digital Scholarship

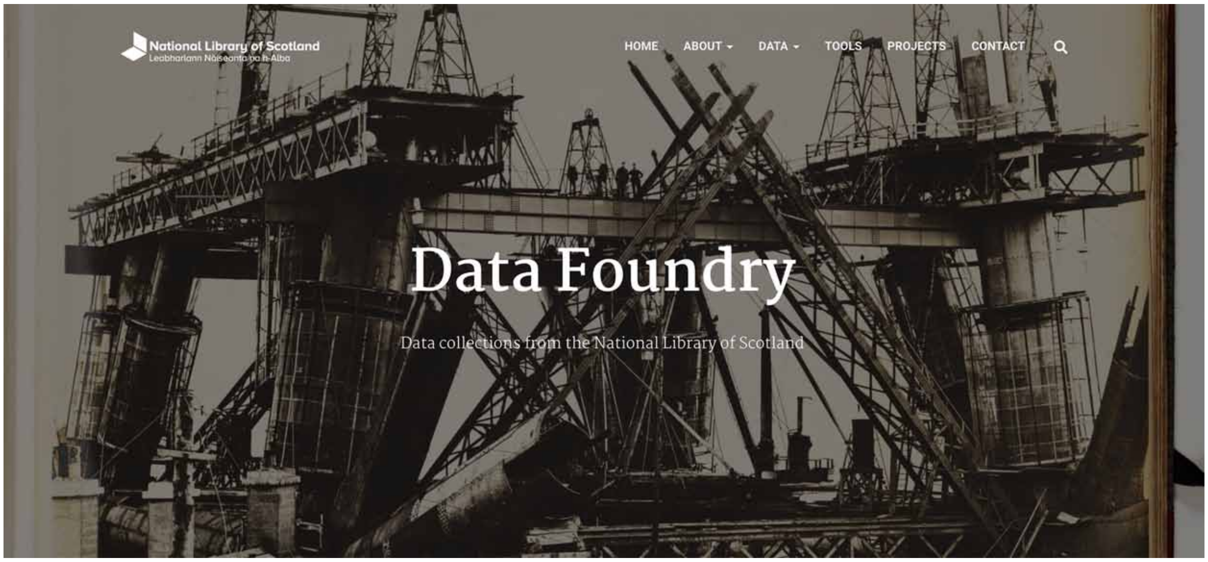
Service has also released a number of Jupyter Notebooks exploring some of the collections, enabling anyone to dig deeper into the datasets.

Delivering collections as data involves a number of challenges, including copyright and licensing - some collections simply cannot be made available as datasets for reuse under open licenses, for example. Identifying collections, which are suitable to be released under an open license can be time-consuming and painstaking work. Technical decisions around file formats to make available and agreeing on standards for datasets can cause issues when there are legacy collections, which have digitized to different standards. Furthermore, it requires consideration about the end-user, and how to make large files available in an accessible way.

All of this points to the need for library resources, from developers to curators and copyright specialists, to be directed towards digital scholarship and DS, which is not always feasible. In the case of the Data Foundry, this sometimes leads to delays between digitization and ultimate publication as a dataset, and means that in-house experimentation with datasets is limited to when time allows. However, the Data Foundry continues to be regularly updated with new datasets and projects, which underpin the library's Digital Scholarship Service offerings.

While the Data Foundry (see Figure 2) is a data-delivery platform from a national library, it currently largely serves an academic, research audience, with future plans to expand the library's Digital Scholarship Service in a more systematic way towards a public audience including schools, businesses, and creatives. Publishing collections as data in this way enables libraries--whether national or academic—to open their collections to new research questions through new methods and technologies.

*Figure 2. Screenshot of The Data Foundry*



Data Foundry

Data collections from the National Library of Scotland

## Case 4: Automation Tool Development for Improving Systematic Review Speed - A Case Study from Australia.

### Project Summary

The Institute of Evidence-Based Healthcare (IEBH), based at Bond University on the Gold Coast in Australia, is a world leader in evidence synthesis, which conducts many systematic reviews (SRs) every year. As SRs are very resource intensive, they take, on average, five staff, 67 weeks to complete, at an average cost of USD $141,000. In 2016, the IEBH launched the Systematic Review Accelerator (SRA) as shown in Figure 3. Additional information about the Systematic Review Accelerator can be found at https://sr-accelerator.com/.

The SRA is a suite of in-house, custom-built, systematic review automation tools, freely available for anyone in the world to use. Initially the SRA focused on the searching tasks of an SR, which required significant involvement from the IEBH's Information Specialist. This led to the SRAs most used tool, the Polyglot Search Translator. Further details about the Polyglot Search Translator can be found at https://sr-accelerator.com/#/polyglot. Since the development of the Polyglot Search Translator, the SRA has expanded outside of searching, to other SR tasks. The IEBH's Information Specialist now runs the automation team and has overseen an increase in usage from 12,000 pageviews in 2016, to over 100,000 pageviews in 2020, totaling nearly 400,000 pageviews since its launch. This has primarily been achieved by presenting at international as well as national conferences and by running training sessions in collaboration with professional bodies, such as the Medical Library Association (MLA) and Health Libraries Australia (HLA). The SRA has also been one of the key components of the IEBH's successful Two-Week Systematic Review (2weekSR) program. This is a program which combines improved workflows, multidisciplinary teams, and automation tools to reduce the time needed to complete an SR, in most cases down to two weeks. More information about the 2weekSR program can be found on the IEBH website athttps://iebh.bond.edu.au/education-services/2-week-systematic-reviews-2weeksr.

### Project Highlights

- The IEBH Information Specialist is heavily involved in SRs, especially 2weekSRs, and uses the SRA tools to create highly focused searches in extremely short time frames.
- The SRA tools reduce the workload of research teams conducting SRs, and are designed in a way to highlight this, directly allowing SR teams to see the impact an information specialist can make om the production of an SR.

*Table 3. Data science skills, workflow, and program outcomes*

| Data Science Skills and Workflow | Program Outcomes |
|---|---|
| • The SRA is a suite of tools used to help accelerate the speed of production of SRs; four of these tools are designed for use by information specialists.<br>• The first search tool is the Word Frequency Analysis Software, which counts the number of times words appear in the title, abstract, and keywords (subject terms) of key, user selected references.<br>• The second tool is the SearchRefinery, which creates a visual representation of the search string and maps it to key, user-selected references.<br>• The third tool is the Polyglot Search Translator, which translates search strings across multiple, commonly searched databases.<br>• The fourth tool is the Deduplicator, which identifies and categorizes duplicate references allowing for their easy identification and removal. | • To date, ten 2weekSRs have been produced using the SRA tools.<br>• This tool allows the creation of basic search strings in a fast timeframe.<br>• The SearchRefinery helps create focused search strings, reducing the screening burden on SR teams.<br>• The Polyglot greatly improves the speed with which searches can be run, allowing a SR to progress rapidly.<br>• The Deduplicator increases the speed with which deduplication can be done, allowing the SR to progress forwards at a much faster pace. |
| **Data Science tools used**: The Systematic Review Accelerator (SRA) | |

## Impact and Lessons Learned

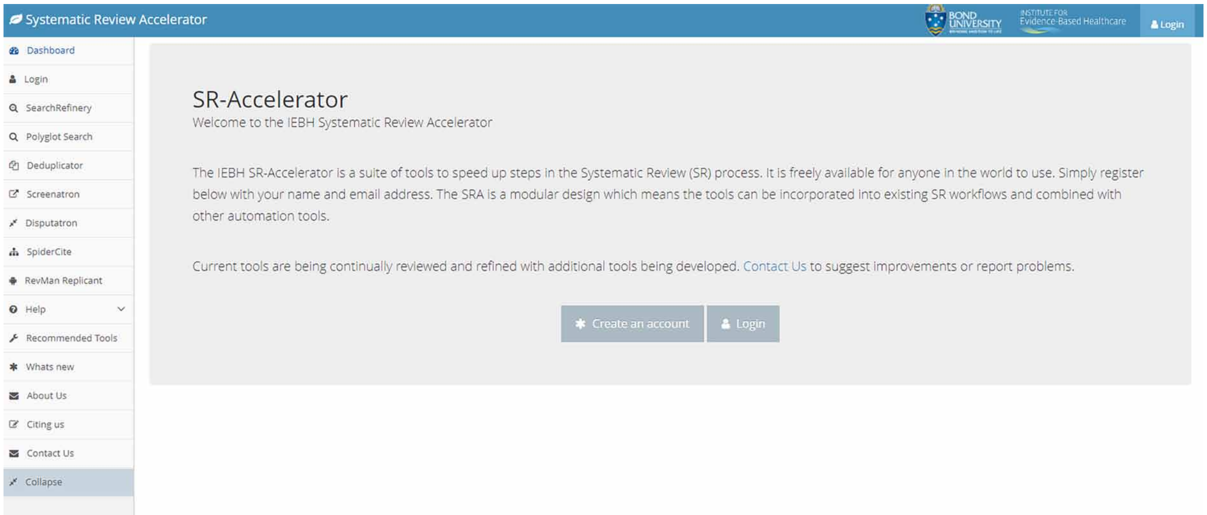
The creation and use of the SRA tools, especially those focused on searching, has allowed the IEBH Information Specialist to play a key role in the development of improved methods for the conduct of SRs. The IEBH Information Specialist, along with colleagues from the IEBH, together developed and published their novel 2weekSR methodology (Clark et al., 2020).

The SRA tools are designed to help skilled professionals use their skills in a more productive and efficient manner. This was, and still is, an important design principle of the SRA. These tools require skilled and experienced people to use them. Demonstrated throughout training sessions and demonstrations has been that use of these tools without proper skills and training does not help, and can, in fact, hinder performance. However, in the right hands, these tools can substantially enhance the role people, such as information specialists, can play in the development of evidence syntheses.

In the early days of development, designing and implementing the SRA tools came with some challenges. Many of the earliest SRA tools were focused on the search task of a SR. Initially, it was difficult to gain interest in investing in the automation of search tasks. Fortunately, grant money provided by Health Libraries Australia (HLA) allowed the initial development of some of these search tools. Their quick adoption by the information specialist community clearly showed this was an area of underserved need. This money, combined with heavy usage statistics, provided the impetus needed for continued development of automated search tools in the SRA.

The reach and impact of these tools has surprised even the author. As well as winning multiple, national Australian health library awards, they have been used to support information professionals all around the world. For instance, the oldest and most popular tool, the Polyglot Search Translator has been accessed over 12,000 times in 2021. While the newest tool, the Deduplicator, released September 2021, was accessed 1,300 times in its first month.

*Figure 3. Screenshot of the Systematic Review Accelerator dashboard*
*Note: For further information, go to https://sr-accelerator.com/*



SR-Accelerator

Welcome to the IEBH Systematic Review Accelerator

The IEBH SR-Accelerator is a suite of tools to speed up steps in the Systematic Review (SR) process. It is freely available for anyone in the world to use. Simply register below with your name and email address. The SRA is a modular design which means the tools can be incorporated into existing SR workflows and combined with other automation tools.

Current tools are being continually reviewed and refined with additional tools being developed. Contact Us to suggest improvements or report problems.

## Case 5: Digital Preservation System Architecture to Support Business Analysis Using Big Data Analysis in Academic libraries in Jordan

### Project Summary

This case study presents some hypothetical use cases for digital preservation activities that focus on the needs of researchers in big data analyses at academic libraries in Jordan.  This case study builds on preliminary work on preservation at the University of Jordan library. The digital preservation process there mainly concerns the preservation of theses and dissertations in the University of Jordan digital repository. Documents, which come from all the Arab universities in the region, are processed and deposited in the University of Jordan digital repository. The University of Jordan library was selected by the Association of Arab Universities, and since 1986, has been the Theses Deposit Center for the Arab region (Library of Jordan University, 2017), which makes it the main center responsible for handling the preservation of theses and dissertations in the Arab region. This also imposes a leading role for handling research data associated with theses and dissertations. Managing research data associated with theses and dissertations is beneficial for researchers, data repositories, the scientific community, and the public. It facilitates and encourages more connection and collaboration between scientists. It allows researchers to share resources, build upon other works rather than repeating their work, and inspires innovation in research (Hamad et al., 2021). The repository holds more than 85,000 theses and dissertations. Theses are deposited either by acquisitioning the CD or by uploading the theses to the webpage provided by the University of Jordan (Library of Jordan University, 2017).

The proposals are based on the preservation architecture developed through co-operation between Cambridge University Library, the Bodleian Library Oxford University, and informed digital image management at the Wellcome Collection, UK (Gerrard et al., 2018). Scalability is a big problem for digital preservation for the management and future use of the digital collection. The University of Jordan is likely to face future challenges in making the research data associated with individual doctoral theses and dissertations available for re-use, particularly when their collaborating universities may have different research data management plan requirements for their students. Therefore, this project aimed to develop a use case for the transformation to open data in the collaborative university repository.

First, defining what is meant by use cases is needed. In object-oriented systems analysis and modelling, use cases describe the interaction between the users of the system and the high-level functions within a system. The concern is what is done, the goal of the use case, and the steps involved, the actors, and the system responses. The use case specifies the functionality of the system as the user sees it, and helps analysts working with users to define the requirements (Bittner & Spence, 2003; Oracle, 2007). An example of the development of a use case diagram for a collection management function is discussed by Urquhart and Tbaishat (2017). The model presented by Gerrard et al. (2018) presents a scalable preservation architecture with various components.

## Project Highlights

- Conducted an initial "big picture" by interviewing the employer at the Computer Applications Section.
- Reviewed Theses/Dissertations submission process.
- Observed the digitization process and digital preservation process.
- Reviewed metadata extraction, assignment process, and metadata database construction for metadata preservation.

*Table 4. Data science skills, workflow, and program outcomes*

| Data Science Skills and Workflow | Program Outcomes |
|---|---|
| • Management of processes to handle date-related file metadata.<br>• Scheduling engine.<br>• Global identification engine.<br>• Metadata indexing engine.<br>• Risk engine, using business analytical data from processes such as file format recognition.<br>• Risk management. | • Use of PREMIS** preservation metadata standards.<br>• Multiple instances of the self-contained micro-service processes to be generated on demand and run in parallel.<br>• Sets of IDs for digital resources to allow definition of clear subsections of the collection.<br>• Free-text search of resources, provide a source of Big Data for research.<br>• Highlighting risks and possible mitigation measures.<br>• Ensuring data privacy and confidentiality guidelines are followed. |
| **Data Science tools used**: Microsoft® Excel®, Oracle® DBMS and SQL™, Microsoft® Visio™<br>**PREMIS** is PREsevation Metadata: Implementation Strategies (https://www.loc.gov/standards/premis/) | |

## Impact and Lessons Learned

The requirements listed under Data Science Skills and Workflow complement the three major components of the "big data" preservation system architecture: multi-tiered storage; metadata database; and a reporting warehouse for business analytics. Surrounding the core are the common microservices associated with preservation. Gerrard et al. (2018) emphasize the importance of defining where the boundaries around digital preservation and access lie—at organization, research domain, and sector levels. The OAIS standard refers to an archive consisting of an organization of people and systems, with responsibility to preserve information and make it available to a designated community. The emphasis is slightly different although both approaches stress the importance of metadata.  The PREMIS is the international standard for metadata to support the preservation of digital objects to ensure long-term usability of the preserved objects (The Library of Congress, 2021). Additional information on OAIS is available at http://oais.info. The transformation to open data in the collaborative university repository is likely to require:

- Survey of the types of research datasets within doctoral theses/dissertations and where these are presented (e.g., in appendices, in the text, etc.) Schöpfel et al. (2015) point out the problems of privacy and third-party copyright that might be revealed, as well as inadequate descriptions of data and data sets, lack of structure and organization of the data, and inadequate format (e.g. data and text glued together in a pdf file).
- Discussions over the management of the thesis/dissertation as a distinct information object—and the associated metadata, and/or separation of the research data from the text, with possible links to the data repository from the electronic thesis/dissertation metadata.
- Redesign of processes and workflows, as digital preservation for big data requires more input at the initial stages of research proposal approval (generally within a data management plan).

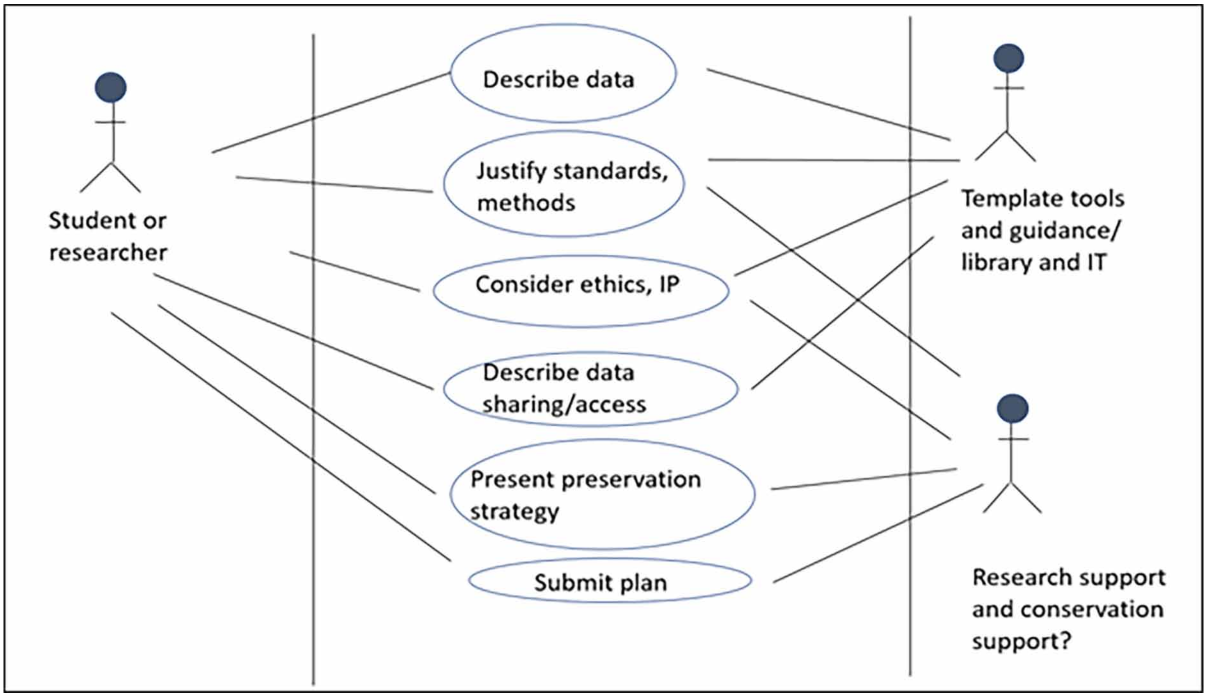
Academic library staff in Jordan has recognized a data repository for research data management as an important part of library infrastructure (Hamad et al., 2021). In a new system that allows for separation of the research data from the rest of the thesis/dissertation, use case descriptions should provide a description of the interaction between the users of the system (actors, such as students or library staff), and the high level functions of the system (e.g. submit research data management plan, check research data management plan, approve data management plan, submit thesis/dissertation, validate submission, assign identifiers, check research data sets, perform data curation, assign metadata describing content, assign metadata for preservation information).

Example. Use case description for use case with goal “submit research data management plan.” These steps are likely to appear on most templates (provided by funders or data preservation specialists) for preparing a research data management plan, but there will be variations in the details.

- Describe data to be collected or created.
- Describe and justify standards or methodologies for data collection and management of the data.
- Demonstrate that ethical guidelines and intellectual property considerations will be met.
- Describe plans for data sharing and access.
- Present preservation strategy.
- Submit completed research data management plan.

Use case diagrams, such as the diagram in Figure 4, should help the University of Jordan library and IT services, working together with the collaborating universities, decide on the main actor roles and functions of a new system. As the diagram indicates, the roles of research support and preservation support may be merged or separated. Library and IT services may support the provision of template tools and guidance. If guidance suggests that the research data management plan should be updated, the system needs to support that functionality, as well as ensuring the plans for preservation and access are in line with current guidance, and with the arrangements made between the University of Jordan and collaborating universities.

*Figure 4. Use case diagram "Submit a research data management plan"*


Student or researcher
Describe data
Justify standards, methods
Consider ethics, IP
Describe data sharing/access
Present preservation strategy
Submit plan
Template tools and guidance/ library and IT
Research support and conservation support?

## Case 6: "We feel fine": Big Data Observations of Citizen Sentiment About State Institutions and Social Inclusion—A Case Study from Canada

### Project Summary

This project was a collaboration between researchers at three different iSchools (led by the University of British Columbia, Canada, School of Information), which studied the relationship between citizen trust in state institutions and social protest. As archivists, librarians, and information professionals increasingly provide reference services to researchers using computational techniques to explore collections, this project was motivated, in part, by the research team's desire to obtain a better understanding of such techniques in order to better train students in how to support new computational research techniques. Using visual analysis of approximately 11 million sentiment classified Tweets from the period of the 2014 Brazilian World Cup, the study explored: 1). How Brazilian citizens felt about their state institutions at the time; 2). How these feelings connected to their sentiments about Brazilian Federal and State Government and politicians and; 3). How such sentiments translated into collective behaviors. Results revealed that the 2014 World Cup protests in Brazil sprang from a wide range of grievances coupled with a relative sense of deprivation compared with emergent comparative "standards." This sense of grievance gave rise to sentiments that activated online protests, which may have led to other forms of social protest, such as demonstrations.

## Project Highlights

- Conducted an initial “big picture” harvest of 11 million Tweets.
- Conducted iterative visual exploratory analysis of the corpus of Tweets using Natural Language Processing to extract relevant search terms representing the key focal topics of the study.
- Used search terms to harvest Tweets and classify them by sentiment.

*Table 5. Data science skills, workflow, and program outcomes*

| Data Science Skills and Workflow | Program Outcomes |
|---|---|
| ● Coding.<br>● Data extraction.<br>● Predictive modelling.<br>● Natural language processing.<br>● Sentiment analysis.<br>● Data visualization and visual analytics. | ● The completed analysis offers the strongest evidence in favor of the “relative deprivation” theory of social protest. In social theory, relative deprivation defines a state in which an individual or group has a sense of lacking something that another individual or group has and to which they feel some entitlement.<br>● Exploration and hypothesis testing using Twitter data with sentiment opinion trends provides an example of a visual exploratory analysis that benefits from integrating statistical learning modeling. |
| **Data Science Tools Used**: Python™, Oracle® Javascript™, Apache® Cassandra™ NoSQL database, STACKS (open source data collection tool), SentiStrength (sentiment classifier), Sifter™ (for data collection), IN-SPIRE (for visual text analysis), Natural Language Toolkit (NLTK) package (for n-gram categorization). | |

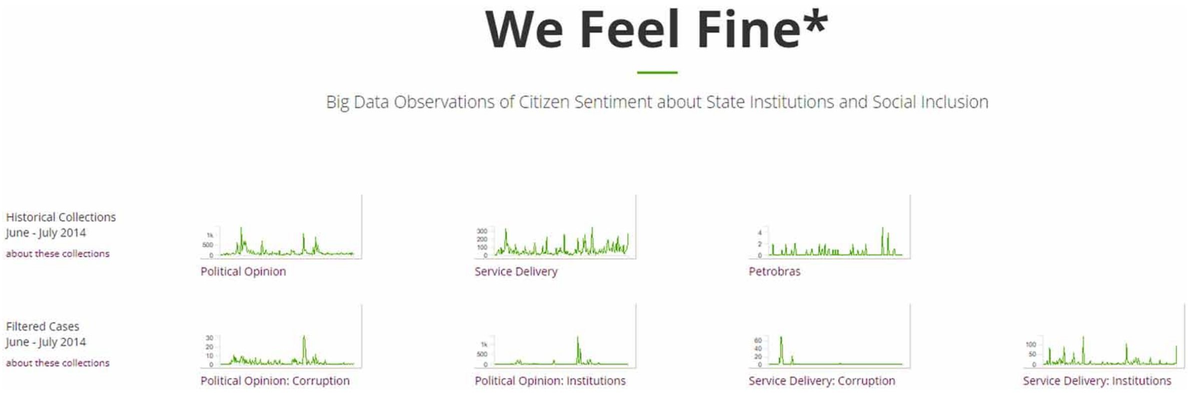
## Impact and Lessons Learned

The approach taken in the Canadian study provided novel insights into issues of concern to Brazilian citizens and their feelings about the quality of service delivery, state institutions, politicians, political parties, and policy choices around the time of the 2014 World Cup. By connecting the analysis to various explanatory hypotheses on the relationship between citizen trust and social protest, the study also found support for theories of relative deprivation as a cause of protest, and for theories of digitally-mediated modes of contestation (Calderon et al., 2015).

The use of big data analytics made it possible to observe the protests from a distance, both in terms of space (i.e. research team did not travel to Brazil), and time (i.e. study used historical Tweets). As such, the study provided insight into the actual experiences of a subset of Brazilian citizens as expressed naturalistically in their own Tweets at a specific point in time. This approach differs from the collection of survey data, which is not naturalistic and may prime citizens with questions that frame their experiences in ways that do not reveal their own thoughts and attitudes. On the other hand, this type of big data analytics does not offer a representative sample of views. The approach also differs from ethnographic field studies, which require the researcher to spend time in the field collecting detailed observational data. As such, big data analytics serves to complement these other approaches to understanding development issues in context.

*Figure 5. Graphical depictions of big data observations of citizen sentiment about state institutions and social inclusion (Calderon et al., 2015)*

*Note: Further information is available at http://www.sfu.ca/~ncaldero/wb-vis/index.htmlChallenges with Global LIS Data Science Projects*



We Feel Fine*

Big Data Observations of Citizen Sentiment about State Institutions and Social Inclusion

## CONCLUSION

While a plethora of DS examples and applications exist globally, locating literature that highlights projects facilitated by, or partnered with members of the LIS field, is challenging. When literature on LIS DS projects has been published, most appear to have originated in the United States, demonstrating a lack of global examples and voices. Additionally, publications may involve library collections or data, but analyses are often led by computer scientists or other non-librarians. The lack of formally published literature examples may exist for a few reasons, namely a lack of formal training, skills, funding, or infrastructure dedicated to DS activities. While a number of LIS schools have recently incorporated DS courses and specializations, these areas have not been traditionally incorporated into LIS curricula, so many current practitioners are only beginning to learn a variety of DS skills, tools, and techniques. In an area where experiential learning is critical and for which there is constant change (e.g., computer programming, artificial intelligence, machine learning), LIS practitioners may still be in the learning phase around DS, which could impact their ability to share their DS efforts through scholarly venues.

Due to long existing budget constraints on libraries globally, training opportunities should be offered at low or no cost and be focused on roles and applications for librarians and library users. Workshops or bootcamps are a few examples of training opportunities in which librarians and users can participate. While no longer offered, one example that originated from the Harvard-Smithsonian Center for Astrophysics John G. Wolbach Library and the Harvard Library, Data Science Training for Librarians (DST4L), is a three day, in-person workshop aimed at upskilling librarians in DS techniques to meet the growing demands of their communities (DST4L, 2016). In-person workshops and bootcamps offer invaluable opportunities for networking and hands-on instruction, but the best method of reaching librarians and users globally will be by offering open, self-paced resources. Major organizations such as the Association of Research Libraries (ARL), the International Federation of Library Associations and Institutions (IFLA), and the Ligue des Bibliothèques Européennes de Recherche Association of European Research

Libraries (LIBER) have a unique opportunity to coordinate efforts for efficiency, reduce duplication, and increase access to bootcamp style instructional opportunities. Other ideas beyond workshops and bootcamps can be found in open conferences specific to library-specific DS projects. Many conferences have shifted to being offered remotely, as a result of the 2020 Covid-19 pandemic, which greatly reduces operational costs and can facilitate participation for users globally. Finally, providing various avenues for funding to support DS initiatives through mini-grants or awards can help facilitate DS inquiry, which can be especially helpful to those in low to middle income countries where funding may not be as robust.

Initiating a DS program can be resource intensive, both from a funding and infrastructure standpoint. Libraries have been increasingly facing collections challenges, which include the significant amount of funding required to support acquisition of data sets. For those hoping to delve into DS techniques using data sets, significant challenges can be found related to cost, licensing (e.g., agreement on who can access data), how long data can be held, and how it can be maintained. Robust infrastructure (often provided by campus Information Technology partners), is critical; some institutions may not have the funding or the campus infrastructure (physically or in the Cloud) to fully accommodate.

The need to provide readily available training opportunities to increase capacity around DS skill building is growing. From provision of carpentries (Library Carpentry, n.d.), community-learning settings for LIS professionals to acquire software and data skills, to offering learning modules and workshops on programs such as R, Python™ and Tableau®, libraries are seeking additional ways to provide their staff with DS skill building opportunities. For institutions that do not have an established DS program, school, or center, or where resources for professional development are limited, it is imperative that library staff have mechanisms to learn new skills without having to take on the additional burden of funding these types of initiatives.

## FUTURE RECOMMENDATIONS

Numerous libraries have established digital repositories for scholars to disseminate and preserve their work (Big Ten Academic Alliance, n.d.; Jisc, n.d.). However, currently, no central repository exists to specifically highlight DS-related projects being undertaken with and by LIS experts on a global scale. To mitigate the challenge of identifying these types of projects, facilitate knowledge sharing, and build capacity around how libraries can partner in DS initiatives, an open repository that highlights LIS DS Projects and Case Studies could be established. Learning from the creation of platforms, such as CochraneCrowd (Cochrane Crowd, 2020) and PREReview (PREreview, 2022), global, crowd-sourced communities, where individuals can meet, collaborate, aid in evidence synthesis, and preprint review tasks, can help establish what an open repository for DS projects could entail and how a participatory process can be infused to encourage ideation, finding collaborators, seeking review for projects, and DS materials to assist with skill building.

An international DS repository could show case examples of DS integration in multiple disciplines and further knowledge and understanding of DS applications in LIS. Research Libraries UK (n.d.) has started the process of highlighting case studies around digital scholarships. Similarly, an organization such as the IFLA could be central in creating an open LIS DS repository that could expand access and increase awareness of how libraries are involved and can collaborate in DS. Creating mechanisms that increase opportunities for cross-institutional collaboration can help build capacity and bring together diverse skill sets and perspectives, which is an excellent way to encourage global collaboration and partnership. Another way to further global LIS DS awareness and to normalize DS work in libraries is to establish dedicated columns in library journals, which focus on the intersection of libraries and DS.

Whether it is through an organizational body or the scholarly domain, it is evident that priority needs to be placed on establishing a mechanism to share information and increase availability of DS projects so that colleagues across the globe can learn from, with, and by one another, and advance the field of DS in libraries.

## REFERENCES

Al-Barashdi, H., & Al-Karousi, R. (2019). Big data in academic libraries: Literature review and future research directions. *Journal of Information Studies & Technology (JIS&T), 2018*(2). doi:10.5339/jist.2018.13

Bain, J. (2020). *TIGER: Using artificial intelligence to discover our collections.* State Library of New South Wales. https://www.sl.nsw.gov.au/blogs/tiger-using-artificial-intelligence-discover-our-collections

Big Ten Academic Alliance. (n.d.). *Big Ten Academic Alliance Open Access Repositories*. Big Academic Alliance. https://btaa.org/library/scholarly-communication/open-access-repositories

Bittner, K., & Spence, I. (2003). *Use case modeling*. Addison-Wesley Professional.

Blummer, B., & Kenton, J. M. (2018). Big data and libraries: Identifying themes in the literature. *Internet Reference Services Quarterly*, *23*(1–2), 15–40. doi:10.1080/10875301.2018.1524337

Briney, K. (2017). *Data Visualization Camp Instructional Materials: UWN Libraries Instructional Materials*. UWMilwaukee UNW Digital Commons. https://dc.uwm.edu/lib_staff_files/4

Burton, M., Lyon, L., Erdmann, C., & Tijerina, B. (2018). *Shifting to data savvy: The future of data science in libraries*. D-Scholarship @ Pitt. https://d-scholarship.pitt.edu/33891/

Calderon, N. A., Fisher, B., Hemsley, J., Ceskavich, B., Jansen, G., Marciano, R., & Lemieux, V. L. (2015). Mixed-initiative social media analytics at the World Bank. Observations of citizen sentiment in Twitter data to explore" trust" of political actors and state institutions and its relationship to social protest. In *2015 IEEE International Conference on Big Data, (Big Data)* (pp. 1678–1687). IEEE 10.1109/BigData.2015.7363939

Chiware, E., & Mathe, Z. (2015). Academic libraries' role in research data management services: A South African perspective. *South African Journal of Library and Information Science*, *81*(2), 1–10. doi:10.7553/81-2-1563

Clark, J., Glasziou, P., Del Mar, C., Bannach-Brown, A., Stehlik, P., & Scott, A. M. (2020). A full systematic review was completed in 2 weeks using automation tools: A case study. *Journal of Clinical Epidemiology*, *121*, 81–90. doi:10.1016/j.jclinepi.2020.01.008 PMID:32004673

Cochrane Crowd. (2020). *What is Cochrane crowd?* You can make a difference. https://crowd.cochrane.org/

Cox, A. M., Kennan, M. A., Lyon, L., & Pinfield, S. (2017). Developments in research data management in academic libraries: Towards an understanding of research data service maturity. *Journal of the Association for Information Science and Technology*, *68*(9), 2182–2200. doi:10.1002/asi.23781

Data Scientist Training for Libraries (DST4L). (2016). *About.* DST4L at DTU Library once more. http://www.dst4l.info/about.html

Electronic Information for Libraries (EIFL). (n.d.). *EIFL in Ethiopia*. EIFL. https://www.eifl.net/country/ethiopia

Erdmann, C. (2015). Data scientist training for librarians. *Open Science at the Frontiers of Librarianship, 492*, 31–37. http://www.aspbooks.org/a/ volumes/article_details/?paper_id=36774

Federer, L., Clarke, S. C., & Zaringhalam, M. (2020). Developing the librarian workforce for data science and open science. doi:10.31219/osf.io/uycaxosf.io/uycax

Gerrard, D. M., Mooney, J. E., & Thompson, D. (2018). Digital preservation at big data scales: Proposing a step-change in preservation system architectures. *Library Hi Tech*, *36*(3), 524–538. doi:10.1108/LHT-06-2017-0122

Hamad, F., Al-Fadel, M., & Al-Soub, A. (2021). Awareness of research data management services at academic libraries in Jordan: Roles, responsibilities and challenges. *New Review of Academic Librarianship*, *27*(1), 76–96. doi:10.1080/13614533.2019.1691027

Hamad, F., Fakhuri, H., & Abdel Jabbar, S. (2020). Big data opportunities and challenges for analytics strategies in Jordanian Academic Libraries. *New Review of Academic Librarianship*, 1–24. doi:10.1080/13614533.2020.1764071

Hutchinson, T. (2018). Protecting privacy in the archives: Supervised machine learning and born-digital records. In *2018 IEEE International Conference on Big Data (Big Data)*, (pp. 2696–2701). IEEE. 10.1109/BigData.2018.8621929

International Federation of Library Associations and Institutions (IFLA) & the Big Data Special Interest Group. (2018). *A concept data science framework for libraries*. https://www.ifla.org/wp-content/uploads/2019/05/assets/big-data/publications/a_concept_data_science_framework_ for_libraries.pdf

JISC. (n.d.). *Welcome to OpenDOAR*. OpenDOAR. https://v2.sherpa.ac.uk/opendoar/

Kettenis, J. (2007). *Getting started with use case modeling.* An Oracle white paper, Oracle Corporation. https://www.oracle.com/technetwork/developer-tools/jdev/gettingstartedwithusecasemodeling-133857.pdf

Library Carpentry. (n.d.) *About Us*. https://librarycarpentry.org/about/

Library of Jordan University. (2017). *Theses Deposit Center*. University of Jordan. http://library.ju.edu.jo/ EN-library/ENlib_Theses.aspx

Ligue des Bibliothèques Européennes de Recherche (LIBER). (n.d.). *Liber Europe. Welcome to LIBER*. https://libereurope.eu/

Limani, F., Latif, A., & Tochtermann, K. (2018, September). Linked publications and research data: Use cases for digital libraries. In E. Méndez, F. Crestani, C. Ribeiro, G. David, & J. C. Lopes (Eds.), *International Conference on Theory and Practice of Digital Libraries* (pp. 363–367). Springer. 10.1007/978-3-030-00066-0_41

Mani, N. S., Cawley, M., Bruckner, L., Casden, J., Dodd, A., Henley, A., Jansen, M., McGarty, J., McKeehan, M., Morris, S., Triumph, T., Venlet, J., & Williams, J. (2020, March). *University libraries: Data science framework*. University of North Carolina at Chapel Hill. https://cdr.lib.unc.edu/concern/journals/3n204432d

Miller, C. (2018). *Understanding how librarians can support data science and big data*. The Marquee; National Library of Medicine. https://news.nnlm.gov/mar/2018/09/24/understanding-how-librarians-can-support-data-science-and-big-data/

Ortiz-Repiso, V., Greenberg, J., & Calzada-Prado, J. (2018). A cross-institutional analysis of data-related curricula in information science programmes: A focused look at the iSchools. *Journal of Information Science*, *44*(6), 768–784. doi:10.1177/0165551517748149

PREreview. (2022). *Mission and values*. https://www.prereview.org/

Research Libraries UK. (n.d.). *RLUK Strategy 2018-2021: Reshaping Scholarship*. Reshaping Scholarship. https://www.rluk.ac.uk/reshaping-scholarship-rluk-strategy-2018-21/

Schöpfel, J., Primož, J., Prost, H., Malleret, C., Češarek, A., & Koler-Povh, T. (2015, December). Dissertations and data. In *GL17 International Conference on Grey Literature*. https://hal.univ-lille.fr/hal-01285304

Semeler, A. R., Pinto, A. L., & Rozados, H. B. F. (2019). Data science in data librarianship: Core competencies of a data librarian. *Journal of Librarianship and Information Science*, *51*(3), 771–780. doi:10.1177/0961000617742465

The Library of Congress. (2021). PREMIS. *The PREMIS Data Dictionary for Preservation Metadata*. https://www.loc.gov/standards/premis/

University of California San Francisco Library. (n.d.). *Programming for Data Science*. ICSF Library. https://www. library.ucsf.edu/ask-an-expert/data-science/programming/

University of Miami Libraries. (2021). *Workshops*. University of Miami. https://www.library.miami.edu/research/workshops-tutorials.html

University of Tartu Library. (n.d.). *Courses*. Open Science. https://utlib.ut.ee/en/courses

University of Virginia Library. (2022, Spring). *Workshops. Research Services and Sciences*. https://data.library. virginia.edu/training/

Urquhart, C., & Tbaishat, D. (2017). Analysing activities, roles and processes. *Information Systems: Process and Practice, 2,* 71. https://www.facetpublishing.co.uk/page/ detail/?k=9781783302413*

Virkus, S., & Garoufallou, E. (2020). Data science and its relationship to library and information science: A content analysis. *Data Technologies and Applications*, *54*(5), 643–663. doi:10.1108/DTA-07-2020-0167

## KEY TERMS AND DEFINITIONS

**Handwritten Text Recognition (HTR):** Software that uses artificial intelligence to create transcriptions of handwritten documents.

**Jupyter Notebook:** An open source, online web application which lets one write and interact with plain text, live code, images, and charts.

**Knowledge Graph:** A structured (even semantic) definition of resources and the links between them, resulting in a graph. Knowledge graphs represent knowledge about a domain of interest, including both the factual data, and the vocabularies used to describe them.

**Lemming (or Lemmatization):** A text normalization technique within the field of Natural Language Processing (NLP) that considers the context in which a word is being used to help define it and link it to a broader list of related terms. Lemming is typically seen as a step beyond 'stemming' techniques and attempts to identify the 'root' of a word to help contextualize it rather than just its inflections. For example, the lemma of the word 'was' could be identified as 'is' or 'be'.

**Natural Language Processing (NLP):** A process (using computational linguistics, statistical, machine learning, and deep learning) in which computers are able to understand information (both written and spoken text) in a similar manner in which a human would relay information.

**Optical Character Recognition (OCR):** The process of converting images with text (static images) to electronic text that can then be searched, indexed, or otherwise used; this process is usually initiated with human input using training data, but over time, machine learning algorithms are able to process more accurately (after "learning") and are able to function with minimal human input.

**Resource Description Framework (RDF):** A knowledge representation framework, whose model uses relatively simple statements of the form subject-predicate-object. The model not only captures the "resources" of a situation, but also allows the specification of how these resources link to each-other. Multiple such statements can be made to represent a certain situation, thus forming an RDF graph.

**Stemming:** A text normalization technique within the field of Natural Language Processing (NLP) that considers the possible prefixes and suffixes of a specific word to identify other words related to it. For example, the word 'walk' which could be linked to 'walks', 'walking' 'walker', etc.